

FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Modelos de Aprendizagem Computacional para previsão da eficácia de medicamentos avaliados em linhas celulares

Pedro Miguel Santos Ferreira



Mestrado Integrado em Engenharia Informática e Computação

Orientador: Rui Camacho

23 de Fevereiro de 2021

Modelos de Aprendizagem Computacional para previsão da eficácia de medicamentos avaliados em linhas celulares

Pedro Miguel Santos Ferreira

Mestrado Integrado em Engenharia Informática e Computação

Resumo

Assiste-se, atualmente, a um cada vez mais significativo número de doenças no mundo e, como resposta a esta evidência, a uma maior quantidade disponível de medicamentos que visam potenciar a eficácia dos sistemas de saúde. A atual disponibilidade de medicamentos, a nível comercial, é bastante relevante, sendo os seus benefícios, bem como a sua importância terapêutica, absolutamente irrefutáveis. Torna-se assim incontestável, quer para a longevidade, quer para a qualidade da vida humana, a importância de novos medicamentos. O processo de conceção de medicamentos passa por diversas etapas até que se atinja a fase final, que corresponde aos testes clínicos. Uma etapa do processo de conceção de medicamentos é testar *in vivo* o novo medicamento sintetizado. Esta fase, além de ser bastante dispendiosa, pois tem inerentes custos muito elevados, é também dolorosa para os animais que são submetidos aos testes. Para evitar ou pelo menos reduzir os testes em animais, as empresas farmacêuticas e os laboratórios de investigação optam por um procedimento clínico designado biópsia, através do qual obtêm uma pequena amostra de tecido maligno (humano ou animal), mas que é bastante inofensiva para o fornecedor biológico. Com essas amostras, pode-se replicar "*ad infinitum*" as células em laboratório. A cultura de novas células a partir de uma célula original é designada por linha celular. Dessa forma, podem ser testados novos medicamentos, em milhares de doses diferentes, em amostras desses tecidos replicados, de uma maneira bastante menos onerosa, permitindo, por outro lado, monitorizar os seus efeitos bem como a sua eficácia. Estes tecidos replicados também permitem aos investigadores testar novas abordagens e terem um processo adequado para comparação de resultados. O objectivo desta dissertação é a aplicação de *Data Mining* a dados públicos, já utilizados em um estudo original para o problema de regressão, para conseguir prever a eficácia de medicamentos em linhas celulares. Para além dos algoritmos utilizados no estudo original, foram introduzidos algoritmos diferentes, foi também feita uma análise de selecção de atributos (*feature selection*), bem como a transformação do problema de regressão em classificação e respectiva análise. Uma última contribuição foi o enriquecimento dos *data sets* com informação obtida por utilização de técnicas de *Graph Mining*.

Abstract

There is currently a significant increase of diseases in the world and, therefore, there is a greater amount of medication available to provide greater effectiveness in health systems. The current availability of medicines at the commercial level is quite relevant, and their benefits and therapeutic importance are irrefutable. Thus, the importance of new medicines is undeniable for human life. There are several stages during the drug design process until it reaches the final phase that corresponds to clinical tests. One step in the drug design process is to test the new synthesized drug in vivo. This phase is characterized by being an expensive phase in terms of costs as well as painful for the animals that are submitted to the tests. To avoid or at least reduce, animal testing, pharmaceutical companies and research laboratories opt for a procedure called biopsy where they obtain a small sample of malignant tissue (human or animal), which is quite harmless to the biological supplier. These samples can replicate "ad infinitum" cells in the laboratory where each cell replicated in this way is called a cell line. That way, they can test new drugs in thousands of different doses on samples of these replicated tissues in a less expensive way and monitor their effects as well as their effectiveness. These replicated tissues also allow researchers to test new approaches and have an appropriate process for comparing results. The objective of this dissertation is the application of Data Mining to public data, already used in an original study for the regression problem, that reports the results of the effectiveness of drugs using cell lines. In addition to the algorithms used in the original study, different algorithms were introduced, a feature selection was carried out, as well as the transformation of the regression problem into classification one and respective analysis. The final experiment was the addition of drug features provided with Graph Mining to the data set.

Agradecimentos

À Faculdade de Engenharia da Universidade do Porto, por me proporcionar as ferramentas que tanto contribuíram para a minha formação académica e pessoal;

Ao Senhor Professor Rui Camacho, pela paciência, pela disponibilidade, pelos bons conselhos, pelas orientações sempre pertinentes;

Aos meus familiares e aos meus amigos, que acreditaram em mim, fazendo com que esta caminhada se tornasse mais jovial.

Ao meu irmão, à minha cunhada, à Leonor e ao Afonso, meus afilhados, e ainda à minha Madrinha, o meu obrigado a todos, pelo apoio incondicional.

Por último, Aos meus Pais, Manuela e Hélder, que estiveram sempre, sempre, ao meu lado, um Obrigado especial, pela educação, pelo incentivo, pelo carinho e pelo apoio sem limites. Obrigado por TUDO!

Sem vocês, não teria sido a mesma coisa!

Pedro Miguel Santos Ferreira

“Experience teaches only the teachable.”

Aldous Huxley

Conteúdo

1	Introdução	1
1.1	Motivação e Objectivos	1
1.2	Estrutura da Dissertação	2
2	Quimioinformática e Data Mining	3
2.1	Quimioinformática	3
2.1.1	Conceitos de Biologia	5
2.2	Ferramentas de Quimioinformática	6
2.3	Fontes de Informação	7
2.4	<i>Data Mining</i>	8
2.4.1	Objectivos de <i>Data Mining</i>	9
2.4.2	Metodologias de <i>Data Mining</i>	10
2.4.3	Algoritmos de <i>Data Mining</i>	13
2.4.4	Ferramentas de <i>Data Mining</i>	19
2.4.5	Métricas de avaliação	20
2.5	Metodologias de Avaliação	22
2.6	Repositórios Web Relevantes	23
2.7	Trabalhos relacionados	25
3	Metodologia	27
3.1	Descrição dos dados originais	27
3.2	Pré-Processamento dos dados originais	28
3.2.1	Características genómicas	28
3.2.2	Características dos medicamentos	29
3.3	Experiências	30
3.3.1	Implementação e avaliação dos modelos	30
3.4	<i>Feature Selection</i>	31
3.5	Classificação	32
3.6	<i>Graph Mining</i>	33
4	Experiências e Resultados	35
4.1	Conjunto de dados completos	35
4.1.1	Regressão	35
4.1.2	Classificação	40
4.2	<i>Feature Selection</i>	44
4.2.1	Ganho de informação mútua - Top 100	46
4.2.2	Ganho de informação mútua - Top 500	48
4.2.3	Correlação de Pearson - Top 100	50

4.2.4	Correlação de Pearson - Top 500	51
4.3	<i>Graph Mining</i>	51
4.4	Discussão de Resultados	52
4.4.1	Regressão	52
4.4.2	Classificação	52
4.4.3	<i>Feature Selection</i>	53
4.4.4	<i>Graph Mining</i>	53
5	Conclusões e Trabalho Futuro	55
5.1	Satisfação dos Objetivos	55
5.2	Limitações e Trabalho Futuro	56
	Referências	57
A	Características	63
A.1	Características das linhas celulares	63
A.2	Descritores dos compostos	65
A.3	<i>Fingerprints</i> dos compostos	68
B	<i>Feature Selection</i>	77
B.1	Regressão	78
B.1.1	Correlação de Pearson	78
B.1.2	Ganho de informação mútua	84
B.2	Classificação	89
B.2.1	Ganho de informação mútua	89
C	<i>Conjunto de dados completos</i>	95
C.1	Regressão	95
C.2	Classificação	96

Lista de Figuras

2.1	Fases da metodologia CRISP-DM (Chapman et al., 2000)	11
2.2	Fases da metodologia SEMMA (Pereira, 2005)	12
2.3	Diagrama resumido de uma rede neuronal (Lopes, 2018)	14
2.4	Árvore de decisão (Viana Gomes, 2011)	15
2.5	Função logística (Viana Gomes, 2011)	16
2.6	Esquema de um algoritmo baseado em Gradient Boosting. Figura adaptada de <i>A Beginner's guide to XGBoost</i>	17
2.7	Crescimento de árvore XGBoost vs LightGBM. Figura adaptada de XGBoost, LightGBM or CatBoost — <i>which boosting algorithm should I use?</i>	18
2.8	Exemplo do método <i>Hold-Out</i> (Namora, 2017)	22
2.9	Exemplo do método <i>k-fold Cross-Validation</i> (Namora, 2017)	23
2.10	Exemplo do método <i>Leave-one-out</i> (Namora, 2017)	23
3.1	Distribuição dos valores IC50	28
3.2	Visão Geral	30

Lista de Tabelas

2.1	Tarefas de <i>Data Mining</i> (Pereira, 2005)	10
4.1	Parâmetros usados no algoritmo <i>Random Forest</i>	35
4.2	Resultados com descritores moleculares para o <i>Random Forest</i>	35
4.3	Resultados com descritores moleculares e <i>fingerprints</i> para o <i>Random Forest</i>	36
4.4	Resultados do <i>blind test</i> com descritores moleculares para o <i>Random Forest</i>	36
4.5	Resultados do <i>blind test</i> com descritores moleculares e <i>fingerprints</i> para o <i>Random Forest</i>	36
4.6	Parâmetros no algoritmo <i>SVM</i>	36
4.7	Resultados com descritores moleculares para o <i>SVM</i>	37
4.8	Resultados com descritores moleculares e <i>fingerprints</i> para o <i>SVM</i>	37
4.9	Resultados do <i>blind test</i> com descritores moleculares para o <i>SVM</i>	37
4.10	Resultados do <i>blind test</i> com descritores moleculares e <i>fingerprints</i> para o <i>SVM</i>	37
4.11	Parâmetros no algoritmo <i>XGBoost</i>	38
4.12	Resultados com descritores moleculares para o <i>XGBoost</i>	38
4.13	Resultados com descritores moleculares e <i>fingerprints</i> para o <i>XGBoost</i>	38
4.14	Resultados do <i>blind test</i> com descritores moleculares para o <i>XGBoost</i>	38
4.15	Resultados do <i>blind test</i> com descritores moleculares e <i>fingerprints</i> para o <i>XGBoost</i>	39
4.16	Parâmetros utilizados no algoritmo <i>ANN</i>	39
4.17	Resultados com descritores moleculares para o <i>ANN</i>	39
4.18	Resultados com descritores moleculares e <i>fingerprints</i> para o <i>ANN</i>	40
4.19	Resultados do <i>blind test</i> com descritores moleculares para o <i>ANN</i>	40
4.20	Resultados do <i>blind test</i> com descritores moleculares e <i>fingerprints</i> para o <i>ANN</i>	40
4.21	Resultados RF 1D2D <i>Non Stringent</i>	40
4.22	Resultados RF 1D2D <i>Stringent</i>	40
4.23	Resultados RF 1D2DFP <i>Non Stringent</i>	41
4.24	Resultados RF 1D2DFP <i>Stringent</i>	41
4.25	Resultados <i>Blind Test</i> RF 1D2D <i>Non Stringent</i>	41
4.26	Resultados <i>Blind Test</i> RF 1D2D <i>Stringent</i>	41
4.27	Resultados <i>Blind Test</i> RF 1D2DFP <i>Non Stringent</i>	41
4.28	Resultados <i>Blind Test</i> RF 1D2DFP <i>Stringent</i>	41
4.29	Resultados SVM 1D2D <i>Non Stringent</i>	41
4.30	Resultados SVM 1D2D <i>Stringent</i>	41
4.31	Resultados SVM 1D2DFP <i>Non Stringent</i>	42
4.32	Resultados SVM 1D2DFP <i>Stringent</i>	42
4.33	Resultados <i>Blind Test</i> SVM 1D2D <i>Non Stringent</i>	42
4.34	Resultados <i>Blind Test</i> SVM 1D2D <i>Stringent</i>	42
4.35	Resultados <i>Blind Test</i> SVM 1D2DFP <i>Non Stringent</i>	42

4.36	Resultados <i>Blind Test</i> SVM 1D2DFP <i>Stringent</i>	42
4.37	Resultados XGBoost 1D2D <i>Non Stringent</i>	42
4.38	Resultados XGBoost 1D2D <i>Stringent</i>	42
4.39	Resultados XGBoost 1D2DFP <i>Non Stringent</i>	43
4.40	Resultados XGBoost 1D2DFP <i>Stringent</i>	43
4.41	Resultados <i>Blind Test</i> XGBoost 1D2D <i>Non Stringent</i>	43
4.42	Resultados <i>Blind Test</i> XGBoost 1D2D <i>Stringent</i>	43
4.43	Resultados <i>Blind Test</i> XGBoost 1D2DFP <i>Non Stringent</i>	43
4.44	Resultados <i>Blind Test</i> XGBoost 1D2DFP <i>Stringent</i>	43
4.45	Resultados ANN 1D2D <i>Non Stringent</i>	43
4.46	Resultados ANN 1D2D <i>Stringent</i>	43
4.47	Resultados ANN 1D2DFP <i>Non Stringent</i>	44
4.48	Resultados ANN 1D2DFP <i>Stringent</i>	44
4.49	Resultados <i>Blind Test</i> ANN 1D2D <i>Non Stringent</i>	44
4.50	Resultados <i>Blind Test</i> ANN 1D2D <i>Stringent</i>	44
4.51	Resultados <i>Blind Test</i> ANN 1D2DFP <i>Non Stringent</i>	44
4.52	Resultados <i>Blind Test</i> ANN 1D2DFP <i>Stringent</i>	44
4.53	Correlação de Pearson	45
4.54	Ganho de informação mútua para regressão	45
4.55	Ganho de informação mútua para classificação	45
4.56	Resultados da regressão para os descritores moleculares 1D e 2D - GIM100	46
4.57	Resultados da regressão para os descritores moleculares 1D e 2D e FP - GIM100 . .	46
4.58	Resultados da classificação para os descritores moleculares 1D e 2D - GIM100 . .	47
4.59	Resultados da classificação para os descritores moleculares 1D e 2D e FP - GIM100 .	47
4.60	Resultados da regressão para os descritores moleculares 1D e 2D - GIM500	48
4.61	Resultados da regressão para os descritores moleculares 1D e 2D e FP - GIM500 . .	48
4.62	Resultados da classificação para os descritores moleculares 1D e 2D - GIM500 . .	49
4.63	Resultados da classificação para os descritores moleculares 1D e 2D e FP - GIM500 .	49
4.64	Resultados da regressão para os descritores moleculares 1D e 2D - CP100	50
4.65	Resultados da regressão para os descritores moleculares 1D e 2D e FP - CP100 . .	50
4.66	Resultados da regressão para os descritores moleculares 1D e 2D - CP500	51
4.67	Resultados da regressão para os descritores moleculares 1D e 2D e FP - CP500 . .	51
4.68	Resultados da regressão para os descritores moleculares 1D e 2D e fragmentos - GM .	52
C.1	Resultados da regressão para os descritores moleculares 1D e 2D	95
C.2	Resultados da regressão para os descritores moleculares 1D e 2D e FP	96
C.3	Resultados da classificação para os descritores moleculares 1D e 2D	96
C.4	Resultados da classificação para os descritores moleculares 1D e 2D e FP	97

Abreviaturas e Símbolos

ADN	Ácido Desoxirribonucléico
ANN	Artificial Neural Network
ASCII	American Standard Code for Information Interchange
ChEBI	Chemical Entities of Biological Interest
CRISP-DM	Cross Industry Standard Process for Data Mining
CTD	Comparative Toxicogenomics Database
DM	Data Mining
DMAIC	Define, Measure, Analyze, Improve, Control
DT	Decision tree
GM	Graph Mining
IC50	half maximal Inhibitory Concentration (Concentração Inibitória 50)
IDE	Integrated Development Environment
ILP	Inductive Logic Programming
KEGG	Kyoto Encyclopedia of Genes and Genomes
KNIME	Konstanz Information Miner
LR	Logistic regression
MEDLINE	MEDical Literature aNalysis and retrieval system online
MeSH	Medical Subject Headings
MMR	MissMatch repair
MSI	MicroSatellite Instability (Instabilidade de microssatélites)
NCI	National Cancer Institute
NCI60	National Cancer Institute60
SDF	Standard Database Format
SEMMA	Sample, Explore, Modify, Model, Assessment
SMILES	Simplified Molecular Input Line Entry Specification
SVM	Support Vector Machine
WEKA	Waikato Environment for Knowledge Analysis
XML	eXtensible Markup Language

Capítulo 1

Introdução

O objectivo da dissertação é a aplicação de técnicas de *Data Mining* (DM) a dados públicos relativos a experiências de administração de medicamentos em linhas celulares. A finalidade é construir modelos que prevejam a eficácia dos medicamentos nessas linhas celulares (valores IC50), usando os resultados das experiências *in vitro*, a caracterização das linhas celulares e informações sobre o medicamento em uso. Existem outros projectos que seguiram esta abordagem e compararam-se os resultados obtidos com esses projectos. Assim, a partir de técnicas de *Data Mining*, será possível o desenvolvimento de uma metodologia para a construção de modelos precisos que possam ajudar os especialistas a entender o mecanismo que torna um medicamento eficaz no combate às doenças para as quais foi concebido.

1.1 Motivação e Objectivos

Assiste-se, nos dias de hoje, a um sem-número de doenças que afetam as pessoas e que obrigam a uma procura cada vez mais rápida e mais especializada de medicamentos para as combater

Uma doença é uma condição patológica que afeta uma parte ou a totalidade de um organismo. É uma condição médica associada a sinais e sintomas específicos. Uma doença pode ser causada por fatores originários de uma fonte externa, como uma doença infecciosa, ou pode ser causada por disfunções internas, como doenças autoimunes. Um dos métodos mais notáveis na cura de doenças é a utilização de medicamentos. Todos os anos, existe um grande número de medicamentos que contribuem para aumentar a eficiência e a eficácia dos sistemas de saúde, proporcionando uma melhoria na qualidade de vida da população e também provocando uma diminuição na taxa de mortalidade. Os medicamentos são moléculas que, quando introduzidas no corpo humano, produzem um efeito terapêutico (Isabel e Candeias, 2017).

No entanto, o desenvolvimento de medicamentos causa alguma controvérsia na sociedade, nomeadamente nos testes *in vivo* a que o medicamento é sujeito. Ora, esta metodologia, apesar de inevitável, implica custos elevadíssimos e é bastante contestada, uma vez que obriga à utilização

de animais e uso de processos invasivos e dolorosos. No sentido de fazer face a estes constrangimentos e de facilitar o desenvolvimento de novas formas de tratamento, os laboratórios desenvolvem culturas de células cancerosas através do crescimento de células cancerosas *in vitro*. Estas culturas de células cancerosas possuem as mesmas características genéticas do tecido original e são designadas de linhas celulares. A vantagem deste processo é então testar novos medicamentos nas linhas celulares, em milhares de doses diferentes, de uma maneira menos dispendiosa e monitorizar os seus efeitos bem como a sua eficácia. Estes tecidos replicados também permitem aos investigadores testar novas abordagens e terem um processo adequado para comparação de resultados.

Esta dissertação teve em linha de conta um conjunto de dados utilizados no projeto *Genomics in Drug Sensitivity in Cancer* (Garnett et al., 2012) que reportam os resultados dos efeitos de um grande número de compostos químicos em várias linhas celulares de diferentes tipos de cancro. O nível de sensibilidade das células para esses compostos foram medidos, usando o IC50, que mede o grau de eficácia que uma substância tem que inibir uma determinada função biológica. Visto que o número de possibilidades para testar combinações de compostos celulares é bastante extenso e o tempo e os recursos são fatores chave para realizar as experiências em laboratório, é portanto muito útil usar métodos *Machine Learning* para estimar os resultados de novos medicamentos. A finalidade da dissertação é, então, compreender melhor as relações entre eficácia de um medicamento e as linhas celulares, afim de melhorar a administração do medicamento e fornecer tratamentos mais específicos. Para alcançar este objectivo, será realizada a construção de modelos que possa prever a eficácia de um medicamento. Além de ser capaz de prever a eficácia, também é importante que se possam adquirir informações sobre as diferentes características inerentes às linhas celulares e medicamentos, bem como resultados e efeitos alcançados.

1.2 Estrutura da Dissertação

Para além da introdução, onde é exposta a motivação e os objectivos, esta dissertação contém mais 4 capítulos. No Capítulo 2, é descrito o estado da arte onde se descrevem alguns conceitos de quimioinformática e biologia, indispensáveis para a compreensão desta dissertação, as ferramentas e repositórios web normalmente associados a quimioinformática, bem como o objectivo, metodologias e técnicas de *Data Mining*. Ainda no Capítulo 2 são descritas as ferramentas, os métodos e metodologias de avaliação utilizados para o processo de *Data Mining*.

O Capítulo 3 apresenta a metodologia para a realização da parte experimental da dissertação que inclui desde o pré-processamento do conjunto de dados utilizado até aos algoritmos e os processos de avaliação dos modelos construídos.

No Capítulo 4 são mostrados e comentados os resultados das experiências realizadas durante o desenvolvimento deste projeto.

O Capítulo 5 é reservado às conclusões obtidas pelo desenvolvimento do projeto e por sugestões de trabalho futuro.

Capítulo 2

Quimioinformática e Data Mining

Neste capítulo é apresentada uma visão geral do estado da arte, destacando as matérias fundamentais para a dissertação. As primeiras duas secções contêm breves explicações dos conceitos de quimioinformática, como também de biologia indispensáveis para a compreensão do trabalho realizado. Na terceira secção, apresentamos os objectivos, metodologias, técnicas e ferramentas de *Data Mining*. Por fim, nas últimas secções são discutidas as metodologias de avaliação, bem como os repositórios Web relevantes, que merecem especial destaque no presente trabalho.

2.1 Quimioinformática

A Quimioinformática remete para a aplicação de diversos métodos de Ciência de Computação com o intuito de se resolverem problemas químicos. De um modo geral, estes métodos englobam várias técnicas, designadamente de armazenamento, de processamento e de manipulação dos dados químicos. Assim, a Quimioinformática consiste numa área que se foca nas pequenas moléculas, motivo pelo qual uma das suas aplicações está diretamente associada à procura e posterior identificação de novas estruturas que possam, eventualmente, resultar em novos medicamentos ([Gasteiger e Funatsu, 2006](#)).

Um medicamento consiste numa substância química que, e para além de objetivar a melhoria da qualidade de vida do ser vivo, se utiliza na cura, tratamento, diagnóstico ou prevenção de uma determinada doença. De um modo geral, os medicamentos são subdivididos em várias classes, e de acordo com as suas semelhanças, essa subdivisão opera-se ainda a nível da sua finalidade e estrutura química. De referir que cada medicamento possui uma substância que se encontra biologicamente ativa, com o intuito de exercer as funções enquanto fármaco ([Lopes, 2018](#)).

Atentando no conceito de *pathway*, é importante especificar que este remete para um conjunto de interações entre as várias moléculas existentes numa única célula, levando-a a sofrer determinadas mudanças. Relativamente aos medicamentos, os *pathways* contêm informações cruciais, visto que permitem compreender os mecanismos de ação e interações de cada medicamento, bem como o seu metabolismo e reposicionamento. *Pathways* são conjuntos de reações químicas com um objetivo específico, mais precisamente o de alterar a função anómala de um *pathway*. Para se

compreender na totalidade as reações existentes entre os medicamentos e os *pathways* é crucial ter acesso a uma enorme quantidade de informação, motivo pelo qual as bases de dados que contêm os *pathways* e aqueles nos quais os medicamentos atuam são absolutamente vitais (Zeng et al., 2015).

Por fim, é importante salientar que, e para que seja possível compreender e processar uma determinada estrutura química, através de um computador, é crucial que aquela seja descrita através de uma sequência única. Resumindo, os descritores moleculares, que representam as estruturas químicas através de várias informações, resultam de um procedimento matemático que identifica propriedades químicas e as codificadas numa representação simbólica de uma determinada molécula (Alves et al., 2018).

Formatos de arquivo químico

Existem vários tipos de arquivos para representação molecular, os mais relevantes para esta dissertação são apresentados nas secções seguintes.

Descritores Moleculares

Descritores moleculares são representações matemáticas da estrutura de uma molécula de forma a permitir o cálculo estatístico, ou, para outras tecnologias de informação, serem utilizadas na caracterização da molécula. Eles podem ser obtidos através de medições experimentais ou podem ser obtidos por meio da representação simbólica da molécula. Os descritores moleculares podem ser categorizados pelas suas dimensões nomeadamente descritores moleculares 0D, 1D, 2D ou 3D (R. Todeschini e Mannhold, 2009).

SMILES

Simplified molecular-input line-entry system (SMILES) é uma especificação para representar moléculas como *strings* ASCII (Weininger, 1988). Por exemplo, o Dióxido de carbono, que tem dois elementos de oxigénio e um de carbono conectados através de ligações duplas, seria representada como O=C=O.

Arquivos de tabela química

Os arquivos de tabelas químicas são formatos para representar as informações contidas nas tabelas. SDF é um formato para representar informações químicas. SDF significa formato de dados estruturados, contém a estrutura da molécula, além de algumas informações gerais sobre ela (A. Dalby e Laufer, 1992).

IC50

A concentração inibitória 50 (IC50) é uma medida comum da eficácia de um medicamento. Esta concentração representa a quantidade de um inibidor necessária para reduzir para metade os

efeitos de um processo biológico. A sua utilização universal é devida à existência de uma reduzida variação para diferentes ambientes experimentais, assim como se deve igualmente à facilidade no cálculo do erro padrão ([J. S. Soothill e Girling, 1992](#)).

2.1.1 Conceitos de Biologia

Biomarcadores

Não existe uma definição clara do que é um biomarcador. A definição mais ampla é definida pelo *Center for Biomarkers in Imaging* que considera um biomarcador qualquer "fisiológico, bioquímico ou parâmetro molecular associado à presença e gravidade de estados de doença específicos" ([Strimbu e Tavel, 2010](#)). A pressão arterial ou amostras de urina são exemplos do que pode ser considerado um biomarcador sob esta definição. Os biomarcadores podem realizar uma das três seguintes funções:

- Biomarcadores diagnósticos - podem ajudar a diagnosticar uma condição
- Biomarcadores prognósticos - ajudam a prever a agressividade das doenças
- Biomarcadores preditivos - são utilizados para prever a eficácia de um tratamento

Os Biomarcadores preditivos e prognósticos são fundamentais para identificar medicamentos que podem ter como alvo as características específicas de um tumor e alcançar um desempenho terapêutico superior.

Cultura celular

A cultura celular consiste num conjunto de processos que envolve a distribuição e isolamento de células em um ambiente artificial (*in vitro*) e que tem como objetivo estudar a sua fisiologia e bioquímica ([Chaudry, 2004](#)).

Linhas celulares

As células que pertencem a uma cultura celular têm um tempo de vida útil limitado. Para prevenir a sua morte e facilitar os processos experimentais, estas células são sub-cultivadas em linhas celulares ([Masters, 2000](#)). A principal vantagem deste processo é uma redução na manutenção e no custo de utilização. Por outro lado, as consequências incluem perigos de contaminação entre diferentes culturas numa única linha celular como também têm uma representação menos precisa do tecido original. Existem algumas dúvidas acerca da capacidade de avaliar a eficácia de um medicamento, mas existe um consenso científico de que as linhas celulares são atualmente a melhor ferramenta para a avaliação de medicamentos ([S. V. Sharma e Settleman, 2010](#)). As linhas celulares têm um histórico extenso de utilização na pesquisa do cancro. O *National Cancer Institute 60* (NCI60) introduziu a triagem de linhas celulares para testes de medicamentos contra o cancro, desenvolvida na década de 1980, que continha 60 linhas celulares cancerosas para teste. Isto é

considerado um marco científico na pesquisa do cancro. O interesse na pesquisa de linhas celulares cancerosas tem sido consolidado devido ao aumento da compreensão da heterogeneidade do cancro. Nos dias de hoje estão disponíveis linhas celulares cancerosas com um grande número de linhas celulares que podem ser usadas para testar terapêuticas específicas.

Genoma

O genoma é o material genético contido dentro de um organismo. É principalmente composto pelo ácido desoxirribonucléico (ADN), que contém o conjunto de instruções para o desenvolvimento de um organismo. Cada célula contém uma cópia do ADN e existem análises que demonstram a sua influência na formação de tumores (J. R. Pollack e Brown, 1999). De seguida, são apresentados alguns conceitos de ADN relevantes para a dissertação.

Copy number

O *copy number* é o número de repetições em certas secções do ADN. Cerca de 5 a 10% de ADN humano é composto por repetições e essas repetições podem ajudar a caracterizar um genoma, especialmente porque as anomalias do *copy number* são mais comuns em células malignas devido à frequência de mutações (J. R. Pollack e Brown, 1999).

Fenótipo

De acordo com o NCI, o fenótipo são as "características observáveis" de um organismo que derivam da sua expressão genética (Phe, 2020). O fenótipo *wild type* refere-se ao fenótipo de um organismo presente na natureza, em outras palavras - é a versão padrão. Outras variantes do fenótipo são referidas por mutações. No cancro, uma vez que as células sofrem um número significativo de mutações, fazer uma análise ao fenótipo mutador pode fornecer informações sobre as possibilidades de tratamento.

Instabilidade de Microssatélites

A instabilidade de microssatélites (MSI) é um fenómeno causado por um *missmatch repair* (MMR) no ADN. Está comprovado que a sua presença é um factor fiável para prever o cancro (Shokal e Sharma, 2012). O MMR do ADN é um sistema que corrige erros de ADN durante a replicação e recombinação, mas, se existir algum defeito, as células tendem a acumular erros, o que as torna mais propensas a comportamentos anormais.

2.2 Ferramentas de Quimioinformática

O **PaDEL-Descriptor** (<http://www.yapcwsoft.com/dd/padeldescriptor/>) consiste num *software standalone*, desenvolvido em Java, que objetiva calcular os descritores moleculares e as *fingerprints*. No presente, esta ferramenta em particular consegue calcular 1875 descritores e 12 tipos

distintos de *fingerprints*, designadamente com a ajuda do *The Chemistry Development Kit* ¹. Para além do mais, este *software* também pode ser utilizado enquanto extensão do KNIME ou do RapidMiner (Chun Wei Yap, 2010).

Por sua vez, o **Open Babel** ² é um *software* gratuito e *open-source*, sendo frequentemente usado para pesquisas, análises e conversões de diversos dados químicos, tendo suporte para 111 formatos de arquivos químicos. Além disso, este *software* também fornece uma pesquisa por subestrutura, baseado no SMILES (*Simplified Molecular Input Line Entry Specification*), que remete para uma forma de representação de estruturas químicas com recurso a caracteres ASCII e ao cálculo da *fingerprint* de uma determinada molécula, obtida através do mapeamento de todas as subestruturas lineares e de anel de uma molécula. Por fim, este *software* permite também a criação de coordenadas 2D e 3D de uma molécula, as quais são obtidas através do seu identificador SMILES e da conversão entre estas três estruturas distintas (O'Boyle et al., 2011).

2.3 Fontes de Informação

A **IntSide** ³ é uma base de dados que integra informações biológicas e químicas. O conteúdo desta base de dados provém de diversas fontes, mais precisamente das seguintes: Uniprot, SIDER, MedDRA, CTD, DrugBank, STITCH, Gene Ontology, KEGG e ChEBI (Juan-Blanco et al., 2015).

A **CTD** ⁴ consiste numa base de dados que apresenta informações sobre doenças, genes e medicamentos, sendo que as suas principais vertentes remetem para as interações medicamento-doença, gene-doença e medicamento-gene. Não obstante, a CTD também integra informações da ontologia genética Go, tal como da base de dados KEGG sobre as interações moleculares (Davis et al., 2017).

A **MEDLINE** (*Medical Literature Analysis and Retrieval System Online*) é uma base de dados bibliográficos e é atualizada mensalmente, contendo mais de 18 milhões de referências de artigos científicos. Dada a sua extensão, para facilitar a pesquisa e a utilização dos dados, a base de dados indexa todos os dados existentes, de acordo com palavras-chave específicas, nomeadamente de um sistema designado por MeSH (*Medical Subject Headings*) (<https://www.ncbi.nlm.nih.gov/mesh>). Essencialmente, o MeSH é um dicionário de termos diretamente relacionados com a medicina, organizando todos os conteúdos de acordo com uma hierarquia, o que permite proceder a pesquisas com vários níveis de especificidade. Finalmente, o PubMed (<https://www.ncbi.nlm.nih.gov/pubmed/>) remete para um serviço que permite o acesso gratuito a citações, artigos e resumos fornecidos pela MEDLINE, contendo cerca de 26 milhões de citações da literatura biomédica publicada por esta base de dados bibliográficos (Namora, 2017).

¹(<https://sourceforge.net/projects/cdk/>)

²(http://openbabel.org/wiki/Main_Page)

³(<http://intside.irbbarcelona.org/>)

⁴(<http://ctdbase.org/>)

2.4 Data Mining

O *Data Mining* consiste num processo específico que permite descobrir padrões e adquirir conhecimento através de quantidades substanciais de informação. Esta informação provém de bases de dados, armazéns de dados ou outros repositórios de informação, que são inseridos de modo dinâmico neste sistema. Regra geral, a metodologia utilizada engloba 7 passos específicos, os quais culminam na extração dos dados e, conseqüentemente, na aquisição de conhecimento, mais concretamente: (Pei, 2011).

1. **Limpeza dos dados**, que remete para a extração de todos os dados irrelevantes ou incorretos;
2. **Integração dos dados**, associada à combinação dos dados provenientes de várias fontes;
3. **Seleção de dados**, relacionada com a extração dos dados considerados como sendo relevantes;
4. **Transformação dos dados**, associada à realização de várias operações, nomeadamente de síntese ou de agregação dos dados;
5. **Data Mining**, associado ao processo através do qual vários métodos são aplicados, almejando-se a extração dos vários padrões dos dados e a posterior construção de modelos;
6. **Avaliação dos modelos**, relacionada com a identificação dos padrões que realmente representam conhecimento;
7. **Utilização do conhecimento**, nomeadamente do conhecimento que é adquirido através deste processo para posterior aplicação no uso pretendido

Segundo (Efraim Turban, 2017), o *DM* contribui para a melhoria do processo de tomada de decisão das organizações em geral, visto que permite a exploração de novas oportunidades, seja a nível de produtos, seja a nível de mercados, através da transformação dos dados numa estratégia organizacional. Por conseguinte, o *DM* tem sido amplamente aplicado em várias áreas, designadamente:

1. Informática, para a previsão de eventuais avarias de hardware e violação da rede;
2. Televisão e rádio, para a previsão de programas que possam ser mais rentáveis no horário nobre, maximizando os lucros através da publicidade mais apropriada;
3. Marketing, para a previsão das compras de eventuais produtos e posterior segmentação demográfica dos clientes;
4. Farmácias, com a análise das vendas e o investimento em determinados produtos e/ou marcas;
5. Medicina, com melhorias a nível da identificação de terapias de sucesso para vários tratamentos e da deteção de padrões de sintomas de diversas patologias.

2.4.1 Objectivos de *Data Mining*

Os objetivos do *DM* remetem para duas categorias distintas, designadamente para a previsão e para a descrição, as quais compreendem várias tarefas. A previsão compreende as tarefas de classificação e de regressão, enquanto a descrição engloba as tarefas de associação, de segmentação, de sumarização e de visualização.

A tarefa de classificação está associada à construção de um modelo com o intuito de classificar diversos dados, distribuindo-os em várias classes ou categorias. Assim sendo, cada dado é examinado e classificado de acordo com uma classe previamente definida. No entanto, esta tarefa pressupõe uma certa indeterminabilidade, tendo em consideração que integra o objetivo de previsão ([Freitas, 2000](#)).

A regressão, por sua vez, é utilizada para a definição de um valor para uma determinada variável, contínua e desconhecida. Assim sendo, esta tarefa em particular engloba resultados contínuos, enquanto a classificação, por outro lado, engloba resultados discretos. Não obstante, a regressão pode ser utilizada na execução de uma tarefa de classificação, tendo em consideração que diferentes intervalos de valores contínuos correspondem, na verdade, a classes distintas ([Pereira, 2005](#)).

Atentando nas tarefas de descrição, a associação remete para a determinação dos itens que concorrem numa mesma transação. Um dos exemplos mais comuns remete para os produtos adquiridos no supermercado, que, e regra geral, são colocados juntos num carrinho, o que permite prever a existência de regras de associação entre vários itens/objetos. A segmentação, por outro lado, está associada à divisão de uma população heterogénea em diversos segmentos homogéneos. Nesta tarefa, não existem classes previamente definidas, visto que os registos são agrupados consoante a sua semelhança ([Pereira, 2005](#)) ([Pei, 2011](#)).

A sumarização, e tal como a própria designação permite prever, está associada à utilização de diversos métodos com o intuito de se encontrar uma descrição compacta e objetiva para um determinado conjunto de dados/objetos. Os métodos mais modernos de sumarização resultam de regras de resumo, bem como da descoberta de relações funcionais entre as variáveis, sendo sempre aplicados no âmbito da análise exploratória dos dados e na geração automática de vários relatórios. Por fim, a tarefa de visualização refere-se à apresentação dos resultados obtidos em todo o processo de Data Mining, designadamente através de uma forma visual (gráficos). Esta tarefa pretende facilitar a interpretação dos resultados do processo, motivo pelo qual se descrevem informações bastante complexas através de formas visuais ([Pereira, 2005](#)).

Na Tabela 2.1 pode ver-se uma listagem de todas estas tarefas de *DM*, sendo que para além da sua designação se apresenta uma breve descrição de cada tarefa.

Tarefas de <i>Data Mining</i>	Descrição da tarefa
1. Classificação	Construção de um modelo para classificar dados de acordo com classes previamente definidas.
2. Regressão	Definição de um valor para uma determinada variável contínua desconhecida.
3. Associação	Determinar os itens que tendem a ser adquiridos juntos numa determinada transação.
4. Segmentação	Divisão da população heterogénea em vários grupos homogéneos.
5. Sumarização	Utilização de métodos para descobrir uma definição compacta e objetiva para cada grupo de dados.
6. Visualização	Apresentação dos resultados obtidos através de formas visuais (gráficos).

Tabela 2.1: Tarefas de *Data Mining* (Pereira, 2005)

2.4.2 Metodologias de *Data Mining*

Tendo em consideração que os projetos de *Data Mining* devem seguir um processo de gestão bastante estruturado, e que o próprio *DM* se enquadra num contexto de uma metodologia, é fundamental evidenciar três metodologias em particular: CRISP-DM, SEMMA e DMAIC. A metodologia de **C***ross* **I***ndustry* **S***tandard* **P***rocess* **f***or* **D***ata* **M***ining* (CRISP-DM) foi lançada no ano de 1999 após a sua conceção, em 1996, por quatro empresas (NCR Corporation, SPSS, OHRA e Daimler-Benz). Essencialmente, esta metodologia de *Data Mining* compreende seis fases específicas, as quais representam um ciclo que deve ser seguido (Figura 2.1). Apesar da sequência deste ciclo não ser rígida, depende significativamente do resultado que é obtido em cada fase. As primeiras fases são as mais interativas, objetivando uma compreensão superior das necessidades do negócio em questão, bem como da disponibilidade dos dados (Efraim Turban, 2017)

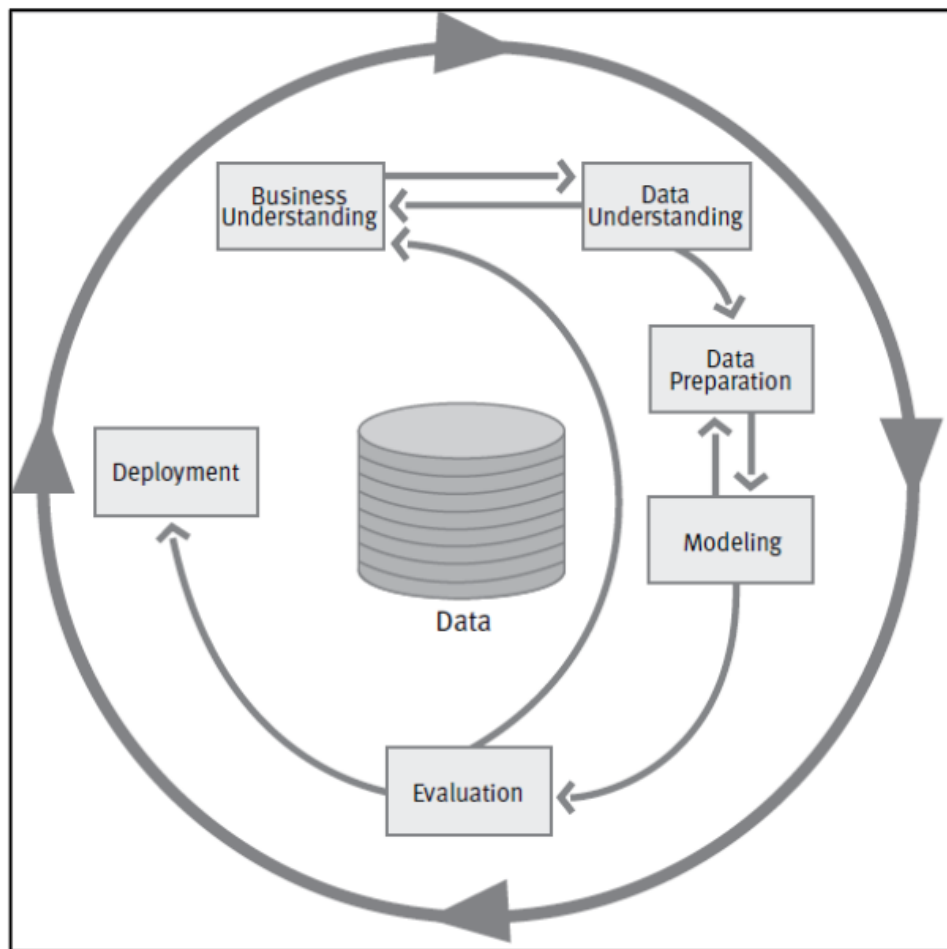


Figura 2.1: Fases da metodologia CRISP-DM (Chapman et al., 2000)

De acordo com (Chapman et al., 2000), as fases da metodologia CRISP-DM desenvolvem-se do seguinte modo:

1. **Compreensão do negócio.** Nesta fase procede-se à análise dos objetivos e requisitos temporais, técnicos e funcionais de acordo com a perspetiva do negócio. Efetua-se, também, o enquadramento dos objetivos e restrições aquando da formulação do problema de *Data Mining* e cria-se uma estratégia para alcançar todos os objetivos previamente definidos;
2. **Compreensão dos dados.** Nesta fase procede-se à recolha e posterior análise dos dados;
3. **Preparação dos dados.** Nesta fase aplica-se um conjunto de atividades para construir um conjunto de dados que possa ser analisado através das ferramentas de modelação. Selecionam-se todos os casos e variáveis para serem analisados;
4. **Modelação.** Nesta fase selecionam-se e aplicam-se as técnicas de *Data Mining* congruentes com os objetivos. Eventualmente, também se procede à calibração do modelo para se otimizarem os resultados;

5. **Avaliação.** Nesta fase efetua-se a avaliação e/ou revisão das atividades que foram previamente realizadas e verifica-se a sua contribuição na concretização dos objetivos estabelecidos;
6. **Implementação.** Nesta fase desenvolvem-se atividades que conduzem à organização do conhecimento e à sua posterior disponibilização (como relatórios). Eventualmente, os resultados obtidos poderão levar ao desenvolvimento de um novo projeto de *Data Mining*.

Já a metodologia **Sample, Explore, Modify, Model, Assessment** (SEMMA) engloba cinco fases distintas, auxiliando no desenvolvimento de um projeto durante todas as suas fases, desde a definição do problema até à sua implementação (Figura 2.2). Assim, a SEMMA disponibiliza uma estrutura para a conceção, criação e posterior evolução dos projetos de *Data Mining*, objetivando o seu sucesso e eficácia (Azevedo e Santos, 2008).

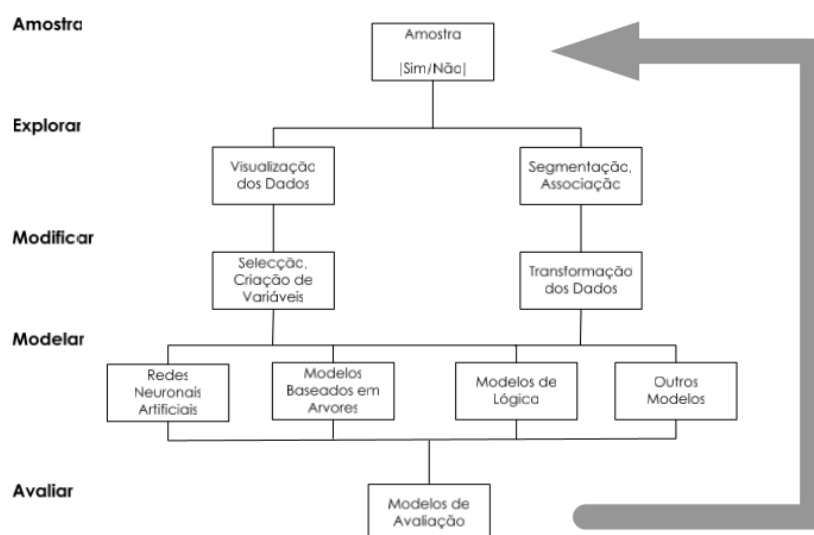


Figura 2.2: Fases da metodologia SEMMA (Pereira, 2005)

Finalmente, a última metodologia que merece especial destaque, apesar de ser pouco abordada na literatura dedicada a esta área, é a metodologia **Define, Measure, Analyze, Improve, Control** (DMAIC), baseada na metodologia *Six Sigma* que é bastante estruturada, focando-se essencialmente na eliminação de qualquer desperdício e dos problemas a nível do controlo da qualidade de diversas áreas de um determinado negócio (Efraim Turban, 2017).

Os passos desta metodologia remetem para a sua própria designação, contando, portanto, com cinco passos em particular. Estes passos representam todo o fluxo do procedimento, sendo perentória a definição de objetivos, de medidas de aplicação e de mecanismos de *feedback*. Esta

metodologia é também utilizada no âmbito do *Data Mining* devido ao facto de melhorar a qualidade e os processos de controlo, sendo possível obter uma eficiência superior nos vários projetos de *Data Mining* a desenvolver ([Efraim Turban, 2017](#)).

2.4.3 Algoritmos de Data Mining

Ao longo dos últimos anos, a investigação em torno das técnicas de *Data Mining* evoluiu significativamente, tendo resultado na publicação de vários trabalhos nesta mesma área. Entre as técnicas mais eficazes, devido ao facto de apresentarem a melhor relação entre eficácia e simplicidade a nível da sua implementação, salientam-se as seguintes: redes neuronais artificiais, árvores de decisão, regressão logística, SVM(*Support Vector Machine*), *XGBoost* e ainda *LightGBM*. De um modo geral, as três primeiras técnicas são as que reúnem um consenso geral na comunidade científica, nomeadamente no que diz respeito à sua viabilidade na modelação de modelos preditivos. Por sua vez, o SVM é bastante eficiente e adequada, especialmente para a análise de dados com ruído.

Redes Neuronais Artificiais

As **redes neuronais artificiais** (Figura 2.3) consistem em modelos matemáticos usados na previsão, baseando-se no funcionamento e na estrutura das redes neuronais artificiais biológicas. Efetivamente, e de forma muito similar a estas últimas, as redes neuronais artificiais são constituídas por neurónios ou nodos, que remetem para uma componente devidamente capaz de aceitar diversos *inputs*, processando-os com o intuito de gerar um resultado, que tanto pode ser uma previsão do modelo, como um input para um novo neurónio. Todos os neurónios são agrupados por camadas (*layers*) e conectam-se através de um conjunto de ligações que determinam a influência entre cada neurónio. De salientar que a definição das camadas de neurónios e das ligações depende do tipo de rede neuronal que é, de facto, utilizada, podendo divergir consoante o modelo adotado. Entre os principais tipos de redes neuronais artificiais mais utilizados, merecem especial destaque os seguintes:

1. *Feedforward neural network* ([Auer et al., 2008](#));
2. *Kohonen self-organizing network* ([T. Kohonen, 2007](#));
3. *Radial basis function (RBF) network* ([Buhmann, 2003](#));
4. *Learning Vector Quantization* ([Kohonen, 1995](#)).

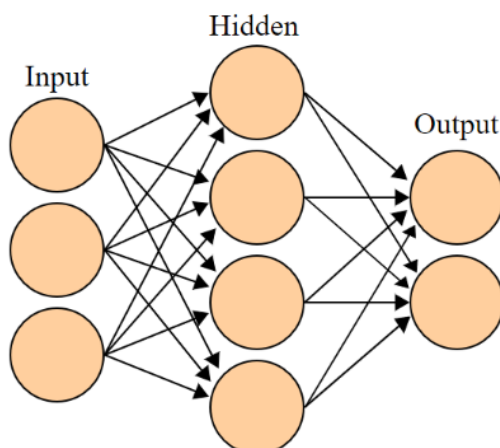


Figura 2.3: Diagrama resumido de uma rede neuronal (Lopes, 2018)

Árvores de decisão

As **árvores de decisão** são uma das técnicas de *Data Mining* mais utilizadas, devido essencialmente à sua simplicidade de implementação e à facilidade de compreensão dos resultados finais. De facto, esta técnica apresenta os resultados obtidos numa estrutura recursiva, onde se expressa a sequência do processo de classificação de um determinado item ou observação, essencialmente caracterizado por um conjunto de atributos que integra uma classe específica. Resumidamente, nesta técnica cada folha/nodo terminal representa uma classe em particular, enquanto cada nodo intermédio representa um teste específico, resultando numa subárvore. Assim, a classificação de um novo item ou observação é realizada através da navegação pela árvore, designadamente desde a sua raiz até uma folha, sendo que a folha representa a classe que classifica esse mesmo item (Quinlan, 1987).

Na Figura 2.4 é apresentado um exemplo bastante simplista de uma árvore de decisão, onde se modela a decisão de uma questão em particular, ainda que esta não remeta para um problema real: Se se deve jogar, ou não, um jogo de futebol num determinado dia, tendo em consideração dois atributos (Temperatura e Humidade) e duas classes (Jogar e Não Jogar).

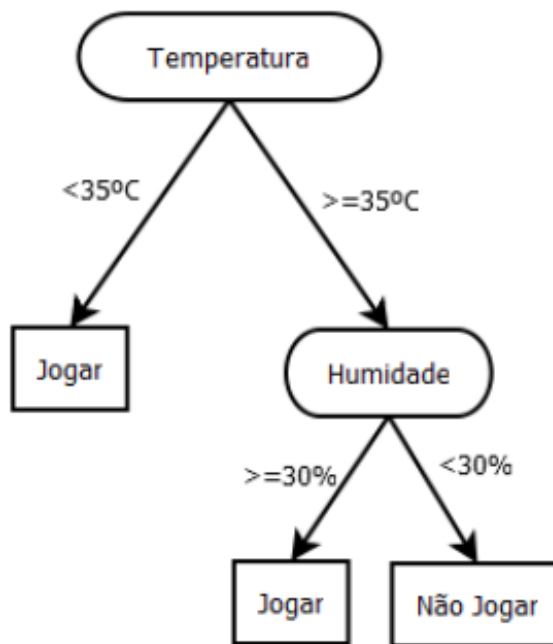


Figura 2.4: Árvore de decisão (Viana Gomes, 2011)

Basicamente, cada árvore de decisão é construída com base num conjunto de dados de treino, procedendo-se à divisão desse conjunto, e de acordo com o resultado obtido num determinado teste, a uma das suas variáveis. O processo em causa termina quando todas as observações de um nodo pertencem a uma mesma classe ou quando já não se obtém um aumento de valor nas previsões realizadas. É ainda importante acrescentar que as árvores de decisão podem ser de dois tipos: de classificação ou de regressão, dependendo do tipo de resultado que é esperado. Por conseguinte, a árvore de decisão é de classificação quando o objetivo é determinar a classe a que a observação pertence, e é de regressão quando o resultado remete para um número real (Loh, 2011).

Regressão logística

A **regressão logística**, por sua vez, consiste numa técnica estatística que é utilizada na previsão de ocorrência de um determinado evento, partindo de variáveis independentes como dados de entrada, e adaptando-as a uma função logística específica $f(z) = \frac{e^z}{e^z + 1} = \frac{1}{1 + e^{-z}}$

(representada na Figura 2.5), sendo que $f(z)$ representa a probabilidade de ocorrência de um determinado evento e z representa a contribuição de todas as variáveis independentes para essa mesma ocorrência. De referir que o valor de z é calculado através de todas as variáveis, e dos seus respectivos coeficientes de regressão, podendo ser usada a seguinte fórmula simplificada: $z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_k x_k$, sendo que β_0 consiste na interceção, representando o valor de z quando o valor das variáveis independentes é igual a zero. O x_1 e x_n representa todas as variáveis

independentes, enquanto o β_1 e β_n representa os seus respectivos coeficientes de regressão (Viana Gomes, 2011).

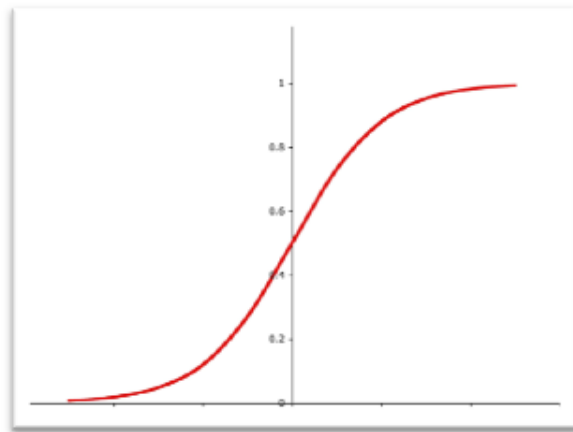


Figura 2.5: Função logística (Viana Gomes, 2011)

Essencialmente, esta função logística (Figura 2.5) pode receber qualquer valor como *input*, pois devolve sempre um valor enquanto resultado que corresponde a uma probabilidade que varia entre 0 e 1. No concernente à cardinalidade que a variável dependente pode assumir, é crucial referir que a regressão logística se divide em dois subtipos distintos: regressão logística binomial e regressão logística multinomial. A primeira é usada quando a variável de resposta/dependente assume apenas dois valores (1 ou 0), o que indica sucesso ou insucesso, respetivamente. A segunda, por sua vez, é usada quando a variável dependente remete para uma variável nominal, cujas categorias não podem ser ordenadas de um modo que faça sentido (Viana Gomes, 2011).

XGBoost

XGBoost ou *Extreme Gradient Boosting* é uma biblioteca baseado em árvores de decisão que implementa algoritmos *Machine Learning*, sob a estrutura de *Gradient Boosting*. Uma característica diferenciadora deste algoritmo é que os modelos são treinados tendo em conta o erro obtido em modelos anteriores (Guilherme, 2020). O *Gradient Boosting* não tenta prever diretamente o objectivo, ele faz uma previsão do erro de acordo com o erro residual obtido em modelos anteriores, assim ao longo da execução de diferentes iterações do mesmo processo vai sendo obtido menos erro. Os modelos que utilizam esta biblioteca têm uma abordagem mais iterativa, ou seja, em vez dos modelos serem treinados de forma isolada são treinados sucessivamente onde o novo modelo corrige os erros do modelo anterior (Oliveira, 2020). Esta abordagem é vantajosa, pois os modelos que vão sendo adicionados focam-se na correção do erro do anterior, ao contrário dos tradicionais métodos ensemble, onde os modelos são executados de forma isolada e, por isso, podem estar sempre a cometer o mesmo erro. Podemos observar o esquema na Figura 2.6

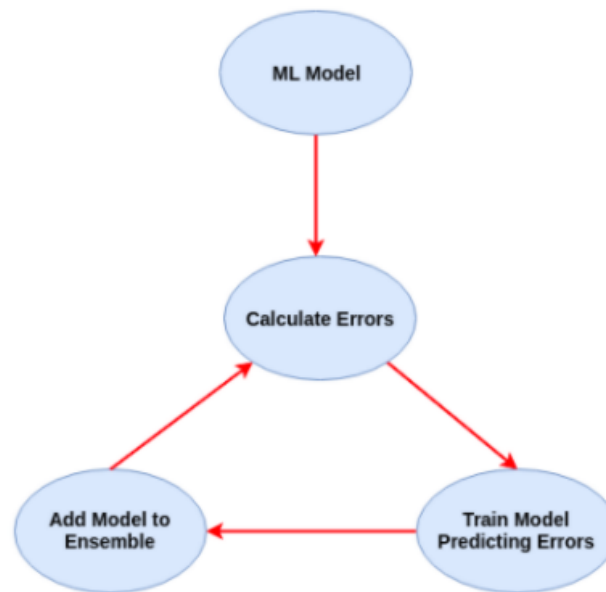


Figura 2.6: Esquema de um algoritmo baseado em Gradient Boosting. Figura adaptada de *A Beginner's guide to XGBoost*⁵

LightGBM

Tal como o *XGBoost*, o *LightGBM* é uma ferramenta que utiliza *gradient boosting* e os seus algoritmos são baseados em aprendizagem de árvore. O *LightGBM* possui as seguintes vantagens:

- Maior eficiência e menor velocidade de treino
- Consumo de memória reduzido
- Maior rigor
- Firme aprendizagem paralela e GPU
- Resiste a uma enorme quantidade de dados

As especificidades das otimizações são uma das diferenças entre *XGBoost* e *LightGBM*, o algoritmo *LightGBM* contém parâmetros que permite escolher como é realizado o crescimento da árvore. Na Figura 2.7 apresenta-se a diferença do crescimento em árvore de ambos os algoritmos.

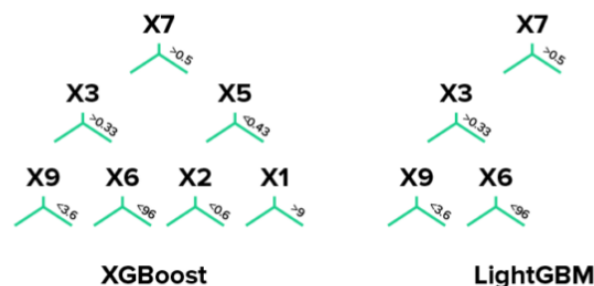


Figura 2.7: Crescimento de árvore XGBoost vs LightGBM. Figura adaptada de XGBoost, LightGBM or CatBoost — *which boosting algorithm should I use?* ⁶

Support Vector Machine

Finalmente, o **Support Vector Machine** (SVM) está associado a uma técnica estatística que pode ser utilizada em vários problemas de regressão e de classificação. Para além do mais, o SVM é um algoritmo supervisionado capaz de obter excelentes resultados, especialmente nos casos em que os dados a analisar apresentam uma elevada cardinalidade de classes independentes. Assim, esta técnica é bastante útil, sendo usada em várias áreas, tal como é o caso da área da otimização de software, da genética, da bioinformática e do reconhecimento de padrões (Corinna Cortes, 1995).

Apesar de esta técnica ter sido originalmente desenhada para classificações binárias, já existem algumas abordagens à implementação eficiente de SVM multi-classe. Ainda que a utilização de vários classificadores binários exija mais capacidade a nível de processamento, o uso de uma função que consiga discriminar diversas *labels* em simultâneo elimina a necessidade de se usarem vários classificadores, o que potencia, de igual modo, a utilização da dependência entre as várias *labels* em análise. Por conseguinte, Crammer e Singer (Crammer, 2002) criaram uma metodologia específica de SVM multi-classe, abordando o problema enquanto único problema de otimização, e não como vários problemas a nível da classificação binária.

Algoritmos Relacionais

Algoritmos relacionais são algoritmos que têm como *input* dados estruturados, como por exemplo grafos ou ficheiros XML, em vez de uma única tabela de dados. A seguir é apresentado o *Graph Mining*, utilizado no desenvolvimento da dissertação.

Graph Mining

Graph Mining é um conjunto de ferramentas e técnicas usadas para analisar as propriedades de grafos do mundo real. Os algoritmos de GM são uma família de algoritmos que tentam encontrar subestruturas comuns em um grafo. Como as moléculas podem ser representadas por grafos, o GM é frequentemente utilizado em cenários relacionados com a química, como por exemplo, encontrar subestruturas comuns de um determinado conjunto de medicamentos. Por esta razão,

os algoritmos de *GM* são relevantes para este projeto e entre eles, foi escolhido o *gSpan* (*graph-based Substructure pattern mining*) (Yan e Han, 2002) devido ao seu desempenho e viabilidade. O *gSpan*, dado um suporte mínimo e um tamanho mínimo (número de vértices) dos padrões, permite a mineração dos mesmos para um conjunto de grafos. Outro algoritmo eficiente é o *GASTON* (*GrAph, Sequences and Tree extractiON*) (Nijssen e Kok, 2004), que tenta aprimorar a eficiência baseada na eficiência já obtida em métodos anteriores, usando o "*quickstart principle*" que considera que as estruturas são partes umas das outras.

2.4.4 Ferramentas de *Data Mining*

O **Weka** (*Waikato Environment for Knowledge Analysis*) é uma ferramenta *open-source* de *Data Mining*, tendo sido desenvolvida em Java, e disponibiliza um vasto conjunto de algoritmos, bem como diversas ferramentas de pré-processamento de dados. Entre os algoritmos fornecidos pela ferramenta, é possível encontrar alguns de regressão, *clustering*, classificação e regras de associação. Além do mais, esta ferramenta também permite carregar dados através de diversos ficheiros, bases de dados ou urls, visto que suporta vários tipos de formatos (ARFF, CSV, C4.5 e LIBSVM), e exportar ou reutilizar vários modelos (Lopes, 2018).

Similarmente, o Weka também permite acrescentar diversos *plugins*, os quais possuem funcionalidades mais específicas e em diversas áreas, tal como é o caso do *plugin* BioWEKA, referente à área da bioinformática, biologia e bioquímica. Por fim, esta ferramenta também engloba algoritmos de avaliação, permitindo, portanto, a comparação entre algoritmos distintos ou, inclusive, entre diferentes conjuntos de dados (Gewehr, Szugat & Zimmer, 2007).

O **Orange** (<http://orange.biolab.si/>) é um *software* livre escrito em *Python*, sendo frequentemente utilizado para o *DM*. Essencialmente, este programa consiste numa ferramenta que disponibiliza uma interface muito intuitiva, permitindo a sua utilização por parte de todos os utilizadores, inclusive por parte daqueles que não têm qualquer tipo de experiência em *DM*. Apesar de ser bastante útil e intuitiva, esta ferramenta apenas suporta um tipo de base de dados, mais precisamente o *MySQL*. Porém, entre as suas principais funcionalidades destacam-se as seguintes:

1. Programação visual, escolhendo de forma inteligente os canais de comunicação e os *widgets* a usar;
2. Visualização, visto que conta com vários tipos de visualizações, desde gráficos a mapas sensoriais;
3. Cobre todas as tarefas de análise de dados mais importantes, para além de ser especializada em *add-ons*;
4. É uma ferramenta open source com uma comunidade de apoio bastante ativa;
5. É uma plataforma independente, podendo ser utilizada em vários sistemas operativos distintos (Gomes, 2014).

O **RStudio** (<https://www.rstudio.com/>), por sua vez, remete para uma ferramenta *open-source* IDE (*integrated development environment*) desenvolvida em C++, utilizando a linguagem R e o *framework* Qt. Este *software* está disponível em duas versões distintas:

1. RStudio Server, que permite o acesso ao RStudio através de um browser de internet;
2. RStudio Desktop, que pode ser utilizado localmente (Lopes, 2018)

O **KNIME** (*Kontanz Information Miner*) (<https://www.knime.org/>) consiste numa ferramenta que contempla um ambiente de desenvolvimento integrado (IDE) e um sistema de *plug-ins*, que tanto serve para o desenvolvimento como para o *DM*, tendo sido escrita em JAVA e baseada na plataforma Eclipse. No entanto, a versão do KNIME sem qualquer *plugin* engloba algoritmos de integração, visualização e transformação de dados (Gomes, 2014).

Finalmente, o **RapidMiner** (<https://rapidminer.com/>) também consiste numa ferramenta desenvolvida em Java que suporta todas as fases do processo de *DM*. Este software utiliza XML internamente com o intuito de uniformizar todos os seus dados, tendo em consideração que estes podem ser extraídos de inúmeras fontes, tal como é o caso do Access, Excel, IBM DB2, Oracle, ficheiros de texto e Microsoft SQL Server. Além disso, o RapidMiner também consegue executar todos os algoritmos executados pelo Weka, disponibilizando ainda muitos outros, o que diminui a necessidade de se escrever um código em específico. O resultado final é visualizado automaticamente através de diversos formatos, nomeadamente densidade, 3D e gráficos de barras (Lopes, 2018).

2.4.5 Métricas de avaliação

A avaliação do desempenho de um determinado classificador baseia-se na contagem dos exemplos de teste incorreta e corretamente previstos pelo classificador em questão. A matriz de confusão consiste num método que facilita a visualização do número de classificações corretas, bem como do número de classificação preditas para cada classe em específico, de um determinado aglomerado de exemplos. Por conseguinte, consiste num instrumento bastante útil para analisar a qualidade de qualquer classificador, designadamente na tarefa de reconhecimento dos exemplos de categorias distintas (Pei, 2011)

Regra geral, quando um conjunto de dados apenas detém duas categorias, uma é considerada como positiva e a outra como negativa. As entradas da matriz de confusão são frequentemente referidas como falsos negativos (*false negatives* – FN), verdadeiros negativos (*true negatives* – TN), verdadeiros positivos (*true positives* – TP) e falsos positivos (*false positives* – FP). Assim, com base nas entradas da matriz de confusão o número total de previsões corretas realizadas pelo classificador é TP+TN, enquanto o número total de previsões incorrectas é FP+FN (Bramer, 2007).

No entanto, este método não se limita a problemas de classificação com duas categorias, sendo também utilizado em problemas de classificação multi-classe. Neste caso, o resultado de um determinado conjunto de testes é apresentado enquanto matriz de confusão bidimensional, contendo uma coluna e uma linha para cada categoria. De referir que cada elemento da matriz demonstra

o número de exemplo de teste para a qual a categoria verdadeira corresponde à linha, enquanto a categoria predita corresponde à coluna. Além do mais, a performance de um determinado classificador também pode ser expressa através da sua taxa de erro (Witten et al., 2005)

- **Accuracy:** A *accuracy* é a métrica que diz quando o modelo acerta nas previsões. Por exemplo, se o modelo obter uma *accuracy* de 80% significa que em 10 acerta 8 previsões. Traduz-se na seguinte fórmula:

$$Accuracy = \frac{TP_c + TN_c}{TP_c + TN_c + FP_c + FN_c} \quad (2.1)$$

- **Precisão:** a precisão para a categoria c obtém-se através da expressão que se segue, a qual mede a percentagem das atribuições corretas entre todos os documentos que são atribuídos a esta categoria.

$$Precisão_c = \frac{TP_c}{TP_c + FP_c} \quad (2.2)$$

- **Sensibilidade (*recall*):** a sensibilidade para a categoria c é obtida através da expressão seguinte, a qual indica a percentagem de atribuições corretas em c e entre todos os documentos que foram atribuídos a esta categoria.

$$Recall_c = \frac{TP_c}{TP_c + FN_c} \quad (2.3)$$

Segundo Faceli et al. (Ferreira, 2012), a precisão mede a exatidão de um determinado classificador, enquanto a *recall* mede a sua completude. Ainda que a análise destas duas medidas, e em separado, não seja discutida, a média harmónica ponderada de ambas resulta na medida- $F\beta$:

- Medida- $F\beta$:

$$F\beta_c = \frac{(\beta + 1) * Precisão_c * Recall_c}{\beta * Precisão_c + Recall_c} \quad (2.4)$$

Porém, é frequente considerar-se que $\beta=1$, atribuindo-se, portanto, a mesma importância às medidas de precisão e de *recall*, motivo pelo qual a fórmula passa a ser a seguinte:

$$F_{1c} = \frac{2 * Precisão_c * Recall_c}{Precisão_c + Recall_c} \quad (2.5)$$

Para problemas de regressão, é usado um conjunto diferente de medidas.

- **A raiz quadrada média do erro (RMSE):** mede a distância entre os valores observados e previstos do modelo.

$$RMSE = \sqrt{\left(\frac{1}{n}\right) \sum_{i=1}^n (y_i - x_i)^2} \quad (2.6)$$

onde y_i : valores observados e x_i : valores previstos

- **O Coeficiente de Correlação de Pearson:**

$$R = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (2.7)$$

onde x_i : valores observados, y_i : valores previstos, $\bar{x}$: média dos valores observados, $\bar{y}$: média dos valores previstos

- **Coeficiente de Determinação:**

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - x_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (2.8)$$

onde y_i : valores observados x_i : valores previstos $\bar{y}$: média dos valores observados.

Ao contrário da raiz quadrada média do erro, o Coeficiente de Correlação de Pearson e o Coeficiente de Determinação são independentes da escala, onde 1 representa uma previsão perfeita e 0 significa que não há absolutamente nenhuma correlação linear.

2.5 Metodologias de Avaliação

As metodologias de avaliação são essencialmente 3, mais precisamente:

1. *Hold-Out*;
2. *k-fold Cross-Validation*;
3. *Leave-one-out* (Caso particular de CV em que um exemplo é retirado de cada vez).

A primeira metodologia consiste na divisão do conjunto de dados em dois subconjuntos diferentes: o de treino e o de teste, sendo que as proporções de cada são variáveis. Regra geral, o conjunto de treino é bastante superior ao conjunto de teste. A Figura 2.8 apresenta um exemplo do método *Hold-Out*, utilizando-se 70% dos exemplos para treino e 30% para teste (Namora, 2017).

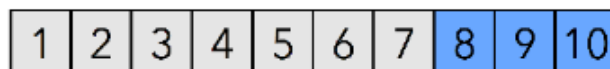


Figura 2.8: Exemplo do método *Hold-Out* (Namora, 2017)

No método *k-fold Cross-Validation* os dados são divididos em K subconjuntos, todos com tamanho igual. Cada modelo é construído através de um subconjunto, o qual representa os dados de teste, enquanto os restantes K-1 subconjuntos representam os dados de treino. A Figura 2.9 apresenta um exemplo deste método de avaliação (Namora, 2017).

Por fim, no método *Leave-one-out* utiliza-se apenas um dos seus exemplos para o teste e todos os outros exemplos enquanto treino. Por conseguinte, criam-se tantos modelos quantos

1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10

Figura 2.9: Exemplo do método *k-fold Cross-Validation* (Namora, 2017)

exemplos/dados existem. Regra geral, este método é utilizado quando os dados disponíveis são bastante limitados. A Figura 2.10 apresenta um exemplo deste método em particular (Namora, 2017).

1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10

Figura 2.10: Exemplo do método *Leave-one-out* (Namora, 2017)

2.6 Repositórios Web Relevantes

De um modo geral, existem vários repositórios web com bases de dados relativas à eficácia dos medicamentos, sendo que alguns são de livre acesso e outros englobam algum custo por parte do investigador. Os repositórios web que merecem especial destaque no presente trabalho são os seguintes:

1. ChemSpider;
2. PubChem;
3. ChEBI
4. DrugBank
5. KEGG
6. Chemoinformatics
7. PACT-F (*Pre-clinical And Clinical Trials knowledge base on bioavailability*);

O **ChemSpider** consiste numa repositório web gratuito que apresenta várias estruturas químicas abordadas em diversos estudos, contando com registos de mais de 81 milhões de estruturas químicas e com 278 fontes de dados (Che, 2020).

O **PubChem** (<https://pubchem.ncbi.nlm.nih.gov/>) foi criado no ano de 2004 e consiste, atualmente, num repositório público que contém informações acerca das substâncias químicas e suas atividades biológicas. A informação deste repositório é produzida por mais de 350 contribuidores, englobando agências governamentais, laboratórios de várias universidades e empresas farmacêuticas. Em 2015, o PubChem dispunha de mais de 157 milhões de descrições de substâncias químicas, 60 milhões de estruturas químicas únicas e cerca de 1 milhão de descrições de ensaios biológicos. De um modo geral, a informação deste repositório está organizada em três bases de dados que se interligam entre si, mais precisamente:

1. *Compound*, que contém estruturas químicas únicas extraídas da base de dados de substâncias (mais de 60 milhões);
2. *BioAssay*, que contém descrições dos ensaios biológicos acerca das substâncias químicas (mais de 1 milhão);
3. *Substance*, que contém descrições de substâncias (mais de 254 milhões) (Kim et al., 2016).

O **ChEBI** (*Chemical Entities of Biological Interest*) (<http://www.ebi.ac.uk/chebi/>) consiste num dicionário gratuito de entidades moleculares que se foca essencialmente em compostos químicos “pequenos”, apresentando também uma classificação ontológica onde todas as relações entre entidades moleculares e os seus pais e/ou filhos são especificados. O ChEBI resultou da combinação de várias fontes de informação, designadamente do IntEnz (Integrated relational Enzyme database), do KEGG COMPOUND e do Chemical Ontology (Degtyarenko et al., 2008).

A **DrugBank** (<https://www.drugbank.ca/>) remete para uma base de dados online, que para além de ser gratuita é bastante abrangente, visto que combina dados detalhados acerca de cada medicamento, bem como algumas das informações acerca da sua descrição, estrutura e do *pathway* em que atua. Nesta base de dados é possível procurar por categoria, medicamento, reação, gene, *pathway*, proteína alvo, indicações de uso ou classe (Lopes, 2018).

A **KEGG** (*Kyoto Encyclopedia of Genes and Genomes*) (<https://www.genome.jp/kegg/>) é uma base de dados que contém informação acerca dos *pathways* metabólicos, designadamente a partir de uma variedade de espécies. É das bases de dados mais completas e utilizadas, contendo mais de 15000 componentes e 7742 medicamentos (Kanehisa et al., 2016).

O **PACT-F** consiste numa base de dados baseada na estrutura, apresentando os resultados obtidos em diversos ensaios clínicos, tanto com testes em humanos, como em animais. Atualmente, este repositório remete para a maior base de dados, nomeadamente a nível da biodisponibilidade, pois contém cerca de 8296 registos de ensaios clínicos, os quais estão detalhadamente descritos e contém toda a estrutura química dos compostos. Para além do mais, o PACT-F contém todas as informações necessárias para o desenvolvimento de modelos preditivos, permitindo, portanto, o estudo da biodisponibilidade dos novos compostos e influenciando simultaneamente a otimização dos medicamentos em geral (PAC, 2020).

2.7 Trabalhos relacionados

Existem alguns estudos na área *Data Mining* para previsão da eficácia de medicamentos avaliados em linhas celulares. No entanto, é importante referenciar os trabalhos feitos anteriormente que tiveram como base o conjunto de dados utilizados nesta dissertação. Destaca-se então o trabalho realizado por (Menden et al., 2013) e por (Ladeiras, 2015). Nestes dois trabalhos foi realizado um conjunto de experiências usando o mesmo conjunto de dados e pré-processamento realizado nesta dissertação.

No (Menden et al., 2013), foram utilizados modelos de regressão, implementando os algoritmos *Artificial Neural Networks* e *Random Forest*, como base das experiências, sem ajuste de parâmetros. Foi realizado o método *8-fold cross-validation* ao conjunto de dados utilizado, um para a versão *Stringent* (sem linhas celulares repetidas nas dobras de treino e teste) outro para a versão *Non Stringent*. (Menden et al., 2013) obteve para o *Artificial Neural Networks* o valor quadrático médio (RMSE) de 0.83 na previsão dos valores IC50 no *test fold* e 0,84 com o algoritmo *Random Forest*. Para o *blind test* foi obtido o valor de 0.97.

Na dissertação do (Ladeiras, 2015) foi feita uma replicação das experiências do (Menden et al., 2013). Para além da replicação, foi utilizado o algoritmo *Support Vector Machine* e foram repetidas as experiências com ajuste de parâmetros para os algoritmos *Random Forest* e *Support Vector Machine*. O problema de regressão foi transformado em um problema de classificação e expandido com recurso à Programação em lógica indutiva (ILP). Por fim, foram adicionados ao conjunto de dados novas características obtidas pelo gSpan. Verificou-se que este processo não ampliou significativamente o desempenho. Também a transformação em problema de classificação e ILP não obtiveram um aumento significativo da performance.

Outros trabalhos que também mereceram atenção antes de iniciar este projeto foram *A Landscape of Pharmacogenomic Interactions in Cancer* realizado por (Iorio et al., 2016) e *A community effort to assess and improve drug sensitivity prediction algorithms* realizado por (Costello et al., 2014).

Capítulo 3

Metodologia

Neste capítulo, são descritos os métodos e a abordagem seguidos na fase experimental. Primeiramente, explicita-se o modo como está apresentado o conjunto de dados originais e o seu pré-processamento. Posteriormente, clarificam-se os métodos e os algoritmos aplicados ao problema de regressão e efetua-se o processo de *feature selection*. Após a análise ao problema de regressão, é feita a transformação para um problema de classificação. A experiência final descrita é a adição de características de medicamentos obtidas com *Graph Mining* ao conjunto de dados usado.

3.1 Descrição dos dados originais

O conjunto de dados originais usados nesta dissertação foi replicado do projeto original *Genomics of Drug Sensitivity in Cancer (GDSC)* (Garnett et al., 2012)¹. Neste projeto podem ser encontradas duas versões do conjunto de dados, a versão 1.0 e 1.1. A versão 1.0 contém 639 linhas celulares diferentes com as respectivas características:

- Instabilidade de microssatélite (MIS)
- Tecido correspondente
- Tipo de cancro
- 77 propriedades de mutação genética

O conjunto de dados 1.0 também inclui 131 medicamentos e os valores IC50 para as várias linhas celulares e pares de medicamentos, que se traduzem, então, em 83.709 potenciais valores de IC50. Para além dos dados de base do GDSC, é também necessário obter as propriedades dos 131 medicamentos. Estas propriedades são geradas através do software PaDEL-Descriptor e resultam na obtenção dos descritores moleculares e das *fingerprints* dos medicamentos. Nesta versão 1.0 dos dados, apenas estão disponíveis cerca de 58% dos valores IC50 para cada par célula-medicamento (Figura 3.1), estando os restantes pares indisponíveis. A versão 1.1 do conjunto de dados, que

¹Disponível em <https://www.cancerrxgene.org/>

incorpora cerca de 15,578 pares de célula-medicamento, foi disposto para criar um conjunto de *blind test* adicional. De seguida, apresenta-se o pré-processamento dos dados originais. O pré-processamento segue a mesma abordagem implementada por (Menden et al., 2013).

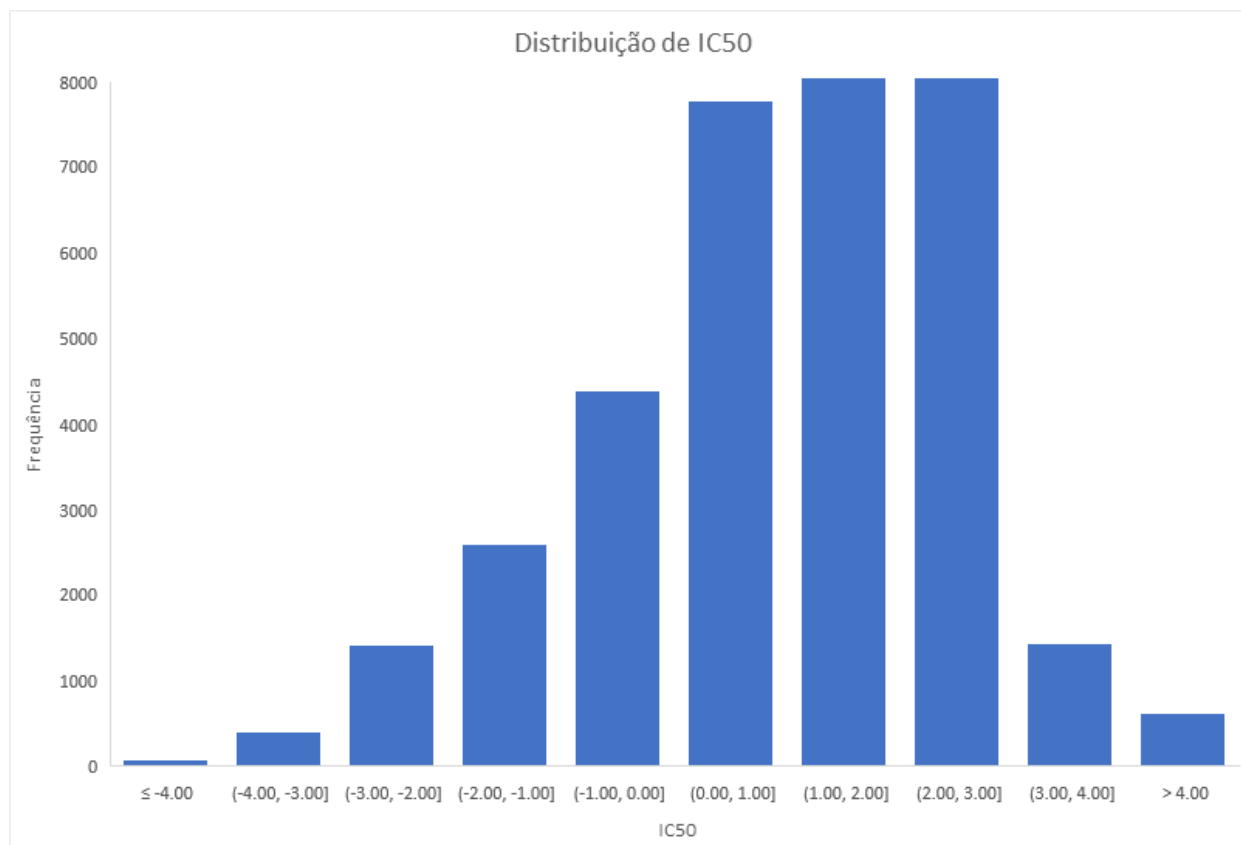


Figura 3.1: Distribuição dos valores IC50

3.2 Pré-Processamento dos dados originais

3.2.1 Características genómicas

Cada linha celular é caracterizada por 155 características genómicas - os dados de mutação genética consistem na instabilidade de microssatélite, a *sequence variation* e a *copy number variation*. Como primeiro passo de pré-processamento, as propriedades dos oncogenes foram transformadas em binárias e ternárias. De seguida explicamos como é apresentada a estrutura dos dados originais e o posterior pré-processamento.

- O valor de Instabilidade de microssatélite (MSI), já assim representado, é 1 se instável e 0 se estável.
- O fenótipo e o *copy number* são emparelhados em um objeto: "X :: Y" onde X é o fenótipo e Y o *copy number*.

- A *sequence variation* é "wt"(*wild type*), ou uma mutação do tipo (p*).
- O *copy number* é ≥ 8 (amplificação), "0" (exclusão) ou $0 < cn < 8$ (*wild type*)

Os valores do *copy number variation* e a *sequence variation* foram então processados do seguinte modo:

- O *copy number* é 0 se for *wild type*, 1 se for uma amplificação e -1 se for uma exclusão.
- A *sequence variation* é 0 se for *wild type* ou 1 se for uma mutação do tipo (p*).

As características genómicas com mais de 15 propriedades não especificadas foram removidas, assim como qualquer propriedade da *sequence variation* e do *copy number variation* indisponível para qualquer oncogene. Depois de remover todas as propriedades indisponíveis, a quantidade total de propriedades foi reduzida de 155 para 142 características genómicas.

3.2.2 Características dos medicamentos

Como descrito na secção da descrição dos dados originais, é então necessário obter as propriedades dos 131 medicamentos presentes. Estas propriedades são geradas através do software PaDEL-Descriptor. Este software necessita do *SMILES* de cada composto para obter as características. O *SMILES* foi então gerado com recurso à API disponível no site Pubchem², através do identificador Pubchem correspondente a cada medicamento. Obtendo o *SMILES*, foi possível, com o software *PaDEL-Descriptor*, gerar os descritores moleculares 1D e 2D e as *fingerprints* para cada medicamento. Alguns exemplos de descritores moleculares obtidos são o número de vezes que cada átomo é encontrado, o índice de Zagreb e para as *fingerprints* a existência ou não existência de ligação N-N. Saliente-se que, durante este processo, alguns descritores e *fingerprints* não conseguiram ser calculados pelo *PaDEL-Descriptor*, sendo assim eliminados. Foram também removidos do conjunto de dados todos os medicamentos para os quais as informações *SMILES* estavam indisponíveis. Os medicamentos ficaram, por conseguinte, reduzidos de 131 para 110 com 1603 características. Após serem removidas todas as características não disponíveis, bem como as que obtiveram o mesmo resultado para todos os medicamentos, o conjunto de dados ficou com 790 características de medicamentos.

Quanto aos valores de IC₅₀, como realizado em (Menden et al., 2013), foram convertidos de logaritmo natural para logaritmo de base 10, para fins de comparação. A variação dos valores de IC₅₀ convertidos é então [-8,22;6,63]. Após este pré-processamento, existem 142 propriedades genómicas e 790 propriedades de medicamentos, resultando em um total de 932 propriedades mais o valor IC₅₀. O conjunto de dados final resultou em 40691 instâncias para a versão 1.0 e 15578 adicionais para a versão 1.1. A lista completa de propriedades está disponível no apêndice A.

²Disponível em <http://pubchem.ncbi.nlm.nih.gov/>

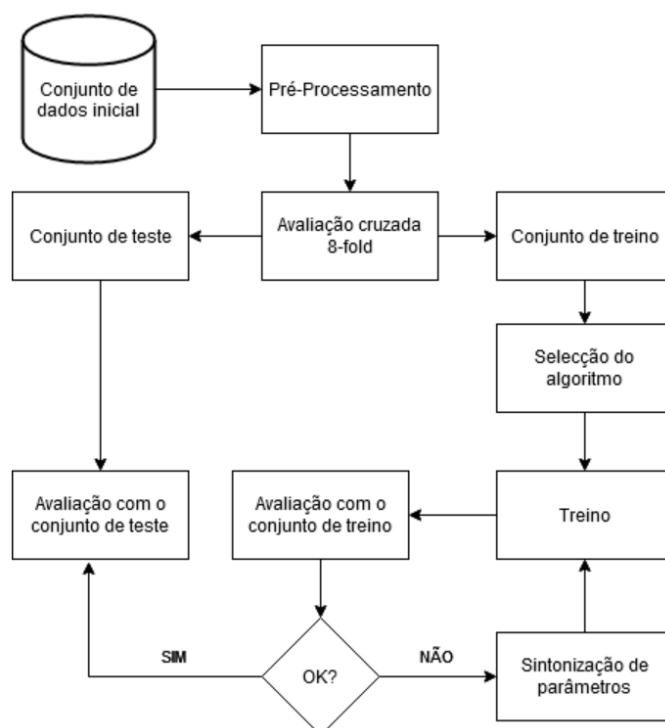


Figura 3.2: Visão Geral

3.3 Experiências

3.3.1 Implementação e avaliação dos modelos

Na nossa abordagem, numa primeira instância, foram construídos modelos para o problema de regressão, aplicando os algoritmos *Artificial Neural Networks*, *Random Forest*, *Support Vector Machine* e *XGBoost*, com ajuste de parâmetros. Uma das diferenças entre os trabalhos mencionados e esta dissertação foi a divisão do conjunto de dados originais em:

- Dados em que os medicamentos apenas apresentam os descritores moleculares 1D 2D
- Dados em que os medicamentos apresentam os descritores moleculares 1D 2D e as *fingerprints*.

Nos trabalhos relacionados mencionados anteriormente, os treinos e testes foram realizados no conjunto de dados em que os medicamentos apresentam os descritores moleculares 1D 2D e as *fingerprints*. Pretendemos com esta abordagem isolar assim os descritores moleculares e verificar a sua influência na performance dos modelos para a previsão dos valores IC50 para os pares linhas celulares-medicamentos.

Outra diferença foi a utilização da técnica de *feature selection* que permite encontrar e remover características redundantes no conjunto de dados. Este processo tem algumas vantagens,

nomeadamente reduzir o risco de *overfitting* do modelo de treino, uma melhor visualização e interpretação do conjunto de dados como também uma redução no tempo de treino do modelo. O conjunto de dados resultante deste processo também foi usado para construir modelos para o problema de regressão com os mesmos algoritmos já mencionados.

Para a avaliação dos modelos, foi realizado um *8-fold cross-validation*. Assim, os dados da versão 1.0 do conjunto de dados foi dividida em 8 compartimentos diferentes. Em concordância com (Menden et al., 2013), também duas versões para os conjunto de dados foram geradas:

- *Non-Stringent*: Divisão aleatória das instâncias em 8 compartimentos com tamanho igual
- *Stringent*: Divisão de modo que cada compartimento de treino e teste não compartilhem nenhuma linha celular.

A versão *Stringent* tem como objectivo mostrar o poder preditivo dos modelos. Os modelos foram então treinados e testados com o conjunto de dados *Stringent* e *Non Stringent*. Também fizemos um *blind test*, usando as instâncias adicionais da versão 1.1 do conjunto de dados como o conjunto de teste (com os extra 15578 exemplos). Para a avaliação de desempenho dos algoritmos presentes nesta secção foram medidos:

- Raíz Quadrada do Erro Médio Quadrático (*RMSE*)
- Coeficiente de correlação de *Pearson*
- Coeficiente de determinação.

Estas métricas são calculadas através dos valores IC50 reais e previstos de cada compartimento de teste. Os resultados apresentados no capítulo das experiências e resultados mostram a média e o desvio padrão para cada métrica.

3.4 Feature Selection

O *Feature Selection* (selecção de atributos) é um dos conceitos mais importantes em *Machine Learning* pois tem um enorme impacto no desempenho dos modelos. As características dos dados que são utilizados para treinar os modelos de *ML* têm uma grande influência no desempenho que pode ser alcançado nos modelos. Pode-se, deste modo, através da selecção de atributos, verificar quais as que fornecem o maior ganho entre as informações das linhas celulares e as características dos medicamentos, como também entre as suas subcategorias. Este processo pode, assim, permitir a remoção de características irrelevantes aos dados ou deduzir melhorias para os mesmos (G. James e Tibshirani, 2013).

Para efetuar este processo, foi utilizada a técnica de coeficiente de correlação, mais precisamente a de correlação de *Pearson*, para uma correlação linear. Correlação é um termo estatístico que se refere à possibilidade de duas variáveis terem uma relação linear entre si. Outra técnica

utilizada foi o ganho de informação mútua, que se refere ao ganho de informação de cada característica no contexto de uma variável alvo, neste caso o IC50. Através destas duas técnicas, verificaram-se quais as características que mais influenciam o valor alvo IC50 nos dados analisados. Seguidamente foi efetuado um corte drástico, onde foram eliminados um grande número de características que obtiveram baixa correlação e ganho de informação mútua. Foram assim gerados 4 novos conjuntos de dados: Conjunto de dados com apenas as 100 e 500 melhores características para a técnica correlação de Pearson e para a técnica ganho de informação mútua. Estes novos conjuntos de dados foram então usados para o problema de regressão e classificação.

3.5 Classificação

Após a análise de regressão, o problema foi transformado em um problema de classificação para explorar o desempenho num diferente tipo de modelo. Para efetuar esta transformação, converteram-se os atributos de dados contínuos dos valores IC50 em um conjunto finito de intervalos e associou-se a cada intervalo um valor específico. Este processo é designado por discretização de dados.

Para o conjunto finito de intervalos, efetuou-se o cálculo dos tercís para a distribuição dos IC50, que foram usados como valores de limite superior e inferior na decisão da classe. Optou-se, então, por dividir os dados em classe "Bom" abaixo do limite inferior e "Mau" acima do limite superior. Os restantes elementos foram eliminados.

A discretização de dados resultou, pois, num total de 27118 exemplos para a versão 1.0 e 10427 para a versão 1.1 dos dados.

Para a avaliação, tal como na regressão, estes dados de classificação foram então divididos em dois conjuntos de dados, os que continham apenas os descritores moleculares 1D e 2D dos medicamentos e outro conjunto com os que continham tanto os descritores moleculares como as *fingerprints*. Para a construção dos modelos de classificação implementamos os mesmos algoritmos utilizados no problema de regressão:

- *Random Forest*
- *SVM*
- *XGBoost*
- *Artificial Neural Networks*

Também se procedeu à divisão entre dados *Stringent* e *Non Stringent* para cada um dos conjuntos de dados. Para o problema de classificação foi realizada a *8-fold cross-validation*. Para a avaliação de desempenho dos algoritmos presentes nesta secção mediram-se a *accuracy*, *precision*, *recall* e o *f1-score* através da obtenção das matrizes de confusão.

3.6 *Graph Mining*

Com o auxílio do Graph Mining foi possível encontrar subestruturas comuns aos medicamentos utilizados no conjunto de dados. Estas subestruturas podem ajudar os especialistas a perceber o efeito dos medicamentos nas linhas celulares. As subestruturas encontradas foram posteriormente adicionadas às restantes características dos medicamentos presentes no conjunto de dados.

Primeiramente, para encontrar as subestruturas, os medicamentos foram representados como grafos e de seguida foi aplicado um algoritmo de *Graph Mining* aos dados dos grafos. Estas subestruturas têm algumas semelhanças com a estrutura das *fingerprints*, sendo o objectivo deste processo perceber se, adicionando ao conjunto de dados as subestruturas encontradas, estas poderão ser mais importantes que as *fingerprints* como também avaliar o desempenho da abordagem em geral.

O algoritmo seleccionado para este projecto foi o *gSpan* (*graph-based Substructure pattern mining*) devido ao seu desempenho e viabilidade. As subestruturas foram calculadas usando implementação *gspan-mining* 0.2.3 do Python. Os *inputs* para a execução do *gSpan* foram os dados dos grafos das moléculas dos medicamentos, gerados através das informações do ficheiro SMILES. Estes dados eram apresentados em um ficheiro SDF, tendo estes sido convertidos para o formato *gSpan* através do Prolog. O número de suporte no algoritmo *gSpan* variou entre 20 e 100. Este valor representa o número mínimo de vezes que um subgrafo aparece em um conjunto de grafos. O tamanho mínimo dos fragmentos variou entre 3 e 20 vértices. O número de fragmentos encontrado foi limitado entre 200 e 500. Os resultados do *gSpan* foram posteriormente convertidos em colunas de grafos no conjunto de dados originais onde 1 indica a presença do subgrafo no medicamento e 0 que não está presente. O novo conjunto de dados gerados foi então usado para construir modelos para o problema de regressão com o algoritmo *XGBoost* e *Random Forest*.

Capítulo 4

Experiências e Resultados

Neste capítulo são apresentadas todas as experiências descritas no capítulo anterior e também se apresenta e discute os seus respectivos resultados.

4.1 Conjunto de dados completos

4.1.1 Regressão

A seguir são apresentados os resultados para o problema de regressão usando o conjunto de dados completo. Para os resultados foi seleccionado o modelo com o conjunto de parâmetros que obtiveram o menor erro. Para os algoritmos utilizados foi usada a biblioteca de Python "*Sklearn*".

4.1.1.1 *Random Forest*

A Tabela 4.1 apresenta o ajuste de parâmetros para o algoritmo *Random Forest* :

Atributos	Valores
<i>n_estimators</i>	128,256,512
<i>max_features</i>	auto, sqrt, log2

Tabela 4.1: Parâmetros usados no algoritmo *Random Forest*

Descritores Moleculares 1D e 2D

São apresentados na Tabela 4.2 os resultados para o conjunto de dados que contém os descritores moleculares 1D e 2D dos medicamentos.

	RMSE	RP	R2
<i>Non Stringent</i>	0.85 +/- 0.028	0.84 +/- 0.010	0.70 +/- 0.017
<i>Stringent</i>	0.89 +/- 0.029	0.82 +/- 0.011	0.67 +/- 0.018

Tabela 4.2: Resultados com descritores moleculares para o *Random Forest*

Descritores Moleculares 1D e 2D e *fingerprints*

São apresentados na Tabela 4.3 os resultados para o conjunto de dados que contém os descritores moleculares 1D e 2D e as *fingerprints* dos medicamentos.

	RMSE	RP	R2
<i>Non Stringent</i>	0.85 +/- 0.029	0.84 +/- 0.011	0.70 +/- 0.018
<i>Stringent</i>	0.88 +/- 0.033	0.82 +/- 0.011	0.68 +/- 0.018

Tabela 4.3: Resultados com descritores moleculares e *fingerprints* para o *Random Forest*

Blind Test

Descritores Moleculares 1D e 2D

São apresentados na Tabela 4.4 os resultados para o conjunto de dados que contém os descritores moleculares 1D e 2D.

	RMSE	RP	R2
<i>Non Stringent Blind Test</i>	0.95 +/- 0.003	0.79 +/- 0.001	0.62 +/- 0.002
<i>Stringent Blind Test</i>	0.86 +/- 0.004	0.83 +/- 0.002	0.69 +/- 0.002

Tabela 4.4: Resultados do *blind test* com descritores moleculares para o *Random Forest*

Descritores Moleculares 1D e 2D e *fingerprints*

São apresentados na Tabela 4.5 os resultados para o conjunto de dados que contém os descritores moleculares 1D e 2D e as *fingerprints* dos medicamentos.

	RMSE	RP	R2
<i>Non Stringent Blind Test</i>	0.95 +/- 0.003	0.79 +/- 0.001	0.62 +/- 0.002
<i>Stringent Blind Test</i>	0.88 +/- 0.003	0.83 +/- 0.001	0.69 +/- 0.002

Tabela 4.5: Resultados do *blind test* com descritores moleculares e *fingerprints* para o *Random Forest*

4.1.1.2 *Support Vector Machines*

A Tabela 4.6 mostra o ajuste de parâmetros para o algoritmo *SVM*:

Atributos	Valores
<i>Kernel</i>	rbf
<i>gamma</i>	1e-3, 1e-4
<i>C</i>	1, 2

Tabela 4.6: Parâmetros no algoritmo *SVM*

Descritores Moleculares 1D e 2D

São apresentados na Tabela 4.7 os resultados para o conjunto de dados que contém os descritores moleculares 1D e 2D dos medicamentos.

	RMSE	RP	R2
<i>Non Stringent</i>	0.94 +/- 0.040	0.80 +/- 0.010	0.64 +/- 0.017
<i>Stringent</i>	0.96 +/- 0.034	0.79 +/- 0.010	0.62 +/- 0.015

Tabela 4.7: Resultados com descritores moleculares para o SVM

Descritores Moleculares 1D e 2D e fingerprints

São apresentados na Tabela 4.8 os resultados para o conjunto de dados que contém os descritores moleculares 1D e 2D e as *fingerprints* dos medicamentos.

	RMSE	RP	R2
<i>Non Stringent</i>	0.84 +/- 0.033	0.84 +/- 0.010	0.71 +/- 0.017
<i>Stringent</i>	0.85 +/- 0.037	0.83 +/- 0.012	0.70 +/- 0.019

Tabela 4.8: Resultados com descritores moleculares e *fingerprints* para o SVM

Blind Test**Descritores Moleculares 1D e 2D**

São apresentados na Tabela 4.9 os resultados para o conjunto de dados que contém os descritores moleculares 1D e 2D.

	RMSE	RP	R2
<i>Non Stringent Blind Test</i>	0.98 +/- 0.002	0.77 +/- 0.001	0.59 +/- 0.002
<i>Stringent Blind Test</i>	0.94 +/- 0.002	0.80 +/- 0.001	0.63 +/- 0.002

Tabela 4.9: Resultados do *blind test* com descritores moleculares para o SVM

Descritores Moleculares 1D e 2D e fingerprints

São apresentados na Tabela 4.10 os resultados para o conjunto de dados que contém os descritores moleculares 1D e 2D e as *fingerprints* dos medicamentos.

	RMSE	RP	R2
<i>Non Stringent Blind Test</i>	0.88 +/- 0.002	0.82 +/- 0.001	0.67 +/- 0.001
<i>Stringent Blind Test</i>	0.82 +/- 0.002	0.84 +/- 0.001	0.70 +/- 0.001

Tabela 4.10: Resultados do *blind test* com descritores moleculares e *fingerprints* para o SVM

4.1.1.3 XGBoost

A Tabela 4.11 apresenta o ajuste de parâmetros para o algoritmo *XGBoost*:

Atributos	Valores
<i>n_estimators</i>	50, 100, 200
<i>min_child_weight</i>	1, 5, 10
<i>gamma</i>	0.5, 3

Tabela 4.11: Parâmetros no algoritmo *XGBoost*

Descritores Moleculares 1D e 2D

São apresentados na Tabela 4.12 os resultados para o conjunto de dados que contém os descritores moleculares 1D e 2D dos medicamentos.

	RMSE	RP	R2
<i>Non Stringent</i>	0.83 +/- 0.035	0.85 +/- 0.010	0.71 +/- 0.017
<i>Stringent</i>	0.86 +/- 0.032	0.83 +/- 0.011	0.69 +/- 0.019

Tabela 4.12: Resultados com descritores moleculares para o *XGBoost*

Descritores Moleculares 1D e 2D e *fingerprints*

São apresentados na Tabela 4.13 os resultados para o conjunto de dados que contém os descritores moleculares 1D e 2D e as *fingerprints* dos medicamentos.

	RMSE	RP	R2
<i>Non Stringent</i>	0.83 +/- 0.035	0.84 +/- 0.010	0.71 +/- 0.017
<i>Stringent</i>	0.84 +/- 0.032	0.84 +/- 0.009	0.70 +/- 0.015

Tabela 4.13: Resultados com descritores moleculares e *fingerprints* para o *XGBoost*

Blind Test

Descritores Moleculares 1D e 2D

São apresentados na Tabela 4.14 os resultados para o conjunto de dados que contém os descritores moleculares 1D e 2D.

	RMSE	RP	R2
<i>Non Stringent Blind Test</i>	0.92 +/- 0.008	0.80 +/- 0.003	0.64 +/- 0.005
<i>Stringent Blind Test</i>	0.84 +/- 0.005	0.84 +/- 0.002	0.71 +/- 0.004

Tabela 4.14: Resultados do *blind test* com descritores moleculares para o *XGBoost*

Descritores Moleculares 1D e 2D e fingerprints

São apresentados na Tabela 4.15 os resultados para o conjunto de dados que contém os descritores moleculares 1D e 2D e as *fingerprints* dos medicamentos.

	RMSE	RP	R2
<i>Non Stringent Blind Test</i>	0.92 +/- 0.003	0.80 +/- 0.001	0.64 +/- 0.002
<i>Stringent Blind Test</i>	0.86 +/- 0.005	0.84 +/- 0.002	0.70 +/- 0.004

Tabela 4.15: Resultados do *blind test* com descritores moleculares e *fingerprints* para o *XGBoost*

4.1.1.4 ANN

A tabela 4.16 mostra o ajuste de parâmetros para o algoritmo *ANN* :

Atributos	Valores
<i>activation</i>	relu, linear
<i>neurons</i>	5,8,10,20
<i>init_mode</i>	uniform, normal
<i>optimizer</i>	adam
<i>hidden_layers</i>	5
<i>Epochs</i>	50,80,200
<i>Batch_size</i>	500,1000,1500

Tabela 4.16: Parâmetros utilizados no algoritmo *ANN*

Descritores Moleculares 1D e 2D

São apresentados na Tabela 4.17 os resultados para o conjunto de dados que contém os descritores moleculares 1D e 2D dos medicamentos.

	RMSE	RP	R2
<i>Non Stringent</i>	0.87 +/- 0.037	0.83 +/- 0.012	0.69 +/- 0.020
<i>Stringent</i>	0.90 +/- 0.047	0.82 +/- 0.019	0.67 +/- 0.031

Tabela 4.17: Resultados com descritores moleculares para o *ANN*

Descritores Moleculares 1D e 2D e fingerprints

São apresentados na Tabela 4.18 os resultados para o conjunto de dados que contém os descritores moleculares 1D e 2D e as *fingerprints* dos medicamentos.

	RMSE	RP	R2
<i>Non Stringent</i>	0.85 +/- 0.040	0.84 +/- 0.013	0.70 +/- 0.021
<i>Stringent</i>	0.87 +/- 0.047	0.83 +/- 0.015	0.68 +/- 0.024

Tabela 4.18: Resultados com descritores moleculares e *fingerprints* para o ANN

Blind Test

Descritores Moleculares 1D e 2D

São apresentados na Tabela 4.19 os resultados para o conjunto de dados que contém os descritores moleculares 1D e 2D.

	RMSE	RP	R2
<i>Non Stringent Blind Test</i>	0.93 +/- 0.021	0.79 +/- 0.009	0.63 +/- 0.015
<i>Stringent Blind Test</i>	0.87 +/- 0.033	0.83 +/- 0.014	0.69 +/- 0.022

Tabela 4.19: Resultados do *blind test* com descritores moleculares para o ANN

Descritores Moleculares 1D e 2D e *fingerprints*

São apresentados na Tabela 4.20 os resultados para o conjunto de dados que contém os descritores moleculares 1D e 2D e as *fingerprints* dos medicamentos.

	RMSE	RP	R2
<i>Non Stringent Blind Test</i>	0.91 +/- 0.017	0.81 +/- 0.007	0.65 +/- 0.011
<i>Stringent Blind Test</i>	0.88 +/- 0.008	0.83 +/- 0.003	0.69 +/- 0.006

Tabela 4.20: Resultados do *blind test* com descritores moleculares e *fingerprints* para o ANN

4.1.2 Classificação

Para a avaliação de desempenho dos algoritmos no problema de classificação calculamos a accuracy, precision, recall e o F1-Score através da obtenção das matrizes de confusão.

4.1.2.1 Random Forest

Descritores Moleculares 1D e 2D

Accuracy	0.8910
Precision	0.8875
Recall	0.8954
F1-Score	0.8914

Tabela 4.21: Resultados RF
1D2D *Non Stringent*

Accuracy	0.8770
Precision	0.8745
Recall	0.8813
F1-Score	0.8779

Tabela 4.22: Resultados RF
1D2D *Stringent*

Descritores Moleculares 1D e 2D e fingerprints

Accuracy	0.8923
Precision	0.8905
Recall	0.8945
F1-Score	0.8925

Tabela 4.23: Resultados RF
1D2DFP *Non Stringent*

Accuracy	0.8850
Precision	0.8845
Recall	0.8824
F1-Score	0.8834

Tabela 4.24: Resultados RF
1D2DFP *Stringent***Blind test****Descritores Moleculares 1D e 2D**

Accuracy	0.8645
Precision	0.8519
Recall	0.8808
F1-Score	0.8660

Tabela 4.25: Resultados *Blind Test*
RF 1D2D *Non Stringent*

Accuracy	0.8863
Precision	0.8804
Recall	0.8897
F1-Score	0.8852

Tabela 4.26: Resultados *Blind Test*
RF 1D2D *Stringent***Descritores Moleculares 1D e 2D e fingerprints**

Accuracy	0.8639
Precision	0.8517
Recall	0.8794
F1-Score	0.8653

Tabela 4.27: Resultados *Blind Test*
RF 1D2DFP *Non Stringent*

Accuracy	0.8843
Precision	0.9012
Recall	0.8697
F1-Score	0.8851

Tabela 4.28: Resultados *Blind Test*
RF 1D2DFP *Stringent***4.1.2.2 Support Vector Machine****Descritores Moleculares 1D e 2D**

Accuracy	0.8436
Precision	0.8979
Recall	0.7758
F1-Score	0.8310

Tabela 4.29: Resultados SVM
1D2D *Non Stringent*

Accuracy	0.8364
Precision	0.9001
Recall	0.7583
F1-Score	0.8217

Tabela 4.30: Resultados SVM
1D2D *Stringent*

Descritores Moleculares 1D e 2D e fingerprints

Accuracy	0.8956
Precision	0.8874
Recall	0.9062
F1-Score	0.8965

Tabela 4.31: Resultados SVM
1D2DFP *Non Stringent*

Accuracy	0.8911
Precision	0.8861
Recall	0.8948
F1-Score	0.8901

Tabela 4.32: Resultados SVM
1D2DFP *Stringent*

Blind test**Descritores Moleculares 1D e 2D**

Accuracy	0.8357
Precision	0.8869
Recall	0.7687
F1-Score	0.8214

Tabela 4.33: Resultados *Blind Test*
SVM 1D2D *Non Stringent*

Accuracy	0.8420
Precision	0.8974
Recall	0.7670
F1-Score	0.8259

Tabela 4.34: Resultados *Blind Test*
SVM 1D2D *Stringent*

Descritores Moleculares 1D e 2D e fingerprints

Accuracy	0.8860
Precision	0.8737
Recall	0.9013
F1-Score	0.8871

Tabela 4.35: Resultados *Blind Test*
SVM 1D2DFP *Non Stringent*

Accuracy	0.8939
Precision	0.9048
Recall	0.8868
F1-Score	0.8954

Tabela 4.36: Resultados *Blind Test*
SVM 1D2DFP *Stringent*

4.1.2.3 XGBoost**Descritores Moleculares 1D e 2D**

Accuracy	0.9027
Precision	0.9026
Recall	0.9026
F1-Score	0.9026

Tabela 4.37: Resultados XGBoost
1D2D *Non Stringent*

Accuracy	0.8950
Precision	0.8944
Recall	0.8969
F1-Score	0.8956

Tabela 4.38: Resultados XGBoost
1D2D *Stringent*

Descritores Moleculares 1D e 2D e fingerprints

Accuracy	0.9059
Precision	0.9064
Recall	0.9052
F1-Score	0.9058

Tabela 4.39: Resultados XGBoost
1D2DFP *Non Stringent*

Accuracy	0.8974
Precision	0.9023
Recall	0.8883
F1-Score	0.8953

Tabela 4.40: Resultados XGBoost
1D2DFP *Stringent***Blind test****Descritores Moleculares 1D e 2D**

Accuracy	0.8721
Precision	0.8821
Recall	0.8771
F1-Score	0.8770

Tabela 4.41: Resultados *Blind Test*
XGBoost 1D2D *Non Stringent*

Accuracy	0.8979
Precision	0.8943
Recall	0.8991
F1-Score	0.8967

Tabela 4.42: Resultados *Blind Test*
XGBoost 1D2D *Stringent***Descritores Moleculares 1D e 2D e fingerprints**

Accuracy	0.8756
Precision	0.8720
Recall	0.8790
F1-Score	0.8755

Tabela 4.43: Resultados *Blind Test*
XGBoost 1D2DFP *Non Stringent*

Accuracy	0.8936
Precision	0.9140
Recall	0.8747
F1-Score	0.8939

Tabela 4.44: Resultados *Blind Test*
XGBoost 1D2DFP *Stringent***4.1.2.4 ANN****Descritores Moleculares 1D e 2D**

Accuracy	0.8866
Precision	0.8878
Recall	0.8854
F1-Score	0.8847

Tabela 4.45: Resultados ANN
1D2D *Non Stringent*

Accuracy	0.8789
Precision	0.8804
Recall	0.8783
F1-Score	0.8792

Tabela 4.46: Resultados ANN
1D2D *Stringent*

Descritores Moleculares 1D e 2D e fingerprints

Accuracy	0.8968
Precision	0.8968
Recall	0.8969
F1-Score	0.8968

Tabela 4.47: Resultados ANN
1D2DFP *Non Stringent*

Accuracy	0.8902
Precision	0.8982
Recall	0.8770
F1-Score	0.8874

Tabela 4.48: Resultados ANN
1D2DFP *Stringent*

Blind test**Descritores Moleculares 1D e 2D**

Accuracy	0.8629
Precision	0.8586
Recall	0.8676
F1-Score	0.8629

Tabela 4.49: Resultados *Blind Test*
ANN 1D2D *Non Stringent*

Accuracy	0.8752
Precision	0.8774
Recall	0.8686
F1-Score	0.8727

Tabela 4.50: Resultados *Blind Test*
ANN 1D2D *Stringent*

Descritores Moleculares 1D e 2D e fingerprints

Accuracy	0.8705
Precision	0.8653
Recall	0.8761
F1-Score	0.8706

Tabela 4.51: Resultados *Blind Test*
ANN 1D2DFP *Non Stringent*

Accuracy	0.8821
Precision	0.9055
Recall	0.8598
F1-Score	0.8820

Tabela 4.52: Resultados *Blind Test*
ANN 1D2DFP *Stringent*

4.2 Feature Selection

Para efetuar o processo de *feature selection* foi utilizada a técnica de coeficiente de correlação, mais precisamente a correlação de Pearson. Os 10 melhores resultados para o problema de regressão podem ser observados na tabela 4.53.

Atributo	Correlação
PubchemFP581	-0.5067
PubchemFP642	-0.5067
SPC-6	-0.4785
MDEC-34	-0.4729
SPC-5	-0.4692
ETA_EtaP	-0.4647
SPC-4	-0.4528
SC-5	-0.4446
ATSp5	-0.4434
WPOL	-0.4411

Tabela 4.53: Correlação de Pearson

Atributo	Ganho
ETA_EtaP_B	0.7189
MLFER_E	0.7116
ETA_Eta_F_L	0.7086
ETA_dBetaP	0.7071
SPC-4	0.7049
ETA_EtaP_L	0.6979
ETA_Shape_P	0.6974
BCUTp-1h	0.6961
SP-1	0.6959
ETA_Eta_R_L	0.6958

Tabela 4.54: Ganho de informação mútua para regressão

Outra técnica utilizada foi o ganho de informação mútua, que se refere ao ganho de informação de cada característica no contexto de uma variável alvo, neste caso o IC50. Os 10 melhores resultados para o problema de regressão podem ser observados na Tabela 4.54.

Para o problema de classificação, os 10 melhores resultados para a técnica do ganho de informação mútua podem ser observados na Tabela 4.55. Os resultados completos para estas técnicas são apresentados no Anexo B.

Atributo	Ganho
ETA_Eta	0.4678
VABC	0.4673
ATSc2	0.4673
ETA_dEpsilon_A	0.4673
ATSp2	0.4672
ETA_Epsilon_4	0.4672
ETA_EtaP_L	0.4671
SP-4	0.4670
VP-5	0.4670
SP-2	0.4669

Tabela 4.55: Ganho de informação mútua para classificação

Foi verificado através dos resultados obtidos que os descritores moleculares 1D e 2D obtiveram um desempenho consistente tendo a maior parte obtido valores relevantes. As *fingerprints* e as características genômicas tiveram um desempenho inferior comparativamente aos descritores moleculares sendo apenas uma pequena percentagem considerada relevante. Existe uma quantidade significativa de características de baixa qualidade que apresentam apenas ruído. Estas foram eliminadas do conjunto dados aquando da criação dos 4 novos conjunto de dados: Os Top 100 e Top 500 resultados para a correlação de Pearson e os Top 100 e Top 500 resultados para o ganho de informação mútua.

A seguir são apresentados os resultados para o problema de regressão e classificação para os 4 novos conjunto de dados otimizados pelo *feature selection*. Para os algoritmos utilizados foi usada a biblioteca de Python "Sklearn" e o mesmo ajuste de parâmetros realizado no conjunto de dados completo.

4.2.1 Ganho de informação mútua - Top 100

4.2.1.1 Regressão

Descritores moleculares 1D e 2D

1D 2D		Non Stringent	Stringent	Non Stringent Blind Test	Stringent Blind Test
RF	RMSE	0.83 +/- 0.030	0.84 +/- 0.032	0.86 +/- 0.001	0.82 +/- 0.001
	RP	0.85 +/- 0.011	0.84 +/- 0.011	0.83 +/- 0.001	0.85 +/- 0.001
	R2	0.72 +/- 0.018	0.70 +/- 0.012	0.68 +/- 0.001	0.72 +/- 0.001
SVM	RMSE	1.17 +/- 0.044	1.17 +/- 0.029	1.19 +/- 0.002	1.16 +/- 0.001
	RP	0.67 +/- 0.011	0.66 +/- 0.009	0.63 +/- 0.001	0.67 +/- 0.001
	R2	0.44 +/- 0.014	0.43 +/- 0.011	0.40 +/- 0.001	0.44 +/- 0.001
XGBoost	RMSE	0.83 +/- 0.031	0.84 +/- 0.032	0.86 +/- 0.001	0.82 +/- 0.001
	RP	0.85 +/- 0.011	0.84 +/- 0.111	0.83 +/- 0.001	0.85 +/- 0.001
	R2	0.72 +/- 0.019	0.70 +/- 0.018	0.68 +/- 0.001	0.72 +/- 0.001
ANN	RMSE	0.83 +/- 0.029	0.84 +/- 0.031	0.86 +/- 0.001	0.82 +/- 0.002
	RP	0.85 +/- 0.011	0.84 +/- 0.010	0.83 +/- 0.001	0.85 +/- 0.001
	R2	0.72 +/- 0.018	0.70 +/- 0.017	0.68 +/- 0.001	0.72 +/- 0.001

Tabela 4.56: Resultados da regressão para os descritores moleculares 1D e 2D - GIM100

Descritores moleculares 1D e 2D e fingerprints

1D 2D FP		Non Stringent	Stringent	Non Stringent Blind Test	Stringent Blind Test
RF	RMSE	0.83 +/- 0.031	0.83 +/- 0.034	0.86 +/- 0.001	0.84 +/- 0.001
	RP	0.85 +/- 0.011	0.84 +/- 0.007	0.83 +/- 0.001	0.84 +/- 0.001
	R2	0.72 +/- 0.018	0.71 +/- 0.013	0.68 +/- 0.001	0.71 +/- 0.001
SVM	RMSE	1.17 +/- 0.044	1.16 +/- 0.038	1.19 +/- 0.002	1.19 +/- 0.001
	RP	0.67 +/- 0.011	0.66 +/- 0.007	0.63 +/- 0.001	0.66 +/- 0.001
	R2	0.44 +/- 0.015	0.43 +/- 0.010	0.40 +/- 0.001	0.44 +/- 0.001
XGBoost	RMSE	0.83 +/- 0.031	0.83 +/- 0.034	0.86 +/- 0.001	0.84 +/- 0.001
	RP	0.85 +/- 0.011	0.84 +/- 0.007	0.83 +/- 0.001	0.84 +/- 0.001
	R2	0.72 +/- 0.019	0.71 +/- 0.012	0.68 +/- 0.001	0.71 +/- 0.001
ANN	RMSE	0.83 +/- 0.029	0.83 +/- 0.032	0.86 +/- 0.002	0.84 +/- 0.002
	RP	0.85 +/- 0.011	0.84 +/- 0.007	0.82 +/- 0.001	0.84 +/- 0.001
	R2	0.72 +/- 0.019	0.71 +/- 0.012	0.68 +/- 0.001	0.71 +/- 0.001

Tabela 4.57: Resultados da regressão para os descritores moleculares 1D e 2D e FP - GIM100

4.2.1.2 Classificação

Descritores moleculares 1D e 2D

1D 2D		Non Stringent	Stringent	Non Stringent Blind Test	Stringent Blind Test
RF	Accuracy	0.8939	0.8910	0.8870	0.8973
	Precision	0.8845	0.8849	0.8737	0.8880
	Recall	0.9062	0.9002	0.9034	0.9062
	f1-score	0.8951	0.8923	0.8883	0.8968
SVM	Accuracy	0.7452	0.7242	0.7306	0.7313
	Precision	0.8442	0.8459	0.8169	0.8462
	Recall	0.6014	0.5517	0.5905	0.5569
	f1-score	0.7011	0.6644	0.6843	0.6683
XGBoost	Accuracy	0.8937	0.8915	0.8869	0.8978
	Precision	0.8857	0.8861	0.8751	0.8892
	Recall	0.9040	0.8996	0.9014	0.9055
	f1-score	0.8947	0.8927	0.8880	0.8972
ANN	Accuracy	0.8922	0.8857	0.8858	0.8914
	Precision	0.8874	0.8873	0.8763	0.8894
	Recall	0.8986	0.8847	0.8972	0.8905
	f1-score	0.8928	0.8856	0.8865	0.8895

Tabela 4.58: Resultados da classificação para os descritores moleculares 1D e 2D - GIM100

Descritores moleculares 1D e 2D e *fingerprints*

1D 2D FP		Non Stringent	Stringent	Non Stringent Blind Test	Stringent Blind Test
RF	Accuracy	0.8960	0.8928	0.8863	0.8936
	Precision	0.8817	0.8881	0.8677	0.9047
	Recall	0.9148	0.8960	0.9101	0.8859
	f1-score	0.8979	0.8919	0.8883	0.8951
SVM	Accuracy	0.7560	0.7341	0.7375	0.7318
	Precision	0.8621	0.8474	0.8270	0.8684
	Recall	0.6092	0.5644	0.5974	0.5627
	f1-score	0.7115	0.6729	0.6913	0.6783
XGBoost	Accuracy	0.8957	0.8922	0.8859	0.8928
	Precision	0.8844	0.8901	0.8711	0.9064
	Recall	0.9103	0.8923	0.9045	0.8824
	f1-score	0.8971	0.8909	0.8874	0.8940
ANN	Accuracy	0.8935	0.8910	0.8844	0.8922
	Precision	0.8871	0.8853	0.8745	0.9027
	Recall	0.9020	0.8957	0.8965	0.8856
	f1-score	0.8942	0.8903	0.8851	0.8939

Tabela 4.59: Resultados da classificação para os descritores moleculares 1D e 2D e FP - GIM100

4.2.2 Ganho de informação mútua - Top 500

4.2.2.1 Regressão

Descritores moleculares 1D e 2D

1D 2D		Non Stringent	Stringent	Non Stringent Blind Test	Stringent Blind Test
RF	RMSE	0.83 +/- 0.030	0.84 +/- 0.032	0.86 +/- 0.001	0.82 +/- 0.001
	RP	0.85 +/- 0.011	0.84 +/- 0.011	0.83 +/- 0.001	0.85 +/- 0.001
	R2	0.72 +/- 0.018	0.70 +/- 0.018	0.68 +/- 0.001	0.72 +/- 0.001
SVM	RMSE	0.94 +/- 0.039	0.95 +/- 0.032	0.97 +/- 0.002	0.93 +/- 0.001
	RP	0.80 +/- 0.012	0.79 +/- 0.010	0.77 +/- 0.001	0.80 +/- 0.001
	R2	0.64 +/- 0.019	0.62 +/- 0.015	0.60 +/- 0.001	0.64 +/- 0.001
XGBoost	RMSE	0.83 +/- 0.031	0.84 +/- 0.032	0.86 +/- 0.001	0.82 +/- 0.001
	RP	0.85 +/- 0.011	0.84 +/- 0.011	0.83 +/- 0.001	0.85 +/- 0.001
	R2	0.72 +/- 0.019	0.70 +/- 0.018	0.68 +/- 0.001	0.72 +/- 0.001
ANN	RMSE	0.83 +/- 0.028	0.85 +/- 0.030	0.86 +/- 0.001	0.82 +/- 0.001
	RP	0.85 +/- 0.011	0.84 +/- 0.010	0.83 +/- 0.001	0.85 +/- 0.001
	R2	0.72 +/- 0.018	0.70 +/- 0.017	0.68 +/- 0.001	0.72 +/- 0.001

Tabela 4.60: Resultados da regressão para os descritores moleculares 1D e 2D - GIM500

Descritores moleculares 1D e 2D e fingerprints

1D 2D FP		Non Stringent	Stringent	Non Stringent Blind Test	Stringent Blind Test
RF	RMSE	0.83 +/- 0.031	0.83 +/- 0.034	0.86 +/- 0.001	0.84 +/- 0.001
	RP	0.85 +/- 0.011	0.84 +/- 0.007	0.83 +/- 0.001	0.84 +/- 0.001
	R2	0.72 +/- 0.018	0.71 +/- 0.013	0.68 +/- 0.001	0.71 +/- 0.001
SVM	RMSE	0.84 +/- 0.032	0.85 +/- 0.035	0.87 +/- 0.001	0.86 +/- 0.001
	RP	0.84 +/- 0.011	0.83 +/- 0.008	0.82 +/- 0.001	0.84 +/- 0.001
	R2	0.71 +/- 0.019	0.70 +/- 0.013	0.67 +/- 0.001	0.70 +/- 0.001
XGBoost	RMSE	0.83 +/- 0.030	0.83 +/- 0.034	0.86 +/- 0.001	0.84 +/- 0.001
	RP	0.85 +/- 0.11	0.84 +/- 0.008	0.83 +/- 0.001	0.84 +/- 0.001
	R2	0.72 +/- 0.018	0.71 +/- 0.013	0.68 +/- 0.001	0.71 +/- 0.001
ANN	RMSE	0.83 +/- 0.028	0.83 +/- 0.033	0.86 +/- 0.001	0.84 +/- 0.002
	RP	0.85 +/- 0.011	0.84 +/- 0.007	0.83 +/- 0.001	0.84 +/- 0.001
	R2	0.72 +/- 0.018	0.71 +/- 0.013	0.68 +/- 0.001	0.71 +/- 0.001

Tabela 4.61: Resultados da regressão para os descritores moleculares 1D e 2D e FP - GIM500

4.2.2.2 Classificação

Descritores moleculares 1D e 2D

1D 2D		Non Stringent	Stringent	Non Stringent Blind Test	Stringent Blind Test
RF	Accuracy	0.8941	0.8916	0.8870	0.8975
	Precision	0.8840	0.8843	0.8733	0.8871
	Recall	0.9072	0.9021	0.9041	0.9076
	f1-score	0.8954	0.8930	0.8837	0.8971
SVM	Accuracy	0.8320	0.8258	0.8225	0.8324
	Precision	0.8861	0.8817	0.8718	0.8833
	Recall	0.7633	0.7542	0.7565	0.7605
	f1-score	0.8187	0.8115	0.8080	0.8158
XGBoost	Accuracy	0.8940	0.8919	0.8872	0.8977
	Precision	0.8866	0.8845	0.8763	0.8869
	Recall	0.9037	0.9026	0.9005	0.9083
	f1-score	0.8949	0.8934	0.8881	0.8974
ANN	Accuracy	0.8939	0.8912	0.8882	0.8979
	Precision	0.8896	0.8881	0.8795	0.8921
	Recall	0.8996	0.8963	0.8986	0.9022
	f1-score	0.8944	0.8920	0.8888	0.8970

Tabela 4.62: Resultados da classificação para os descritores moleculares 1D e 2D - GIM500

Descritores moleculares 1D e 2D e *fingerprints*

1D 2D FP		Non Stringent	Stringent	Non Stringent Blind Test	Stringent Blind Test
RF	Accuracy	0.8939	0.8925	0.8850	0.8919
	Precision	0.8842	0.8910	0.8711	0.9075
	Recall	0.9064	0.8915	0.9024	0.8787
	f1-score	0.8951	0.8912	0.8864	0.8928
SVM	Accuracy	0.8882	0.8799	0.8798	0.8824
	Precision	0.8875	0.8853	0.8720	0.9054
	Recall	0.8891	0.8694	0.8888	0.8605
	f1-score	0.8880	0.8771	0.8801	0.8822
XGBoost	Accuracy	0.8966	0.8930	0.8866	0.8931
	Precision	0.8847	0.8920	0.8717	0.9080
	Recall	0.9121	0.8915	0.9052	0.8808
	f1-score	0.8981	0.8916	0.8881	0.8941
ANN	Accuracy	0.8936	0.8920	0.8843	0.8927
	Precision	0.8849	0.8907	0.8702	0.9075
	Recall	0.9051	0.8908	0.9021	0.8808
	f1-score	0.8948	0.8906	0.8858	0.8938

Tabela 4.63: Resultados da classificação para os descritores moleculares 1D e 2D e FP - GIM500

4.2.3 Correlação de Pearson - Top 100

4.2.3.1 Regressão

Descritores moleculares 1D e 2D

1D 2D		Non Stringent	Stringent	Non Stringent Blind Test	Stringent Blind Test
RF	RMSE	0.83 +/- 0.030	0.84 +/- 0.032	0.86 +/- 0.001	0.82 +/- 0.001
	RP	0.85 +/- 0.011	0.84 +/- 0.011	0.83 +/- 0.001	0.85 +/- 0.001
	R2	0.72 +/- 0.018	0.70 +/- 0.018	0.68 +/- 0.001	0.72 +/- 0.001
SVM	RMSE	1.22 +/- 0.042	1.23 +/- 0.032	1.23 +/- 0.001	1.21 +/- 0.001
	RP	0.63 +/- 0.014	0.62 +/- 0.012	0.60 +/- 0.001	0.62 +/- 0.001
	R2	0.39 +/- 0.018	0.38 +/- 0.014	0.35 +/- 0.001	0.39 +/- 0.001
XGBoost	RMSE	0.83 +/- 0.031	0.84 +/- 0.032	0.86 +/- 0.001	0.82 +/- 0.001
	RP	0.85 +/- 0.011	0.84 +/- 0.010	0.83 +/- 0.001	0.85 +/- 0.001
	R2	0.72 +/- 0.019	0.70 +/- 0.018	0.68 +/- 0.001	0.72 +/- 0.001
ANN	RMSE	0.83 +/- 0.031	0.84 +/- 0.034	0.86 +/- 0.003	0.82 +/- 0.002
	RP	0.85 +/- 0.011	0.84 +/- 0.011	0.83 +/- 0.001	0.85 +/- 0.001
	R2	0.72 +/- 0.019	0.70 +/- 0.018	0.68 +/- 0.001	0.72 +/- 0.001

Tabela 4.64: Resultados da regressão para os descritores moleculares 1D e 2D - CP100

Descritores moleculares 1D e 2D e fingerprints

1D 2D FP		Non Stringent	Stringent	Non Stringent Blind Test	Stringent Blind Test
RF	RMSE	0.83 +/- 0.030	0.83 +/- 0.034	0.86 +/- 0.001	0.84 +/- 0.001
	RP	0.85 +/- 0.011	0.84 +/- 0.001	0.83 +/- 0.001	0.84 +/- 0.001
	R2	0.72 +/- 0.018	0.71 +/- 0.013	0.68 +/- 0.001	0.71 +/- 0.001
SVM	RMSE	1.12 +/- 0.042	1.13 +/- 0.036	1.16 +/- 0.002	1.15 +/- 0.001
	RP	0.69 +/- 0.016	0.68 +/- 0.006	0.65 +/- 0.001	0.69 +/- 0.003
	R2	0.48 +/- 0.022	0.47 +/- 0.008	0.43 +/- 0.001	0.48 +/- 0.001
XGBoost	RMSE	0.83 +/- 0.031	0.83 +/- 0.034	0.86 +/- 0.001	0.84 +/- 0.001
	RP	0.85 +/- 0.011	0.84 +/- 0.008	0.83 +/- 0.001	0.84 +/- 0.001
	R2	0.72 +/- 0.019	0.71 +/- 0.013	0.68 +/- 0.001	0.71 +/- 0.001
ANN	RMSE	0.83 +/- 0.031	0.84 +/- 0.034	0.86 +/- 0.001	0.85 +/- 0.005
	RP	0.85 +/- 0.011	0.84 +/- 0.007	0.83 +/- 0.001	0.84 +/- 0.001
	R2	0.72 +/- 0.018	0.70 +/- 0.012	0.68 +/- 0.001	0.71 +/- 0.001

Tabela 4.65: Resultados da regressão para os descritores moleculares 1D e 2D e FP - CP100

4.2.4 Correlação de Pearson - Top 500

4.2.4.1 Regressão

Descritores moleculares 1D e 2D

1D 2D		Non Stringent	Stringent	Non Stringent Blind Test	Stringent Blind Test
RF	RMSE	0.83 +/- 0.027	0.85 +/- 0.031	0.89 +/- 0.002	0.82 +/- 0.003
	RP	0.85 +/- 0.009	0.84 +/- 0.010	0.81 +/- 0.001	0.85 +/- 0.001
	R2	0.72 +/- 0.016	0.70 +/- 0.017	0.66 +/- 0.002	0.72 +/- 0.002
SVM	RMSE	0.94 +/- 0.040	0.96 +/- 0.031	0.98 +/- 0.002	0.94 +/- 0.001
	RP	0.80 +/- 0.011	0.79 +/- 0.009	0.77 +/- 0.001	0.80 +/- 0.001
	R2	0.64 +/- 0.018	0.62 +/- 0.014	0.59 +/- 0.001	0.64 +/- 0.001
XGBoost	RMSE	0.82 +/- 0.029	0.84 +/- 0.030	0.87 +/- 0.001	0.81 +/- 0.002
	RP	0.85 +/- 0.010	0.84 +/- 0.010	0.82 +/- 0.001	0.85 +/- 0.001
	R2	0.72 +/- 0.016	0.70 +/- 0.017	0.67 +/- 0.001	0.72 +/- 0.001
ANN	RMSE	0.82 +/- 0.027	0.84 +/- 0.030	0.87 +/- 0.003	0.81 +/- 0.002
	RP	0.85 +/- 0.010	0.84 +/- 0.010	0.82 +/- 0.001	0.85 +/- 0.001
	R2	0.72 +/- 0.016	0.70 +/- 0.017	0.67 +/- 0.002	0.72 +/- 0.001

Tabela 4.66: Resultados da regressão para os descritores moleculares 1D e 2D - CP500

Descritores moleculares 1D e 2D e fingerprints

1D 2D FP		Non Stringent	Stringent	Non Stringent Blind Test	Stringent Blind Test
RF	RMSE	0.83 +/- 0.030	0.83 +/- 0.034	0.86 +/- 0.001	0.84 +/- 0.001
	RP	0.85 +/- 0.011	0.84 +/- 0.007	0.83 +/- 0.001	0.84 +/- 0.001
	R2	0.72 +/- 0.018	0.71 +/- 0.013	0.68 +/- 0.001	0.71 +/- 0.001
SVM	RMSE	0.84 +/- 0.032	0.85 +/- 0.036	0.88 +/- 0.001	0.86 +/- 0.001
	RP	0.84 +/- 0.011	0.83 +/- 0.008	0.82 +/- 0.001	0.84 +/- 0.001
	R2	0.71 +/- 0.019	0.70 +/- 0.013	0.67 +/- 0.001	0.70 +/- 0.001
XGBoost	RMSE	0.83 +/- 0.030	0.83 +/- 0.034	0.86 +/- 0.001	0.84 +/- 0.001
	RP	0.85 +/- 0.011	0.84 +/- 0.008	0.83 +/- 0.001	0.84 +/- 0.001
	R2	0.72 +/- 0.018	0.71 +/- 0.013	0.68 +/- 0.001	0.71 +/- 0.001
ANN	RMSE	0.83 +/- 0.031	0.83 +/- 0.032	0.86 +/- 0.001	0.84 +/- 0.003
	RP	0.85 +/- 0.011	0.84 +/- 0.007	0.83 +/- 0.001	0.84 +/- 0.001
	R2	0.72 +/- 0.018	0.71 +/- 0.013	0.68 +/- 0.001	0.71 +/- 0.001

Tabela 4.67: Resultados da regressão para os descritores moleculares 1D e 2D e FP - CP500

4.3 Graph Mining

A seguir são apresentados os resultados para o problema de regressão usando o conjunto de dados originado pela introdução do *Graph Mining*. Para os resultados foi seleccionado o modelo

com o conjunto de parâmetros que obtiveram o menor erro. Para os algoritmos utilizados foi usada a biblioteca de Python "Sklearn".

1D 2D		Non Stringent	Stringent	Non Stringent Blind Test	Stringent Blind Test
RF	RMSE	0.84 +/- 0.034	0.89 +/- 0.023	0.95 +/- 0.003	0.86 +/- 0.002
	RP	0.84 +/- 0.011	0.82 +/- 0.011	0.79 +/- 0.002	0.83 +/- 0.001
	R2	0.70 +/- 0.018	0.67 +/- 0.017	0.62 +/- 0.002	0.69 +/- 0.002
XGBoost	RMSE	0.83 +/- 0.032	0.86 +/- 0.032	0.92 +/- 0.002	0.83 +/- 0.004
	RP	0.85 +/- 0.009	0.83 +/- 0.012	0.80 +/- 0.001	0.84 +/- 0.002
	R2	0.71 +/- 0.015	0.69 +/- 0.019	0.64 +/- 0.001	0.71 +/- 0.003

Tabela 4.68: Resultados da regressão para os descritores moleculares 1D e 2D e fragmentos - GM

4.4 Discussão de Resultados

4.4.1 Regressão

Uma das primeiras conclusões, e já esperada, que se pode retirar, após a análise dos resultados, é a pequena redução do desempenho dos modelos *Stringent* em relação aos modelos *Non Stringent*. Isto verifica-se para todos os algoritmos implementados nesta dissertação e pode ser explicado pela grande diferença entre as características das células dos conjuntos de treino e teste. Visto que nos modelos *Stringent* cada compartimento de treino e teste não compartilham nenhuma linha celular, o *overfitting* é menor do que nos modelos *Non Stringent*.

Também se verificou uma redução de desempenho em relação aos resultados dos *blind test*, devido à introdução de novas linhas celulares e medicamentos que a versão 1.1 apresenta, não presentes no conjunto de dados de treino.

Quanto à divisão em dados que apenas contêm os descritores moleculares 1D e 2D e aqueles que contêm os descritores moleculares 1D e 2D e as *fingerprints* verificou-se que a diferença no desempenho dos modelos foi pouco significativa, tendo os descritores moleculares 1D e 2D com as *fingerprints* obtido um desempenho ligeiramente superior.

4.4.2 Classificação

Para a avaliação de desempenho dos algoritmos no problema de classificação, calculou-se a *accuracy*, *precision*, *recall* e o *f1-score*, através da obtenção das matrizes de confusão.

Verificou-se que a média dos resultados obtidos pelos algoritmos para a métrica de *accuracy* se encontra entre 0.80 e 0.90. Isto significa que, em 100 classificações, os algoritmos acertam entre 80 a 90. Para a métrica *precision*, registou-se que a média de valores se encontra entre 0.80 e 0.90. Isto significa que dos valores considerados "Bom" para os valores de IC50, 80% a 90% são realmente "Bom". Quanto ao *recall*, a média de valores encontra-se entre 0.87 e 0.91. Significa isto que, dos valores IC50 classificados como positivos, 86% a 90% foram classificados

corretamente pelo modelo. Para a métrica *f1-score*, que é uma média harmónica entre a métrica de precision e recall, a média de valores encontra-se entre 0.86 e 0.91.

Ao analisar os resultados obtidos, no geral, entre os conjuntos de dados que contêm os descritores moleculares 1D e 2D e aqueles que contêm os descritores moleculares e as *fingerprints*, a diferença é mínima, verificando-se um maior desempenho para o conjunto de dados que contêm os descritores moleculares e as *fingerprints* nomeadamente para o algoritmo *SVM* e o *ANN*.

Quanto ao *blind test*, observou-se uma esperada redução em todas as métricas calculadas, devido à introdução de novas linhas celulares e medicamentos que a versão 1.1 apresenta, não presentes no conjunto de dados de treino.

Entre os conjuntos de dados *Stringent* e *Non Stringent*, não existe uma diferença significativa, existindo em alguns algoritmos um desempenho ligeiramente superior no conjunto de dados *Non Stringent*.

4.4.3 Feature Selection

Para os conjuntos de dados que foram otimizados pelo *Feature Selection*, os desempenhos tanto para o problema de regressão como o de classificação foram superiores comparativamente aos conjuntos de dados completo. Conclui-se assim que uma quantidade significativa de características são de baixa qualidade contribuindo apenas com ruído. Entre os algoritmos, para cada técnica de *Feature Selection*, as conclusões dos resultados são iguais às obtidas para o conjunto de dados completo.

4.4.4 Graph Mining

Para o conjunto de dados originado pelo *Graph Mining*, para o problema de regressão, não existiu uma melhoria no desempenho sendo os resultados semelhantes. Para as técnicas da correlação de Pearson e o ganho de informação mútua aplicadas no conjunto de dados originado pelo *Graph Mining*, verifica-se que os resultados dos fragmentos são ligeiramente superiores aos obtidos pelas *fingerprints*, o que significa que as características de *gSpan* adicionadas tiveram, no geral, uma maior importância. Estes resultados são consistentes com uma experiência semelhante realizada pelo (Ladeiras, 2015) apesar de que na sua experiência os dados foram altamente pré-processados.

Capítulo 5

Conclusões e Trabalho Futuro

5.1 Satisfação dos Objetivos

Esta dissertação teve como finalidade construir modelos que previssem a eficácia dos medicamentos em linhas celulares (IC50), usando os resultados de experiências *in vitro*, a caracterização das linhas celulares e informações sobre o medicamento em uso. Este é um assunto que se reveste de grande importância, não apenas pelo seu imenso potencial no que concerne a diminuição do trabalho repetitivo, moroso e dispendioso, feito *in vitro*, mas também constitui um passo importante no tratamento personalizado. Outro propósito foi o de auxiliar a exploração de novas metodologias para aumentar o desempenho e compreensibilidade, o que, consequentemente, amplia a oportunidade de melhoria no tratamento. O objetivo de obter informações e fornecer soluções para melhorar o desempenho dos modelos foi alcançado.

Nesta dissertação foram introduzidos não só métodos para melhorar os resultados de trabalhos anteriores, como também novos algoritmos, que provaram ser relevantes para a obtenção de melhores resultados para os modelos. A separação dos descritores moleculares 1D e 2D das *fingerprints* não foi muito significativa, visto que, para os diferentes algoritmos, em ambas as situações, as métricas obtiveram melhores resultados mas com uma diferença mínima. Quanto ao desempenho dos modelos *Stringent*, verificou-se uma redução em relação aos modelos *Non Stringent*. A introdução do processo de *feature selection* provou ser eficiente e útil na eliminação de atributos que apenas apresentavam ruído no conjunto de dados. Os conjuntos de dados que sofreu este processo obtiveram melhores resultados. Verificou-se, também, com este processo que descritores moleculares 1D e 2D obtiveram um desempenho consistente tendo a maior parte obtido valores relevantes. As *fingerprints* e as características genômicas tiveram um desempenho inferior, comparativamente aos descritores moleculares, sendo apenas uma pequena percentagem considerada relevante. Por fim, o conjunto de dados originado com informação obtida por utilização de técnicas de *Graph Mining*, obteve resultados semelhantes aos dados do conjunto completo, apesar de os resultados dos fragmentos para a técnica de correlação de Pearson e do ganho de informação mútua terem sido superiores aos obtidos pelas *fingerprints*, o que significa que as características de *gSpan* adicionadas tiveram, no geral, uma maior importância.

Apesar das limitações do projeto, a sua otimização apresenta o potencial de demonstrar a importância do uso de metodologias *in silico* para melhorar a qualidade de vida dos pacientes e ajudar no tratamento da doença.

5.2 Limitações e Trabalho Futuro

O principal fator de limitação ao longo do desenvolvimento da dissertação foi o tempo necessário para treino dos modelos nas máquinas disponíveis. Mesmo tendo disponível um número considerável de máquinas para o treino dos modelos, as execuções foram bastante custosas em termos de tempo. A quantidade de memória e a potência computacional necessária para os treinos dos modelos era consideravelmente alta. Devido a esta limitação a sintonização de parâmetros teve que sofrer alterações para assim ser possível o treino dos modelos gerados.

Posto isto, uma das propostas para o trabalho futuro seria fazer um ajuste de parâmetros mais variado, que é apenas limitado pelo poder de computação e tempo associado. Testando os algoritmos com uma maior variação dos parâmetros, resultaria em uma previsão mais precisa dos modelos.

Também, explorar o uso de *Deep Learning* neste problema específico seria relevante, uma vez que existe literatura que indica ganhos de desempenho em problemas no contexto de quiminformática (Min et al., 2017).

Referências

- ChemSpider, 2020. URL <http://www.chemspider.com/>.
- PharmaInformatic, 2020. URL <https://www.pharmainformatic.com/html/pact-f.html>.
- National Cancer Institute, 2020. URL <https://www.cancer.gov/publications/dictionaries/geneticsdictionary>.
- W. D. Hounshell A. K. I. Gushurst D. L. Grier B. A. Leland A. Dalby, J. G. Nourse e J. Laufer. Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *Journal of Chemical Information and Computer Sciences*, 32(3): 244–255, 1992. ISSN 0095-2338. doi: 10.1021/ci00007a012. URL <http://pubs.acs.org/doi/abs/10.1021/ci00007a012>.
- Vinicius M. Alves, Rodolpho C. Braga, Eugene N. Muratov e Carolina Horta Andrade. Cheminformatics: An introduction. *Quimica Nova*, 41(2):202–212, 2018. ISSN 16787064. doi: 10.21577/0100-4042.20170145.
- Peter Auer, Harald Burgsteiner e Wolfgang Maass. A learning rule for very simple universal approximators consisting of a single layer of perceptrons. *Neural Networks*, 21(5):786–795, 2008. ISSN 08936080. doi: 10.1016/j.neunet.2007.12.036.
- Ana Azevedo e Manuel Filipe Santos. *KDD, semma and CRISP-DM: A parallel overview*. tese de doutoramento, 2008.
- Max Bramer. *Principles of Data Mining*. Number February. 2007. ISBN 9781846287664. doi: 10.1007/978-1-84628-766-4.
- Martin D. Buhmann. In *Radial basis functions: Theory and implementations*. Cambridge University Press, 2003.
- Pete Chapman, Clinton Julian, Kerber Randy, Khabaza Thomas, Reinartz Thomas, Shearer Colin e Rudiger Wirth. CRISP-DM 1.0: Step-by-step data mining guide. *CRISP-DM Consortium*, página 76, 2000.
- Arshad Chaudry. Cell Culture. *The Science Creative Quarterly*, 2004. URL :<http://www.scq.ubc.ca/cell-culture/>.
- Chun Wei Yap. PaDEL-Descriptor: An Open Source Software to Calculate Molecular Descriptors and Fingerprints. *Journal of computational chemistry*, 32:1466–1474, 2010. ISSN 1096-987X. doi: 10.1002/jcc.21707.

Vladimir Vapnik Corinna Cortes. Support-Vector Networks. In *Machine Learning*, páginas 273—297. 1995. doi: 10.1023/A:1022627411411.

James C. Costello, Laura M. Heiser, Elisabeth Georgii, Mehmet Gönen, Michael P. Menden, Nicholas J. Wang, Mukesh Bansal, Muhammad Ammad-Ud-Din, Petteri Hintsanen, Suleiman A. Khan, John Patrick Mpindi, Olli Kallioniemi, Antti Honkela, Tero Aittokallio, Krister Wennerberg, James J. Collins, Dan Gallahan, Dinah Singer, Julio Saez-Rodriguez, Samuel Kaski, Joe W. Gray, Gustavo Stolovitzky, Jean Paul Abbuehl, Jeffrey Allen, Russ B. Altman, Shawn Balcome, Alexis Battle, Andreas Bender, Bonnie Berger, Jonathan Bernard, Madhuchhanda Bhattacharjee, Krithika Bhuvaneshwar, Andrew A. Bieberich, Fred Boehm, Andrea Califano, Christina Chan, Beibei Chen, Ting Huei Chen, Jaejoon Choi, Luis Pedro Coelho, Thomas Cokelaer, James C. Collins, Chad J. Creighton, Jike Cui, Will Dampier, V. Jo Davisson, Bernard De Baets, Raamesh Deshpande, Barbara DiCamillo, Murat Dundar, Zhana Duren, Adam Ertel, Haoyang Fan, Hongbin Fang, Robinder Gauba, Assaf Gottlieb, Michael Grau, Yuriy Gusev, Min Jin Ha, Leng Han, Michael Harris, Nicholas Henderson, Hussein A. Hejase, Krisztian Homicsko, Jack P. Hou, Woonchang Hwang, Adriaan P. IJzerman, Bilge Karacali, Sunduz Keles, Christina Kendzioriski, Junho Kim, Min Kim, Youngchul Kim, David A. Knowles, Daphne Koller, Junehawk Lee, Jae K. Lee, Eelke B. Lenselink, Biao Li, Bin Li, Jun Li, Han Liang, Jian Ma, Subha Madhavan, Sean Mooney, Chad L. Myers, Michael A. Newton, John P. Overington, Ranadip Pal, Jian Peng, Richard Pestell, Robert J. Prill, Peng Qiu, Bartek Rajwa, Anguraj Sadasanandam, Francesco Sambo, Hyunjin Shin, Jiuzhou Song, Lei Song, Arvind Sridhar, Michiel Stock, Wei Sun, Tram Ta, Mahlet Tadesse, Ming Tan, Hao Tang, Dan Theodorescu, Gianna Maria Toffolo, Aydin Tozeren, William Trepicchio, Nelle Varoquaux, Jean Philippe Vert, Willem Waegeman, Thomas Walter, Qian Wan, Difei Wang, Wen Wang, Yong Wang, Zhishi Wang, Joerg K. Wegner, Tongtong Wu, Tian Xia, Guanghua Xiao, Yang Xie, Yanxun Xu, Jichen Yang, Yuan Yuan, Shihua Zhang, Xiang Sun Zhang, Junfei Zhao, Chandler Zuo, Herman W.T. Van Vlijmen e Gerard J.P. Van Westen. A community effort to assess and improve drug sensitivity prediction algorithms. *Nature Biotechnology*, 32(12):1202–1212, 2014. ISSN 15461696. doi: 10.1038/nbt.2877.

Koby Crammer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research - JMLR*, 2(2):265–292, 2002. ISSN 1533-7928.

Allan Peter Davis, Cynthia J. Grondin, Robin J. Johnson, Daniela Sciaky, Benjamin L. King, Roy McMorran, Jolene Wiegiers, Thomas C. Wiegiers e Carolyn J. Mattingly. The Comparative Toxicogenomics Database: Update 2017. *Nucleic Acids Research*, 45(D1):D972–D978, 2017. ISSN 13624962. doi: 10.1093/nar/gkw838.

Kirill Degtyarenko, Paula De matos, Marcus Ennis, Janna Hastings, Martin Zbinden, Alan Mcnaught, Rafael Alcántara, Michael Darsow, Mickaël Guedj e Michael Ashburner. ChEBI: A database and ontology for chemical entities of biological interest. *Nucleic Acids Research*, 36 (SUPPL. 1):344–350, 2008. ISSN 03051048. doi: 10.1093/nar/gkm791.

Ramesh Sharda David King Efraim Turban, Dursun Delen. *Business Intelligence: A Managerial Approach*. Pearson Education Limited, 4th edição, 2017. ISBN 9781292220543.

Silva Ferreira. *Classificação Hierárquica da AtividadeEconómicadas Empresas a partir de Texto da Web*. tese de doutoramento, Faculty of Engineering of University of Porto, 2012.

Alex A. Freitas. Understanding the crucial differences between classification and discovery of association rules. *ACM SIGKDD Explorations Newsletter*, 2(1):65–69, 2000. ISSN 19310145. doi: 10.1145/360402.360423.

- T. Hastie G. James, D. Witten e R. Tibshirani. An Introduction to Statistical Learning. *Springer Texts in Statistics*, 103, 2013. ISSN 978-1-4614-7137-0 978-1-4614-7138-7. doi: 10.1007/F978-1-4614-7138-7. URL <http://link.springer.com/10.1007/978-1-4614-7138-7>.
- Mathew J. Garnett, Elena J. Edelman, Sonja J. Heidorn, Chris D. Greenman, Anahita Dastur, King Wai Lau, Patricia Greninger, I. Richard Thompson, Xi Luo, Jorge Soares, Qingsong Liu, Francesco Iorio, Didier Surdez, Li Chen, Randy J. Milano, Graham R. Bignell, Ah T. Tam, Helen Davies, Jesse A. Stevenson, Syd Barthorpe, Stephen R. Lutz, Fiona Kogera, Karl Lawrence, Anne McLaren-Douglas, Xenia Mitropoulos, Tatiana Mironenko, Helen Thi, Laura Richardson, Wenjun Zhou, Frances Jewitt, Tinghu Zhang, Patrick O'Brien, Jessica L. Boisvert, Stacey Price, Wooyoung Hur, Wanjun Yang, Xianming Deng, Adam Butler, Hwan Geun Choi, Jae Won Chang, Jose Baselga, Ivan Stamenkovic, Jeffrey A. Engelman, Sreenath V. Sharma, Olivier Delattre, Julio Saez-Rodriguez, Nathanael S. Gray, Jeffrey Settleman, P. Andrew Futreal, Daniel A. Haber, Michael R. Stratton, Sridhar Ramaswamy, Ultan McDermott e Cyril H. Benes. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, 483(7391):570–575, 2012. ISSN 00280836. doi: 10.1038/nature11005.
- Johann Gasteiger e Kimito Funatsu. Chemoinformatics – An Important Scientific Discipline. *Journal of Computer Chemistry, Japan*, 5(2):53–58, 2006. ISSN 1347-1767. doi: 10.2477/jccj.5.53.
- Tânia Gomes. *Ferramentas Open Source de Data Mining*. tese de doutoramento, Instituto Superior de Engenharia de Coimbra, 2014. URL http://files.isec.pt/DOCUMENTOS/SERVICOS/BIBLIO/Teses/Tese{__}Mest{__}Tania-Gomes.pdf.
- Duarte. Guilherme. Gradient boostings parte 1: Métodos de ensemble learning, 2020. URL <https://datarisk.io/gradient-boostings-parte-1-metodos-de-ensemble-learning/>.
- Francesco Iorio, Theo A. Knijnenburg, Daniel J. Vis, Graham R. Bignell, Michael P. Menden, Michael Schubert, Nanne Aben, Emanuel Gonçalves, Syd Barthorpe, Howard Lightfoot, Thomas Cokelaer, Patricia Greninger, Ewald van Dyk, Han Chang, Heshani de Silva, Holger Heyn, Xianming Deng, Regina K. Egan, Qingsong Liu, Tatiana Mironenko, Xenia Mitropoulos, Laura Richardson, Jinhua Wang, Tinghu Zhang, Sebastian Moran, Sergi Sayols, Maryam Soleimani, David Tamborero, Nuria Lopez-Bigas, Petra Ross-Macdonald, Manel Esteller, Nathanael S. Gray, Daniel A. Haber, Michael R. Stratton, Cyril H. Benes, Lodewyk F.A. Wessels, Julio Saez-Rodriguez, Ultan McDermott e Mathew J. Garnett. A Landscape of Pharmacogenomic Interactions in Cancer. *Cell*, 166(3):740–754, 2016. ISSN 10974172. doi: 10.1016/j.cell.2016.06.017.
- Catarina Isabel e Peixoto Candeias. *Data Mining for Rational Drug Design*. tese de doutoramento, Faculty of Engineering of University of Porto, 2017.
- A. A. Alizadeh M. B. Eisen A. Pergamenschikov C. F. Williams S. S. Jeffrey D. Botstein J. R. Pollack, C. M. Perou e P. O. Brown. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nature Genetics*, 23:41–46, 1999. ISSN 1061-4036. doi: 10.1038/12640.
- R. Ward J. S. Soothill e A. J. Girling. The IC50: An exactly defined measure of antibiotic sensitivity. *Journal of Antimicrobial Chemotherapy*, 29:137–139, 1992. ISSN 0305-7453. doi: 10.1093/jac/29.2.137.

- Teresa Juan-Blanco, Miquel Duran-Frigola e Patrick Aloy. IntSide: A web server for the chemical and biological examination of drug side effects. *Bioinformatics*, 31(4):612–613, 2015. ISSN 14602059. doi: 10.1093/bioinformatics/btu688.
- Minoru Kanehisa, Yoko Sato, Masayuki Kawashima, Miho Furumichi e Mao Tanabe. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, 44(D1):D457–D462, 2016. ISSN 13624962. doi: 10.1093/nar/gkv1070.
- Sunghwan Kim, Paul A. Thiessen, Evan E. Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, Jane He, Siqian He, Benjamin A. Shoemaker, Jiyao Wang, Bo Yu, Jian Zhang e Stephen H. Bryant. PubChem substance and compound databases. *Nucleic Acids Research*, 44(D1):D1202–D1213, 2016. ISSN 13624962. doi: 10.1093/nar/gkv951.
- T. Kohonen. Learning vector quantization. in m. arbib. In *The handbook of brain theory and neural networks*, páginas 537–540. United Kingdom: MIT Press, 1995.
- João Tiago Chaves Miranda Ladeiras. *Response Prediction to Cancer Cell Lines Treatment cells*. tese de doutoramento, University of Porto, 2015.
- Wei Yin Loh. Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1):14–23, 2011. ISSN 19424787. doi: 10.1002/widm.8.
- Marta Milheiro Soeiro Nunes Lopes. *Data Mining para estudos de interação entre fármacos e previsão de Efeitos Adversos de Fármacos*. tese de doutoramento, Faculty of Engineering of University of Porto, 2018. URL <https://repositorio-aberto.up.pt/handle/10216/114187>.
- J. R. Masters. “Human cancer cell lines: Fact and fantasy. *Nature Reviews. Molecular Cell Biology*, 1:233–236, 2000. ISSN 1471-0072. doi: 10.1038/35043102.
- Michael P. Menden, Francesco Iorio, Mathew Garnett, Ultan McDermott, Cyril H. Benes, Pedro J. Ballester e Julio Saez-Rodriguez. Machine Learning Prediction of Cancer Cell Sensitivity to Drugs Based on Genomic and Chemical Properties. *PLoS ONE*, 8(4), 2013. ISSN 19326203. doi: 10.1371/journal.pone.0061318.
- Seonwoo Min, Byunghan Lee e Sungroh Yoon. Deep learning in bioinformatics. *Briefings in bioinformatics*, 18(5):851–869, 2017. ISSN 14774054. doi: 10.1093/bib/bbw068.
- Jéssica Namora. *Previsão de efeitos adversos de medicamentos*. tese de doutoramento, Faculty of Engineering of University of Porto, 2017.
- Siegfried Nijssen e Joost N. Kok. A quickstart in frequent structure mining can make a difference. *KDD-2004 - Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, páginas 647–652, 2004. doi: 10.1145/1014052.1014134.
- Noel M O’Boyle, Michael Banck, Craig A James, Chris Morley, Tim Vandermeersch e Geoffrey R Hutchison. Open Babel. *Journal of Cheminformatics*, 3(33):1–14, 2011. ISSN 17582946. doi: 10.1186/1758-2946-3-33. URL <https://jcheminf.biomedcentral.com/track/pdf/10.1186/1758-2946-3-33>.
- Marisa Oliveira. *Análise de som cardíaco pediátrico para identificação de sopro*. tese de doutoramento, University of Porto, 2020.

- Jiawei Han Micheline Kamber Jian Pei. *Data mining: Data mining concepts and techniques*. Morgan Kaufmann Publishers Inc., 3rd edição, 2011. ISBN 9780123814807. doi: 10.1109/ICMIRA.2013.45.
- João Pereira. *Modelos de Data Mining para multi-previsão: aplicação à medicina intensiva*. tese de doutoramento, University of Minho, 2005. URL <http://hdl.handle.net/1822/25914>.
- J. Ross Quinlan. Generating production rules from decision trees. *Proceedings of the Tenth International Joint Conference on Artificial Intelligence*, 30107:304–307, 1987.
- V. Consonni R. Todeschini e R. Mannhold. , Molecular Descriptors for Chemoinformatics: Volume I: Alphabetical Listing / Volume II: Appendices, References. 2009. ISSN 978-3-527-31852-0.
- D. A. Haber S. V. Sharma e J. Settleman. Cell line-based platforms to evaluate the therapeutic efficacy of candidate anticancer agents. *Nature Reviews Cancer*, 10:241–253, 2010. ISSN 1474-175X, 1474-1768. doi: 10.1038/nrc2820.
- U. Shokal e P. C. Sharma. “Implication of microsatellite instability in human gastric cancers. *The Indian Journal of Medical Research*, 135:599–613, 2012. ISSN 0971-5916.
- Kyle Strimbu e Jorge A. Tavel. What are biomarkers? *Current Opinion in HIV and AIDS*, 5(6): 463–466, 2010. ISSN 1746630X. doi: 10.1097/COH.0b013e32833ed177.
- T. Honkela T. Kohonen. Kohonen network. *Scholarpedia*, 2007. doi: 10.4249/scholarpedia.1568.
- Bruno Miguel Viana Gomes. *Previsao de Churn em Companhias de Seguros*. tese de doutoramento, University of Minho, 2011. URL <http://hdl.handle.net/1822/28301>.
- David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, páginas 31–36, 1988. doi: 10.1021/ci00057a005.
- Ian H Witten, Eibe Frank e Mark a Hall. *Data Mining: Practical Machine Learning Tools and Techniques (Google eBook)*. 2005. ISBN 0080890369. URL <http://books.google.com/books?id=bDtLM8CODsQC{&}pgis=1>.
- Xifeng Yan e Jiawei Han. gspan: Graph-based substructure pattern mining. In *Data Mining. Proceedings. 2002 IEEE International Conference on*, páginas 721–724, 2002. doi: 10.1109/ICDM.2002.1184038.
- Hui Zeng, Chengxiang Qiu e Qinghua Cui. Drug-Path: A database for drug-induced pathways. *Database*, 2015:1–4, 2015. ISSN 17580463. doi: 10.1093/database/bav061.

Anexo A

Características

A.1 Características das linhas celulares

1. MIS	18. CDK4 CN	35. ERBB2 SV
2. AKT2 CN	19. CDK6 CN	36. ERBB2 CN
3. ALK SV	20. CDKN2A SV	37. EZH2 SV
4. ALK CN	21. CDKN2A CN	38. EZH2 CN
5. APC SV	22. CDKN2C SV	39. FAM123B SV
6. APC CN	23. CDKN2C CN	40. FAM123B CN
7. BRAF SV	24. CDKN2a(p14) SV	41. FBXW7 SV
8. BRAF CN	25. CDKN2a(p14) CN	42. FBXW7 CN
9. BRCA1 SV	26. CTNNB1 SV	43. FGFR2 CN
10. BRCA1 CN	27. CTNNB1 CN	44. FGFR3 SV
11. BRCA2 SV	28. CYLD SV	45. FGFR3 CN
12. BRCA2 CN	29. CYLD CN	46. FLCN SV
13. CCND1 CN	30. DICER1 SV	47. FLCN CN
14. CCND2 CN	31. DICER1 CN	48. FLT3 SV
15. CCND3 CN	32. EGFR SV	49. FLT3 CN
16. CDH1 SV	33. EGFR CN	50. GNAS SV
17. CDH1 CN	34. EP300 CN	51. GNAS CN

52. HRAS SV	78. MSH6 SV	104. PIK3R1 SV
53. HRAS CN	79. MSH6 CN	105. PIK3R1 CN
54. IDH1 SV	80. MYC SV	106. PTCH1 SV
55. IDH1 CN	81. MYC CN	107. PTCH1 CN
56. JAK2 SV	82. MYCL1 CN	108. PTEN SV
57. JAK2 CN	83. MYCN CN	109. PTEN CN
58. KDM5C SV	84. NF1 SV	110. RB1 SV
59. KDM5C CN	85. NF1 CN	111. RB1 CN
60. KDM6A SV	86. NF2 SV	112. RET SV
61. KDM6A CN	87. NF2 CN	113. RET CN
62. KDR SV	88. NOTCH1 SV	114. RUNX1 SV
63. KDR CN	89. NOTCH1 CN	115. RUNX1 CN
64. KIT SV	90. NPM1 SV	116. SETD2 SV
65. KIT CN	91. NPM1 CN	117. SETD2 CN
66. KRAS SV	92. NRAS SV	118. SMAD4 SV
67. KRAS CN	93. NRAS CN	119. SMAD4 CN
68. MAP2K4 SV	94. NTRK3 SV	120. SMARCA4 SV
69. MAP2K4 CN	95. NTRK3 CN	121. SMARCA4 CN
70. MDM2 CN	96. PALB2 SV	122. SMARCB1 CN
71. MET SV	97. PALB2 CN	123. SMO SV
72. MET CN	98. PDGFRA SV	124. SMO CN
73. MLH1 SV	99. PDGFRA CN	125. SOCS1 SV
74. MLH1 CN	100. PHOX2B SV	126. SOCS1 CN
75. MLLT3 CN	101. PHOX2B CN	127. STK11 SV
76. MSH2 SV	102. PIK3CA SV	128. STK11 CN
77. MSH2 CN	103. PIK3CA CN	129. SUFU SV

130. SUFU CN	135. TSC1 SV	140. VHL CN
131. TET2 SV	136. TSC1 CN	141. WT1 SV
132. TET2 CN	137. TSC2 SV	142. WT1 CN
133. TP53 SV	138. TSC2 CN	
134. TP53 CN	139. VHL SV	

A.2 Descritores dos compostos

1. nAcid	21. ATSc1	41. BCUTp-1l
2. ALogP	22. ATSc2	42. BCUTp-1h
3. ALogp2	23. ATSc3	43. nBonds
4. AMR	24. ATSc4	44. nBonds2
5. apol	25. ATSc5	45. nBondsS
6. naAromAtom	26. ATSm1	46. nBondsS2
7. nAromBond	27. ATSm2	47. nBondsS3
8. nAtom	28. ATSm3	48. nBondsD
9. nHeavyAtom	29. ATSm4	49. nBondsD2
10. nH	30. ATSm5	50. nBondsT
11. nB	31. ATSp1	51. bpol
12. nC	32. ATSp2	52. C1SP1
13. nN	33. ATSp3	53. C2SP1
14. nO	34. ATSp4	54. C1SP2
15. nS	35. ATSp5	55. C2SP2
16. nP	36. nBase	56. C3SP2
17. nF	37. BCUTw-1l	57. C1SP3
18. nCl	38. BCUTw-1h	58. C2SP3
19. nBr	39. BCUTc-1l	59. C3SP3
20. nI	40. BCUTc-1h	60. C4SP3

61. SCH-3	87. SP-2	113. nsssN
62. SCH-4	88. SP-3	114. naasN
63. SCH-5	89. SP-4	115. ndO
64. SCH-6	90. SP-5	116. nssO
65. SCH-7	91. SP-6	117. naaO
66. VCH-3	92. SP-7	118. nsF
67. VCH-4	93. VP-0	119. ndsssP
68. VCH-5	94. VP-1	120. ndS
69. VCH-6	95. VP-2	121. nssS
70. VCH-7	96. VP-3	122. naaS
71. SC-3	97. VP-4	123. nddssS
72. SC-4	98. VP-5	124. nsCl
73. SC-5	99. VP-6	125. nsBr
74. SC-6	100. VP-7	126. nsI
75. VC-3	101. ECCEN	127. SdS
76. VC-4	102. nHBa	128. mindS
77. VC-5	103. nwHBa	129. maxHBa
78. VC-6	104. nsssB	130. maxwHBa
79. SPC-4	105. ntsC	131. maxdssC
80. SPC-5	106. ndssC	132. maxaasC
81. SPC-6	107. naasC	133. maxaaaC
82. VPC-4	108. naaaC	134. maxdsN
83. VPC-5	109. nssssC	135. maxaaN
84. VPC-6	110. ntN	136. maxsssN
85. SP-0	111. ndsN	137. maxaasN
86. SP-1	112. naaN	138. maxdO

139. maxssO	165. ETA_Beta_ns	191. nHBDOn_Lipinski
140. maxsF	166. ETA_BetaP_ns	192. Kier1
141. maxsCl	167. ETA_dBeta	193. Kier2
142. hmax	168. ETA_dBetaP	194. Kier3
143. hmin	169. ETA_Beta_ns_d	195. nAtomLC
144. ETA_Alpha	170. ETA_BetaP_ns_d	196. nAtomP
145. ETA_AlphaP	171. ETA_Eta	197. nAtomLAC
146. ETA_dAlpha_A	172. ETA_EtaP	198. MLogP
147. ETA_dAlpha_B	173. ETA_Eta_R	199. McGowan_Volume
148. ETA_Epsilon_1	174. ETA_Eta_F	200. MDEC-11
149. ETA_Epsilon_2	175. ETA_EtaP_F	201. MDEC-12
150. ETA_Epsilon_3	176. ETA_Eta_L	202. MDEC-13
151. ETA_Epsilon_4	177. ETA_EtaP_L	203. MDEC-14
152. ETA_Epsilon_5	178. ETA_Eta_R_L	204. MDEC-22
153. ETA_dEpsilon_A	179. ETA_Eta_F_L	205. MDEC-23
154. ETA_dEpsilon_B	180. ETA_EtaP_F_L	206. MDEC-24
155. ETA_dEpsilon_C	181. ETA_Eta_B	207. MDEC-33
156. ETA_Psi_1	182. ETA_EtaP_B	208. MDEC-34
157. ETA_dPsi_A	183. ETA_Eta_B_RC	209. MDEC-44
158. ETA_Shape_P	184. ETA_EtaP_B_RC	210. MDEO-11
159. ETA_Shape_Y	185. fragC	211. MDEO-12
160. ETA_Shape_X	186. nHBAcc	212. MDEO-22
161. ETA_Beta	187. nHBAcc2	213. MDEN-11
162. ETA_BetaP	188. nHBAcc3	214. MDEN-12
163. ETA_Beta_s	189. nHBAcc_Lipinski	215. MDEN-13
164. ETA_BetaP_s	190. nHBDOn	216. MDEN-22

217. MDEN-23	235. nG12Ring	253. nT12Ring
218. MDEN-33	236. nFRing	254. nTG12Ring
219. MLFER_A	237. nF6Ring	255. nRotB
220. MLFER_BH	238. nF8Ring	256. LipinskiFailures
221. MLFER_BO	239. nF9Ring	257. TopoPSA
222. MLFER_S	240. nF10Ring	258. VABC
223. MLFER_E	241. nF11Ring	259. VAdjMat
224. MLFER_L	242. nF12Ring	260. MW
225. PetitjeanNumber	243. nFG12Ring	261. WTPT-1
226. nRing	244. nTRing	262. WTPT-2
227. n3Ring	245. nT4Ring	263. WTPT-3
228. n4Ring	246. nT5Ring	264. WTPT-4
229. n5Ring	247. nT6Ring	265. WTPT-5
230. n6Ring	248. nT7Ring	266. WPATH
231. n7Ring	249. nT8Ring	267. WPOL
232. n8Ring	250. nT9Ring	268. XLogP
233. n9Ring	251. nT10Ring	269. Zagreb
234. n10Ring	252. nT11Ring	

A.3 Fingerprints dos compostos

1. PubchemFP6	8. PubchemFP15	15. PubchemFP22
2. PubchemFP9	9. PubchemFP16	16. PubchemFP23
3. PubchemFP10	10. PubchemFP17	17. PubchemFP24
4. PubchemFP11	11. PubchemFP18	18. PubchemFP25
5. PubchemFP12	12. PubchemFP19	19. PubchemFP30
6. PubchemFP13	13. PubchemFP20	20. PubchemFP33
7. PubchemFP14	14. PubchemFP21	21. PubchemFP34

22. PubchemFP37	48. PubchemFP160	74. PubchemFP218
23. PubchemFP38	49. PubchemFP164	75. PubchemFP219
24. PubchemFP43	50. PubchemFP167	76. PubchemFP227
25. PubchemFP44	51. PubchemFP178	77. PubchemFP228
26. PubchemFP46	52. PubchemFP179	78. PubchemFP232
27. PubchemFP115	53. PubchemFP180	79. PubchemFP233
28. PubchemFP116	54. PubchemFP181	80. PubchemFP241
29. PubchemFP117	55. PubchemFP182	81. PubchemFP246
30. PubchemFP118	56. PubchemFP183	82. PubchemFP247
31. PubchemFP129	57. PubchemFP184	83. PubchemFP248
32. PubchemFP130	58. PubchemFP185	84. PubchemFP252
33. PubchemFP132	59. PubchemFP186	85. PubchemFP255
34. PubchemFP143	60. PubchemFP187	86. PubchemFP256
35. PubchemFP144	61. PubchemFP188	87. PubchemFP257
36. PubchemFP145	62. PubchemFP189	88. PubchemFP258
37. PubchemFP146	63. PubchemFP190	89. PubchemFP259
38. PubchemFP147	64. PubchemFP191	90. PubchemFP260
39. PubchemFP148	65. PubchemFP192	91. PubchemFP261
40. PubchemFP149	66. PubchemFP193	92. PubchemFP262
41. PubchemFP150	67. PubchemFP194	93. PubchemFP274
42. PubchemFP152	68. PubchemFP195	94. PubchemFP276
43. PubchemFP153	69. PubchemFP199	95. PubchemFP283
44. PubchemFP155	70. PubchemFP200	96. PubchemFP284
45. PubchemFP156	71. PubchemFP206	97. PubchemFP285
46. PubchemFP157	72. PubchemFP213	98. PubchemFP286
47. PubchemFP159	73. PubchemFP214	99. PubchemFP287

100. PubchemFP293	126. PubchemFP346	152. PubchemFP376
101. PubchemFP294	127. PubchemFP347	153. PubchemFP377
102. PubchemFP297	128. PubchemFP349	154. PubchemFP378
103. PubchemFP298	129. PubchemFP350	155. PubchemFP379
104. PubchemFP299	130. PubchemFP351	156. PubchemFP380
105. PubchemFP300	131. PubchemFP352	157. PubchemFP381
106. PubchemFP301	132. PubchemFP353	158. PubchemFP382
107. PubchemFP305	133. PubchemFP355	159. PubchemFP383
108. PubchemFP308	134. PubchemFP356	160. PubchemFP384
109. PubchemFP314	135. PubchemFP357	161. PubchemFP385
110. PubchemFP327	136. PubchemFP358	162. PubchemFP386
111. PubchemFP328	137. PubchemFP359	163. PubchemFP387
112. PubchemFP330	138. PubchemFP360	164. PubchemFP388
113. PubchemFP332	139. PubchemFP362	165. PubchemFP389
114. PubchemFP333	140. PubchemFP363	166. PubchemFP390
115. PubchemFP334	141. PubchemFP364	167. PubchemFP391
116. PubchemFP335	142. PubchemFP365	168. PubchemFP392
117. PubchemFP336	143. PubchemFP366	169. PubchemFP393
118. PubchemFP337	144. PubchemFP367	170. PubchemFP394
119. PubchemFP338	145. PubchemFP368	171. PubchemFP395
120. PubchemFP339	146. PubchemFP370	172. PubchemFP396
121. PubchemFP340	147. PubchemFP371	173. PubchemFP397
122. PubchemFP341	148. PubchemFP372	174. PubchemFP398
123. PubchemFP342	149. PubchemFP373	175. PubchemFP399
124. PubchemFP344	150. PubchemFP374	176. PubchemFP400
125. PubchemFP345	151. PubchemFP375	177. PubchemFP403

178. PubchemFP404	204. PubchemFP435	230. PubchemFP464
179. PubchemFP405	205. PubchemFP436	231. PubchemFP465
180. PubchemFP406	206. PubchemFP437	232. PubchemFP466
181. PubchemFP407	207. PubchemFP438	233. PubchemFP467
182. PubchemFP408	208. PubchemFP439	234. PubchemFP469
183. PubchemFP409	209. PubchemFP440	235. PubchemFP470
184. PubchemFP411	210. PubchemFP441	236. PubchemFP471
185. PubchemFP412	211. PubchemFP442	237. PubchemFP472
186. PubchemFP413	212. PubchemFP443	238. PubchemFP473
187. PubchemFP414	213. PubchemFP445	239. PubchemFP474
188. PubchemFP416	214. PubchemFP446	240. PubchemFP475
189. PubchemFP417	215. PubchemFP447	241. PubchemFP476
190. PubchemFP418	216. PubchemFP448	242. PubchemFP477
191. PubchemFP419	217. PubchemFP449	243. PubchemFP480
192. PubchemFP420	218. PubchemFP450	244. PubchemFP481
193. PubchemFP421	219. PubchemFP451	245. PubchemFP482
194. PubchemFP422	220. PubchemFP452	246. PubchemFP483
195. PubchemFP423	221. PubchemFP453	247. PubchemFP484
196. PubchemFP425	222. PubchemFP454	248. PubchemFP485
197. PubchemFP427	223. PubchemFP456	249. PubchemFP486
198. PubchemFP428	224. PubchemFP457	250. PubchemFP487
199. PubchemFP429	225. PubchemFP458	251. PubchemFP489
200. PubchemFP430	226. PubchemFP459	252. PubchemFP490
201. PubchemFP431	227. PubchemFP460	253. PubchemFP493
202. PubchemFP432	228. PubchemFP461	254. PubchemFP494
203. PubchemFP434	229. PubchemFP462	255. PubchemFP495

256. PubchemFP497	282. PubchemFP537	308. PubchemFP572
257. PubchemFP498	283. PubchemFP538	309. PubchemFP573
258. PubchemFP499	284. PubchemFP541	310. PubchemFP574
259. PubchemFP500	285. PubchemFP543	311. PubchemFP575
260. PubchemFP501	286. PubchemFP544	312. PubchemFP577
261. PubchemFP504	287. PubchemFP545	313. PubchemFP578
262. PubchemFP505	288. PubchemFP547	314. PubchemFP579
263. PubchemFP506	289. PubchemFP548	315. PubchemFP580
264. PubchemFP507	290. PubchemFP549	316. PubchemFP581
265. PubchemFP508	291. PubchemFP550	317. PubchemFP582
266. PubchemFP509	292. PubchemFP551	318. PubchemFP583
267. PubchemFP514	293. PubchemFP552	319. PubchemFP584
268. PubchemFP515	294. PubchemFP553	320. PubchemFP585
269. PubchemFP516	295. PubchemFP554	321. PubchemFP586
270. PubchemFP517	296. PubchemFP555	322. PubchemFP588
271. PubchemFP518	297. PubchemFP556	323. PubchemFP589
272. PubchemFP519	298. PubchemFP558	324. PubchemFP591
273. PubchemFP520	299. PubchemFP560	325. PubchemFP592
274. PubchemFP523	300. PubchemFP563	326. PubchemFP593
275. PubchemFP524	301. PubchemFP564	327. PubchemFP594
276. PubchemFP530	302. PubchemFP565	328. PubchemFP595
277. PubchemFP531	303. PubchemFP566	329. PubchemFP596
278. PubchemFP533	304. PubchemFP567	330. PubchemFP597
279. PubchemFP534	305. PubchemFP568	331. PubchemFP598
280. PubchemFP535	306. PubchemFP569	332. PubchemFP599
281. PubchemFP536	307. PubchemFP570	333. PubchemFP600

334. PubchemFP601	360. PubchemFP629	386. PubchemFP660
335. PubchemFP602	361. PubchemFP630	387. PubchemFP661
336. PubchemFP603	362. PubchemFP632	388. PubchemFP662
337. PubchemFP604	363. PubchemFP633	389. PubchemFP664
338. PubchemFP605	364. PubchemFP634	390. PubchemFP665
339. PubchemFP606	365. PubchemFP636	391. PubchemFP666
340. PubchemFP607	366. PubchemFP637	392. PubchemFP668
341. PubchemFP608	367. PubchemFP638	393. PubchemFP669
342. PubchemFP609	368. PubchemFP639	394. PubchemFP670
343. PubchemFP610	369. PubchemFP640	395. PubchemFP671
344. PubchemFP611	370. PubchemFP641	396. PubchemFP673
345. PubchemFP612	371. PubchemFP642	397. PubchemFP674
346. PubchemFP613	372. PubchemFP644	398. PubchemFP675
347. PubchemFP614	373. PubchemFP645	399. PubchemFP676
348. PubchemFP615	374. PubchemFP647	400. PubchemFP677
349. PubchemFP616	375. PubchemFP648	401. PubchemFP678
350. PubchemFP618	376. PubchemFP650	402. PubchemFP679
351. PubchemFP619	377. PubchemFP651	403. PubchemFP680
352. PubchemFP620	378. PubchemFP652	404. PubchemFP681
353. PubchemFP621	379. PubchemFP653	405. PubchemFP682
354. PubchemFP622	380. PubchemFP654	406. PubchemFP683
355. PubchemFP623	381. PubchemFP655	407. PubchemFP684
356. PubchemFP624	382. PubchemFP656	408. PubchemFP685
357. PubchemFP625	383. PubchemFP657	409. PubchemFP686
358. PubchemFP626	384. PubchemFP658	410. PubchemFP687
359. PubchemFP628	385. PubchemFP659	411. PubchemFP688

412. PubchemFP689	438. PubchemFP715	464. PubchemFP756
413. PubchemFP690	439. PubchemFP716	465. PubchemFP757
414. PubchemFP691	440. PubchemFP717	466. PubchemFP758
415. PubchemFP692	441. PubchemFP719	467. PubchemFP759
416. PubchemFP693	442. PubchemFP721	468. PubchemFP761
417. PubchemFP694	443. PubchemFP722	469. PubchemFP762
418. PubchemFP695	444. PubchemFP725	470. PubchemFP763
419. PubchemFP696	445. PubchemFP728	471. PubchemFP764
420. PubchemFP697	446. PubchemFP729	472. PubchemFP766
421. PubchemFP698	447. PubchemFP733	473. PubchemFP767
422. PubchemFP699	448. PubchemFP734	474. PubchemFP770
423. PubchemFP700	449. PubchemFP735	475. PubchemFP771
424. PubchemFP701	450. PubchemFP736	476. PubchemFP772
425. PubchemFP702	451. PubchemFP737	477. PubchemFP776
426. PubchemFP703	452. PubchemFP738	478. PubchemFP777
427. PubchemFP704	453. PubchemFP740	479. PubchemFP778
428. PubchemFP705	454. PubchemFP742	480. PubchemFP779
429. PubchemFP706	455. PubchemFP743	481. PubchemFP780
430. PubchemFP707	456. PubchemFP745	482. PubchemFP782
431. PubchemFP708	457. PubchemFP746	483. PubchemFP784
432. PubchemFP709	458. PubchemFP747	484. PubchemFP785
433. PubchemFP710	459. PubchemFP749	485. PubchemFP788
434. PubchemFP711	460. PubchemFP750	486. PubchemFP791
435. PubchemFP712	461. PubchemFP751	487. PubchemFP792
436. PubchemFP713	462. PubchemFP752	488. PubchemFP796
437. PubchemFP714	463. PubchemFP755	489. PubchemFP797

490. PubchemFP798	501. PubchemFP813	512. PubchemFP827
491. PubchemFP799	502. PubchemFP814	513. PubchemFP829
492. PubchemFP800	503. PubchemFP815	514. PubchemFP830
493. PubchemFP801	504. PubchemFP818	515. PubchemFP833
494. PubchemFP803	505. PubchemFP819	516. PubchemFP834
495. PubchemFP805	506. PubchemFP820	517. PubchemFP835
496. PubchemFP806	507. PubchemFP821	518. PubchemFP839
497. PubchemFP808	508. PubchemFP822	519. PubchemFP840
498. PubchemFP809	509. PubchemFP824	520. PubchemFP860
499. PubchemFP810	510. PubchemFP825	521. PubchemFP861
500. PubchemFP812	511. PubchemFP826	

Anexo B

Feature Selection

B.1 Regressão

B.1.1 Correlação de Pearson

Atributo	Correlação	Atributo	Correlação	Atributo	Correlação
PubchemFP581	-0.506691897	Zagreb	-0.384958403	nHBa	-0.316312005
PubchemFP642	-0.506691897	nC	-0.382776605	PubchemFP346	-0.313919656
SPC-6	-0.478489571	ATSp1	-0.380629582	ETA_Eta_F	-0.31339203
MDEC-34	-0.472921581	PubchemFP380	-0.37994463	hmin	-0.313100936
SPC-5	-0.469183584	VABC	-0.379671296	PubchemFP686	-0.31225779
ETA_EtaP	-0.464741012	VC-3	-0.379293846	PubchemFP366	-0.31010507
SPC-4	-0.452870369	C3SP3	-0.377750492	nT8Ring	-0.309118289
SC-5	-0.444644411	McGowan_Volume	-0.376489481	SC-6	-0.307540554
ATSp5	-0.443384796	VP-2	-0.374571056	MDEO-11	-0.305441905
WPOL	-0.441100205	PubchemFP308	-0.374394768	C1SP3	-0.304429228
VPC-5	-0.439681413	SP-0	-0.373490631	ETA_EtaP_L	-0.303929491
MDEC-13	-0.437787333	apol	-0.371847837	ETA_dBetaP	-0.303350296
VPC-4	-0.435838714	PubchemFP452	-0.371245655	nHBacc_Lipinski	-0.302673074
nssssC	-0.435700522	nBonds2	-0.371245655	VAdjMat	-0.300302631
nBondsS3	-0.433936738	nBonds	-0.370460316	C4SP3	-0.300134556
nBondsS2	-0.433936738	PubchemFP20	-0.370459782	nHBacc	-0.29928344
ATSp4	-0.433289012	SP-2	-0.370072081	VCH-7	-0.296168194
VPC-6	-0.432884531	nHeavyAtom	-0.369240007	ndO	-0.293530726
VC-5	-0.426796495	VP-0	-0.368115982	nHBacc2	-0.293173484
ATSm5	-0.422049345	WTPT-1	-0.367278615	MLFER_L	-0.293121176
ETA_Eta	-0.421630594	PubchemFP335	-0.365775751	ATSm2	-0.293121176
ETA_Eta_L	-0.417512008	MDEC-11	-0.365305393	Kier2	-0.292939193
MDEC-44	-0.412640771	SP-1	-0.365305265	n8Ring	-0.281706739
ETA_dBeta	-0.408997332	ETA_Eta_R_L	-0.363586397	PubchemFP227	-0.278336213
nBondsS	-0.408374022	MDEC-12	-0.36256865	nBondsD2	-0.268290873
WTPT-4	-0.407710879	bpol	-0.362248868	PubchemFP519	-0.266752246
nO	-0.405641712	VP-1	-0.359789167	SCH-7	-0.265598209
ATSp3	-0.403815082	PubchemFP406	-0.359381486	ETA_BetaP_ns	-0.263007537
SP-6	-0.402332995	ETA_Beta_s	-0.358476938	PubchemFP701	-0.260427309
ATSm4	-0.402039973	PubchemFP567	-0.357623308	PubchemFP375	-0.258387767
SP-3	-0.401407789	Kier1	-0.356024897	PubchemFP191	-0.257032783
VP-4	-0.400553346	PubchemFP639	-0.354778473	PubchemFP405	-0.256278942
VP-6	-0.398527506	PubchemFP21	-0.35356544	nG12Ring	-0.255166633
SP-5	-0.397538262	ndssC	-0.352970823	nF8Ring	-0.254042463
SP-4	-0.396362036	ETA_Alpha	-0.352101332	ETA_EtaP_F_L	-0.250765421
MDEO-12	-0.395010722	VC-6	-0.350443266	PubchemFP690	-0.249278909
VP-3	-0.394384344	SC-3	-0.349965496	PubchemFP487	-0.249057351
VP-5	-0.394314445	PubchemFP662	-0.349735446	PubchemFP450	-0.247770048
MDEC-24	-0.394056813	PubchemFP440	-0.349202466	LipinskiFailures	-0.247280912
ATSp2	-0.392384209	MLogP	-0.34899465	VCH-6	-0.247280912
VP-7	-0.391948169	MDEC-33	-0.348860597	PubchemFP699	-0.243999419
ETA_Eta_B_RC	-0.391784192	ETA_Eta_B	-0.347908021	PubchemFP680	-0.242880142
PubchemFP337	-0.391494632	MW	-0.340144206	nRing	-0.242439178
ETA_Eta_R	-0.391449857	ATSm3	-0.339638462	BCUTp-1h	-0.241795163
SP-7	-0.390337728	PubchemFP339	-0.338683562	MDEO-22	-0.241616087
MDEC-14	-0.389711718	PubchemFP681	-0.333597096	MDEN-13	-0.237190179
nAtom	-0.389281542	PubchemFP13	-0.326196361	PubchemFP638	-0.233855929
nH	-0.388262359	nssO	-0.322249632	hmax	-0.233790956
C2SP3	-0.38743702	AMR	-0.320668279	PubchemFP228	-0.233790956
PubchemFP537	-0.38743702	nHBacc3	-0.319071912	PubchemFP132	-0.233790956
PubchemFP462	-0.387243963	fragC	-0.316950319	VC-4	-0.233790956

Atributo	Correlação	Atributo	Correlação	Atributo	Correlação
PubchemFP560	-0.233790956	PubchemFP782	-0.175828732	PubchemFP568	-0.097184703
ECCEN	-0.233368568	PubchemFP352	-0.175828732	PubchemFP466	-0.097038433
PubchemFP338	-0.233172568	PubchemFP684	-0.175212282	PubchemFP429	-0.097038433
PubchemFP614	-0.232083274	nFG12Ring	-0.174817373	PubchemFP419	-0.096804617
nAtomLAC	-0.231956316	nHBDon	-0.174817373	C1SP2	-0.09647371
PubchemFP637	-0.231201717	nF9Ring	-0.174817373	PubchemFP443	-0.09647371
WTPT-3	-0.228154505	PubchemFP484	-0.174817373	PubchemFP22	-0.095112917
PubchemFP582	-0.228154505	ATSc1	-0.172701792	PubchemFP38	-0.094767753
PubchemFP441	-0.227735143	PubchemFP861	-0.171819049	PubchemFP118	-0.09435573
PubchemFP416	-0.227413296	PubchemFP840	-0.171819049	SC-4	-0.094249089
PubchemFP332	-0.22666309	PubchemFP214	-0.165714134	PubchemFP632	-0.092141877
PubchemFP11	-0.226589331	BCUTc-1h	-0.165589017	PubchemFP620	-0.092141877
PubchemFP798	-0.223729182	ndsssP	-0.165080053	nsCl	-0.092083186
TopoPSA	-0.223256092	nP	-0.164905998	nCl	-0.091711494
PubchemFP446	-0.22161625	PubchemFP456	-0.164581189	PubchemFP16	-0.090368785
nTRing	-0.219860966	PubchemFP425	-0.162638987	nAtomP	-0.088826571
ETA_BetaP	-0.21926085	PubchemFP411	-0.161341965	PubchemFP708	-0.087721215
PubchemFP579	-0.218295777	PubchemFP407	-0.160847072	PubchemFP659	-0.087491286
PubchemFP666	-0.218097653	PubchemFP314	-0.160498459	PubchemFP459	-0.085465726
PubchemFP594	-0.217795518	PubchemFP30	-0.158544772	PubchemFP457	-0.084651898
MLFER_BO	-0.215425886	PubchemFP692	-0.154921248	PubchemFP414	-0.084449389
WPATH	-0.213012888	PubchemFP700	-0.154645566	nBondsT	-0.083409802
PubchemFP689	-0.212331681	ATSc2	-0.154485517	ETA_dAlpha_B	-0.08319442
PubchemFP693	-0.211224951	nT7Ring	-0.154485517	ETA_dEpsilon_B	-0.08319442
PubchemFP157	-0.210654821	n7Ring	-0.154485517	MLFER_S	-0.083156752
PubchemFP341	-0.210633715	PubchemFP213	-0.154007872	nddssS	-0.082913489
ETA_Beta	-0.210633715	PubchemFP641	-0.151332574	PubchemFP647	-0.079789521
PubchemFP698	-0.209315135	nsB	-0.151332574	PubchemFP707	-0.079778266
PubchemFP535	-0.208534451	nB	-0.149999299	PubchemFP822	-0.079289645
MLFER_BH	-0.208478441	PubchemFP6	-0.149635235	PubchemFP759	-0.077993919
nF10Ring	-0.208478441	PubchemFP276	-0.149551182	PubchemFP473	-0.077664449
PubchemFP334	-0.203934664	PubchemFP274	-0.147995074	PubchemFP820	-0.077477881
PubchemFP636	-0.203313075	PubchemFP797	-0.147849917	PubchemFP757	-0.077216338
MDEC-23	-0.202884023	PubchemFP640	-0.143962631	PubchemFP33	-0.074059294
PubchemFP596	-0.201652301	PubchemFP556	-0.143669644	PubchemFP293	-0.073675625
PubchemFP824	-0.201406438	PubchemFP524	-0.140196038	maxHBa	-0.073675625
PubchemFP619	-0.199481989	PubchemFP490	-0.13961206	C3SP2	-0.073270589
PubchemFP674	-0.199481989	ATSc3	-0.137949416	PubchemFP826	-0.072395848
nT10Ring	-0.19579455	ETA_Beta_ns_d	-0.137396045	ETA_EtaP_F	-0.072395848
PubchemFP183	-0.192838643	PubchemFP563	-0.136546741	PubchemFP374	-0.071999326
nT11Ring	-0.19268053	SCH-4	-0.136128726	nS	-0.068097746
nF11Ring	-0.192612158	SCH-5	-0.135214646	PubchemFP474	-0.067076559
MDEN-11	-0.192425568	SCH-4	-0.135048508	PubchemFP709	-0.065027948
PubchemFP573	-0.191809104	PubchemFP566	-0.13402141	PubchemFP572	-0.064970044
nFRing	-0.191474678	PubchemFP420	-0.133975872	PubchemFP504	-0.064970044
PubchemFP344	-0.191440491	PubchemFP299	-0.130320339	PubchemFP389	-0.064278772
PubchemFP284	-0.191224013	PubchemFP184	-0.129163001	PubchemFP34	-0.064185307
PubchemFP697	-0.191224013	nWHBa	-0.129137124	nBase	-0.063919951
PubchemFP706	-0.191224013	PubchemFP193	-0.127564	PubchemFP538	-0.062988827
SCH-6	-0.190537465	ATSm1	-0.126777157	ETA_Shape_Y	-0.061297491
BCUTc-1l	-0.190223613	PubchemFP679	-0.125143815	PubchemFP353	-0.060868118
PubchemFP434	-0.190223613	ETA_AlphaP	-0.125101641	PubchemFP152	-0.060827048
PubchemFP704	-0.188409159	VCH-5	-0.123710535	PubchemFP776	-0.060827048
ETA_BetaP_ns_d	-0.187877525	C1SP1	-0.122904666	PubchemFP545	-0.060543404
PubchemFP379	-0.187877525	PubchemFP799	-0.122335545	PubchemFP800	-0.060543404
PubchemFP160	-0.187877525	PubchemFP736	-0.121912673	PubchemFP737	-0.060236256
PubchemFP159	-0.187877525	PubchemFP465	-0.120515628	PubchemFP259	-0.060101004
PubchemFP12	-0.186889947	PubchemFP438	-0.119704467	ndsN	-0.057130301
Kier3	-0.186844104	ATSc5	-0.119518674	PubchemFP333	-0.056835874
ETA_Eta_F_L	-0.186310676	PubchemFP15	-0.11855534	PubchemFP778	-0.05531777
PubchemFP18	-0.185725356	PubchemFP705	-0.11653471	PubchemFP715	-0.053282447
PubchemFP286	-0.184480318	PubchemFP19	-0.115905511	VCH-4	-0.052744683
nT9Ring	-0.183398593	PubchemFP447	-0.115271763	PubchemFP397	-0.051111999
PubchemFP711	-0.182467111	PubchemFP233	-0.107380766	nT6Ring	-0.050270677
nTG12Ring	-0.180657819	PubchemFP232	-0.103239562	PubchemFP696	-0.047435887
nT4Ring	-0.180043604	PubchemFP190	-0.102962586	PubchemFP645	-0.047424216
n4Ring	-0.180043604	PubchemFP373	-0.102811944	PubchemFP599	-0.046359502
PubchemFP129	-0.180043604	nAcid	-0.101859647	ETA_Shape_X	-0.046359502
ATSc4	-0.179670661	PubchemFP777	-0.101100168	nssS	-0.042463231
nT12Ring	-0.179138164	ALogP	-0.100900536	PubchemFP182	-0.041315118
nF12Ring	-0.179132001	PubchemFP393	-0.100308609	ETA_dAlpha_A	-0.041315118
PubchemFP432	-0.1777844	PubchemFP378	-0.099899805	PubchemFP531	-0.041196762
n9Ring	-0.1777844	PubchemFP678	-0.099899805	PubchemFP206	-0.039427629
PubchemFP247	-0.1777844	PubchemFP660	-0.099899805	nAtomLC	-0.03934897
PubchemFP246	-0.17725785	maxdO	-0.099569123	PubchemFP178	-0.038102482
PubchemFP241	-0.175828732	ntN	-0.099268198	PubchemFP534	-0.038102482
PubchemFP654	-0.175828732	PubchemFP676	-0.099253978	n6Ring	-0.038102482
nRotB	-0.175828732	PubchemFP568	-0.097184703	ntsC	-0.038102482

Atributo	Correlação	Atributo	Correlação	Atributo	Correlação
naasC	-0.038102482	PubchemFP751	-0.003730961	PubchemFP261	0.011471155
PubchemFP37	-0.037504151	PubchemFP733	-0.003427206	PubchemFP592	0.011471155
PubchemFP342	-0.035589371	PubchemFP675	-0.003110813	PubchemFP805	0.011471155
PubchemFP294	-0.034410263	PubchemFP670	-0.002688895	PubchemFP742	0.011701917
PubchemFP418	-0.034171061	PubchemFP554	-0.001942818	PubchemFP671	0.012154496
PubchemFP819	-0.034171061	PubchemFP509	-0.001539609	PubchemFP400	0.012194341
nHBDOn_Lipinski	-0.032771593	PubchemFP44	-0.001485828	PubchemFP494	0.012355313
PubchemFP598	-0.03267448	PubchemFP43	-0.001423047	PubchemFP386	0.012405194
PubchemFP591	-0.030460227	PubchemFP330	-0.000743051	PubchemFP477	0.012451949
PubchemFP550	-0.029965616	PubchemFP328	-0.000343263	PubchemFP435	0.01294571
PubchemFP501	-0.02973285	PubchemFP327	-0.000159509	PubchemFP372	0.013131795
PubchemFP803	-0.029495923	PubchemFP297	0.000102266	PubchemFP451	0.013242593
PubchemFP383	-0.028179304	ETA_Epsilon_4	0.000192367	PubchemFP256	0.013541255
PubchemFP362	-0.027808384	nN	0.000217811	CDKN2A_SV	0.013541255
PubchemFP815	-0.027032491	MLFER_E	0.000217811	ETA_EtaP_B	0.013649009
PubchemFP752	-0.027032491	nAromBond	0.000548342	PubchemFP779	0.013693025
PubchemFP657	-0.027032491	PubchemFP180	0.000757622	PubchemFP716	0.01378052
PubchemFP585	-0.02652161	PubchemFP523	0.000845108	PubchemFP395	0.013863364
PubchemFP734	-0.026369877	PubchemFP549	0.001016383	PubchemFP301	0.014072432
PubchemFP665	-0.026074932	PubchemFP516	0.001253765	maxdssC	0.014101775
PubchemFP600	-0.025934924	PubchemFP839	0.001544467	PubchemFP668	0.014238894
PubchemFP710	-0.025406332	PubchemFP189	0.001845359	PubchemFP634	0.014244057
nessN	-0.025080459	PubchemFP695	0.001914797	PubchemFP618	0.01433725
ETA_Epsilon_3	-0.02417167	PubchemFP703	0.002328679	PubchemFP603	0.014389636
PubchemFP533	-0.02417167	PubchemFP825	0.002349841	PubchemFP578	0.014608719
PubchemFP453	-0.023713555	PubchemFP762	0.002628855	PubchemFP552	0.014608719
PubchemFP347	-0.022819574	PubchemFP469	0.002717945	PubchemFP371	0.014608719
ETA_dPsi_A	-0.022267457	PubchemFP195	0.002856847	PubchemFP399	0.014608719
ETA_Psi_1	-0.022244481	PubchemFP664	0.00339078	PubchemFP548	0.01469531
PubchemFP687	-0.021544505	PubchemFP564	0.003577059	MLFER_A	0.014762233
maxdsN	-0.020811513	MDEN-33	0.003626588	PubchemFP186	0.014863878
PubchemFP156	-0.02058379	PubchemFP644	0.0038111	PubchemFP755	0.015273514
PubchemFP155	-0.019048795	PubchemFP467	0.004164702	PubchemFP355	0.015723253
naasN	-0.017991681	PubchemFP475	0.004656951	PubchemFP255	0.015911724
PubchemFP514	-0.017816701	PubchemFP792	0.004946016	maxaasN	0.01599894
PubchemFP605	-0.017744576	PubchemFP729	0.005449635	PubchemFP390	0.016018738
PubchemFP430	-0.017744576	PubchemFP471	0.005588843	PubchemFP388	0.016439517
PubchemFP780	-0.017744576	PubchemFP625	0.005675983	SCH-3	0.016607193
PubchemFP717	-0.017038707	PubchemFP655	0.005716689	n10Ring	0.0168833
PubchemFP650	-0.016341416	PubchemFP530	0.005746497	PubchemFP252	0.017225199
PubchemFP624	-0.016303516	PubchemFP749	0.006127	PubchemFP248	0.017296101
PubchemFP810	-0.015927838	MDEC-22	0.006140378	PubchemFP262	0.017363832
PubchemFP747	-0.015927838	PubchemFP621	0.006143827	MDEN-22	0.017652409
PubchemFP669	-0.015585011	PubchemFP167	0.006271652	PubchemFP442	0.017744418
PubchemFP612	-0.015548016	PubchemFP345	0.006689126	BCUTw-1h	0.017806445
nF6Ring	-0.015251938	PubchemFP682	0.006941658	CDKN2A_CN	0.01799045
PubchemFP117	-0.014469129	PubchemFP860	0.007091033	CDKN2a(p14)_SV	0.018170012
MDEN-12	-0.01379832	PubchemFP829	0.007275652	PubchemFP258	0.018593401
PubchemFP712	-0.013798254	PubchemFP766	0.007275652	CDKN2a(p14)_CN	0.018759966
PubchemFP481	-0.012630937	ETA_dEpsilon_C	0.00729242	PubchemFP574	0.018949477
PubchemFP458	-0.012626452	PubchemFP507	0.008011871	PubchemFP194	0.019032751
PubchemFP408	-0.01252068	PubchemFP412	0.00810638	PubchemFP608	0.019211188
nT5Ring	-0.01252068	PubchemFP359	0.00810638	PubchemFP595	0.01930379
n5Ring	-0.01250879	PubchemFP586	0.008585941	PubchemFP570	0.019925071
PubchemFP580	-0.011997859	PubchemFP351	0.008768127	PubchemFP384	0.019925071
naAromAtom	-0.011297716	BCUTp-1l	0.008926824	PubchemFP364	0.020350675
nBondsD	-0.011297716	n3Ring	0.009112354	PubchemFP713	0.020649818
ETA_EtaP_B_RC	-0.011297716	PubchemFP115	0.009334965	PubchemFP761	0.020754098
PubchemFP199	-0.011297716	PubchemFP784	0.009545703	PubchemFP630	0.021022639
PubchemFP673	-0.011297716	PubchemFP518	0.009621562	PubchemFP544	0.021300224
PubchemFP219	-0.011029902	PubchemFP25	0.009796133	PubchemFP536	0.021300288
PubchemFP218	-0.010660642	PubchemFP610	0.009891331	PubchemFP493	0.021376823
PubchemFP192	-0.009977727	PubchemFP367	0.010118172	PubchemFP398	0.021617233
PetitjeanNumber	-0.009773531	PubchemFP633	0.010178481	PubchemFP394	0.021905878
maxsF	-0.009191724	WTPT-2	0.010411677	PubchemFP482	0.022428428
PubchemFP809	-0.008829678	maxsCI	0.010412076	PubchemFP396	0.022428428
PubchemFP694	-0.008250968	BCUTw-1l	0.010423318	PubchemFP387	0.022545923
PubchemFP515	-0.007143937	PubchemFP377	0.010423318	PubchemFP358	0.022719845
PubchemFP300	-0.007056926	PubchemFP495	0.010423318	PubchemFP609	0.022793561
PubchemFP601	-0.006798176	PubchemFP147	0.010423318	PubchemFP555	0.023258172
PubchemFP376	-0.006605453	PubchemFP500	0.010589292	PubchemFP812	0.023258172
PubchemFP340	-0.006555061	PubchemFP454	0.010713437	PubchemFP508	0.023351816
nsBr	-0.006021363	PubchemFP423	0.010755973	PubchemFP356	0.023368197
nBr	-0.005462432	WTPT-5	0.010827993	ndS	0.023383358
PubchemFP835	-0.004778145	PubchemFP688	0.011138436	PubchemFP360	0.024105985
PubchemFP814	-0.004653145	PubchemFP740	0.011146596	mindS	0.024318677
PubchemFP796	-0.004122904	PubchemFP653	0.011312324	SdS	0.024318677
PubchemFP772	-0.003897097	PubchemFP368	0.011431221	PubchemFP445	0.024927594

Atributo	Correlação	Atributo	Correlação	Atributo	Correlação
PubchemFP746	0.026876031	PubchemFP615	0.05709556	VHL_CN	0.088725089
PubchemFP428	0.026876031	PubchemFP164	0.05709556	PHOX2B_CN	0.088725089
PubchemFP658	0.027136671	PubchemFP464	0.057186924	PALB2_CN	0.088725089
PubchemFP623	0.028850705	naaaC	0.057384099	HRAS_CN	0.088725089
PubchemFP179	0.028850705	ALogp2	0.058020284	PubchemFP791	0.088854701
naaO	0.028850705	PubchemFP130	0.058850594	TSC1_CN	0.089728962
PubchemFP188	0.0290845	PubchemFP833	0.058850594	PubchemFP575	0.089839927
PubchemFP305	0.02913435	PubchemFP770	0.05885312	MLH1_CN	0.089839927
PubchemFP813	0.029530428	PubchemFP801	0.059399637	APC_CN	0.09091275
PubchemFP750	0.029530428	PubchemFP738	0.060087043	KRAS_SV	0.091627986
PubchemFP143	0.030263633	PubchemFP404	0.060262033	PDGFRA_CN	0.092760747
PubchemFP685	0.030772792	PubchemFP652	0.060262033	FAM123B_SV	0.094986201
PubchemFP656	0.030772792	FGFR2_CN	0.060262033	PubchemFP146	0.096513372
PubchemFP604	0.030772792	PubchemFP499	0.060262033	NOTCH1_SV	0.096513372
PubchemFP756	0.031007655	PubchemFP365	0.060262033	BRCA1_SV	0.096513372
PubchemFP336	0.031007655	PubchemFP648	0.060262033	RET_CN	0.097008636
nsI	0.031007655	PubchemFP558	0.060262033	DICER1_CN	0.097257273
nl	0.031007655	PubchemFP17	0.060501152	PubchemFP834	0.097257273
PubchemFP46	0.031007655	PubchemFP721	0.061560655	PubchemFP771	0.097257273
PubchemFP350	0.032032903	PubchemFP622	0.063079519	MLLT3_CN	0.098285446
PubchemFP298	0.032553027	PubchemFP285	0.06323594	RB1_SV	0.098436981
PubchemFP628	0.032573923	nsF	0.063246712	NF2_CN	0.098911754
PubchemFP607	0.032704037	nF	0.064120549	TET2_SV	0.098911754
PubchemFP153	0.032741892	PubchemFP498	0.065236669	VHL_SV	0.099634069
PubchemFP409	0.032801617	PubchemFP489	0.065380755	PubchemFP821	0.100900536
MYCN_CN	0.033638179	MYCL1_CN	0.065945036	PubchemFP758	0.101519276
maxaaaC	0.033638179	PIK3R1_SV	0.065945036	PubchemFP149	0.103603199
PubchemFP788	0.034289768	PubchemFP728	0.067317632	PIK3R1_CN	0.103603199
PubchemFP725	0.034768486	PubchemFP349	0.067459833	SMO_CN	0.108316801
PubchemFP719	0.035080562	maxwHBa	0.067459833	PubchemFP702	0.108316801
PubchemFP616	0.035145761	PubchemFP577	0.067490664	CYLD_SV	0.108563528
PubchemFP422	0.035377036	CDK4_CN	0.068329059	FBXW7_CN	0.108563528
naaN	0.035377036	PubchemFP363	0.0685715	IDH1_SV	0.10958579
maxaaN	0.035903609	SETD2_CN	0.068613164	PTCH1_CN	0.113577115
PubchemFP257	0.035903609	PTEN_CN	0.069155546	PubchemFP460	0.115409396
PubchemFP150	0.036002718	NF2_SV	0.072388699	PubchemFP427	0.115409396
PubchemFP14	0.036694113	CCND3_CN	0.07251383	PubchemFP417	0.115409396
PubchemFP683	0.036790984	PubchemFP287	0.072652141	WT1_CN	0.115409396
PubchemFP613	0.037747878	PubchemFP23	0.073860446	SMARCB1_CN	0.117496013
PubchemFP806	0.037747878	NPM1_CN	0.074547575	PubchemFP651	0.117759602
PubchemFP743	0.038195579	PubchemFP391	0.074599228	PubchemFP606	0.117759602
PubchemFP181	0.039898975	PubchemFP200	0.074846913	PubchemFP541	0.117759602
PubchemFP553	0.040180793	BRCA1_CN	0.074846913	PubchemFP382	0.11809666
ERBB2_CN	0.041407498	PubchemFP144	0.075122264	PubchemFP381	0.118434022
PubchemFP629	0.042504059	MDM2_CN	0.0758707	PubchemFP187	0.11945023
VCH-3	0.043884799	SMAD4_CN	0.0758707	MLH1_SV	0.120526318
MDEN-23	0.044035547	CDH1_SV	0.079103275	CTNNB1_SV	0.121129671
PubchemFP485	0.04418479	PubchemFP480	0.079398766	CYLD_CN	0.12224988
NTRK3_CN	0.045060833	AKT2_CN	0.079508423	PubchemFP260	0.126764434
PubchemFP827	0.045500708	PubchemFP543	0.080186761	MSH2_CN	0.127265485
PubchemFP764	0.045613394	PubchemFP547	0.081260607	MSH6_SV	0.127265485
PubchemFP583	0.046359502	ETA_Epsilon_5	0.081260607	STK11_CN	0.127738174
PubchemFP551	0.046359502	ETA_Epsilon_2	0.08295191	PubchemFP677	0.128784533
PubchemFP505	0.046435753	ETA_Epsilon_1	0.08295191	PubchemFP584	0.128981675
PubchemFP626	0.048362876	STK11_SV	0.084207795	PubchemFP520	0.128981675
PubchemFP589	0.048473321	PubchemFP413	0.084207795	PubchemFP470	0.129061116
PubchemFP565	0.048815673	SOCs1_CN	0.084207795	SUFU_CN	0.130296228
PubchemFP735	0.048815673	MSH6_CN	0.085177095	FLCN_SV	0.130392177
maxsssN	0.048815673	BRCA2_SV	0.085177095	EP300_CN	0.130586393
PubchemFP461	0.048815673	PubchemFP602	0.086225242	KDR_CN	0.130923564
XLogP	0.049093976	NF1_SV	0.086663201	KDM6A_SV	0.133648449
PubchemFP785	0.049093976	PubchemFP116	0.086767683	JAK2_CN	0.133648449
PubchemFP722	0.049157237	GNAS_CN	0.08745825	FAM123B_CN	0.133648449
maxssO	0.050471033	FLT3_SV	0.087546231	SMAD4_SV	0.134539903
PubchemFP449	0.050471033	PubchemFP148	0.088725089	PubchemFP392	0.136568096
PubchemFP483	0.050471033	PubchemFP597	0.088725089	CTNNB1_CN	0.13666223
PubchemFP506	0.051778365	FLCN_CN	0.088725089	BRCA2_CN	0.13666223
PubchemFP472	0.052119338	PubchemFP448	0.088725089	KDM5C_SV	0.137040025
PubchemFP403	0.052119338	PubchemFP436	0.088725089	TSC2_CN	0.137040025
ETA_BetaP_s	0.052119338	MET_CN	0.088725089	TP53_CN	0.137759099
BRAF_SV	0.052119338	CCND2_CN	0.088725089	PubchemFP497	0.137759099
PubchemFP476	0.054208682	PubchemFP185	0.088725089	SMARCA4_CN	0.138705741
PubchemFP9	0.054755801	PubchemFP714	0.088725089	MSH2_SV	0.139522605
PubchemFP283	0.055584773	APC_SV	0.088725089	PubchemFP691	0.139727679
PubchemFP10	0.055688085	PubchemFP763	0.088725089	PubchemFP611	0.14269337
PubchemFP517	0.055914719	PubchemFP431	0.088725089	PubchemFP569	0.14271552
PubchemFP486	0.056826416	KDM5C_CN	0.088725089	NRAS_CN	0.14271552
naaS	0.05709556	TET2_CN	0.088725089	CDKN2C_SV	0.14271552

Atributo	Correlação	Atributo	Correlação	Atributo	Correlação
PubchemFP808	0.144429112	PubchemFP588	0.157878373	MYC_CN	0.258828572
PubchemFP745	0.145799586	PubchemFP421	0.159098642	PubchemFP24	0.262043371
maxaasC	0.145862336	NRAS_SV	0.164872418	PubchemFP385	0.271660313
FGFR3_CN	0.145862336	ERBB2_SV	0.171854303	PubchemFP830	0.281343248
GNAS_SV	0.14706112	SETD2_SV	0.178550547	PubchemFP767	0.288844935
MAP2K4_SV	0.14706112	RUNX1_CN	0.179175132	JAK2_SV	0.304546952
EGFR_SV	0.148822169	EZH2_CN	0.179175132	EZH2_SV	0.313770352
KRAS_CN	0.150248177	FBXW7_SV	0.179175132	PIK3CA_CN	0.410718528
PTEN_SV	0.150248177	HRAS_SV	0.179175132	WT1_SV	NaN
FGFR3_SV	0.150248177	PubchemFP437	0.179175132	TSC2_SV	NaN
BRAF_CN	0.150248177	IDH1_CN	0.179175132	SUFU_SV	NaN
CDKN2C_CN	0.150248177	KIT_CN	0.179175132	SOCS1_SV	NaN
PubchemFP145	0.150248177	NOTCH1_CN	0.179175132	SMO_SV	NaN
ALK_SV	0.151065123	MIS	0.181385024	RUNX1_SV	NaN
C2SP2	0.152004293	TP53_SV	0.187104531	RET_SV	NaN
CDK6_CN	0.152957977	SMARCA4_SV	0.200263711	PTCH1_SV	NaN
ALK_CN	0.153858576	ETA_Beta_ns	0.200673631	PHOX2B_SV	NaN
PubchemFP818	0.15419452	FLT3_CN	0.202477768	PDGFRA_SV	NaN
MAP2K4_CN	0.154257212	CCND1_CN	0.209729216	PALB2_SV	NaN
EGFR_CN	0.154509641	NF1_CN	0.212124758	NTRK3_SV	NaN
PubchemFP439	0.155034404	PubchemFP593	0.213573946	NPM1_SV	NaN
ETA_Shape_P	0.155112157	KDM6A_CN	0.217206822	MYC_SV	NaN
C2SP1	0.155746255	RB1_CN	0.231223088	MET_SV	NaN
PIK3CA_SV	0.156354968	PubchemFP370	0.243678948	KIT_SV	NaN
ETA_dEpsilon_A	0.157163959	PubchemFP357	0.249227118	KDR_SV	NaN
CDH1_CN	0.157163959	TSC1_SV	0.258828572	DICER1_SV	NaN

B.1.2 Ganho de informação mútua

Atributo	Ganho	Atributo	Ganho	Atributo	Ganho
APC_CN	0	WT1_CN	0	SOCS1_SV	0.003117852
APC_SV	0	WT1_SV	0	PubchemFP722	0.00317045
BRAF_CN	0	PubchemFP460	0.000100332	TP53_SV	0.00319092
BRAF_SV	0	SOCS1_CN	0.000107953	CDKN2C_CN	0.003193751
BRCA1_CN	0	FLCN_SV	0.000172734	PubchemFP461	0.003207054
BRCA2_SV	0	DICER1_CN	0.000193203	maxaaaC	0.003210341
CCND2_CN	0	CCND1_CN	0.00026166	PubchemFP575	0.003215299
CCND3_CN	0	MLLT3_CN	0.000306604	PubchemFP115	0.003295589
CDH1_CN	0	PTEN_CN	0.00031123	BRCA1_SV	0.003297507
CDK4_CN	0	BRCA2_CN	0.000342573	HRAS_SV	0.003388682
CYLD_CN	0	PubchemFP738	0.000387655	SdS	0.003433551
CYLD_SV	0	PubchemFP588	0.000486678	PubchemFP360	0.003531995
DICER1_SV	0	EGFR_CN	0.000503474	PubchemFP417	0.00353676
EGFR_SV	0	AKT2_CN	0.000525699	n10Ring	0.003584467
EP300_CN	0	KRAS_SV	0.000535667	PubchemFP436	0.003586773
ERBB2_CN	0	KDM5C_CN	0.000566263	SETD2_SV	0.003628282
EZH2_CN	0	TP53_CN	0.000651385	PubchemFP252	0.003668791
FBXW7_CN	0	NTRK3_SV	0.000680771	MSH2_SV	0.003676863
FBXW7_SV	0	ALK_CN	0.000701553	PubchemFP505	0.003821727
FGFR2_CN	0	FGFR3_SV	0.000706656	NTRK3_CN	0.003923846
FGFR3_CN	0	CDKN2a(p14)_CN	0.000719417	PubchemFP497	0.003955385
FLCN_CN	0	C2SP1	0.000721648	PubchemFP143	0.004123406
FLT3_SV	0	SMARCA4_CN	0.000723406	PubchemFP788	0.004166098
GNAS_CN	0	PubchemFP558	0.000747451	PubchemFP145	0.004317184
HRAS_CN	0	VHL_SV	0.000784993	PubchemFP413	0.004340257
IDH1_CN	0	GNAS_SV	0.000787609	PubchemFP766	0.004365657
IDH1_SV	0	SMARCA4_SV	0.0008728	PubchemFP146	0.004390994
JAK2_CN	0	PubchemFP200	0.000890555	SMAD4_CN	0.004415481
JAK2_SV	0	PubchemFP14	0.000900417	CDKN2A_CN	0.004459938
KDM5C_SV	0	KIT_CN	0.000916637	FAM123B_CN	0.00453218
KDM6A_SV	0	CTNNB1_SV	0.00104565	PubchemFP806	0.004576827
KDR_SV	0	SUFU_SV	0.001053064	PubchemFP483	0.004681918
KRAS_CN	0	KDM6A_CN	0.001055608	PubchemFP551	0.004897912
MAP2K4_CN	0	TET2_SV	0.001070475	PubchemFP10	0.004933909
MAP2K4_SV	0	PubchemFP771	0.001071251	PubchemFP117	0.00501893
MDM2_CN	0	ERBB2_SV	0.001097706	PubchemFP428	0.0050574
MET_CN	0	PDGFRA_CN	0.001115551	PubchemFP421	0.005066966
MET_SV	0	SETD2_CN	0.001139057	PubchemFP164	0.005126608
MIS	0	RUNX1_SV	0.00125641	PubchemFP404	0.005191671
MLH1_CN	0	CDK6_CN	0.001360842	PubchemFP827	0.005280881
MLH1_SV	0	PubchemFP648	0.001475809	PubchemFP609	0.005310655
MSH2_CN	0	EZH2_SV	0.001489342	PubchemFP116	0.005315482
MSH6_CN	0	PIK3R1_CN	0.001510571	PubchemFP9	0.005356633
MSH6_SV	0	PIK3CA_CN	0.00152663	PubchemFP808	0.005401624
MYC_CN	0	RET_SV	0.00157533	CDKN2a(p14)_SV	0.00544525
MYC_SV	0	MYCN_CN	0.001600545	ndS	0.00559793
NF1_CN	0	PubchemFP743	0.001701154	PubchemFP480	0.005665845
NF1_SV	0	CDKN2A_SV	0.001757988	PubchemFP691	0.005787194
NF2_CN	0	PubchemFP448	0.001760216	PubchemFP569	0.005831249
NF2_SV	0	CDKN2C_SV	0.001763508	PubchemFP746	0.005884171
NOTCH1_SV	0	SMAD4_SV	0.001852159	PubchemFP702	0.006146904
NPM1_CN	0	naaO	0.001856212	PubchemFP508	0.006152394
NPM1_SV	0	NOTCH1_CN	0.002018709	n3Ring	0.006164544
PALB2_CN	0	CDH1_SV	0.002072707	PubchemFP423	0.00618309
PALB2_SV	0	PubchemFP390	0.002177042	PubchemFP285	0.006216692
PHOX2B_CN	0	KIT_SV	0.002200494	PubchemFP767	0.006287614
PHOX2B_SV	0	PubchemFP785	0.0023533	PubchemFP610	0.006531901
PIK3R1_SV	0	PubchemFP336	0.00236325	PubchemFP486	0.006579802
PTCH1_CN	0	PDGFRA_SV	0.002468475	PubchemFP489	0.006753509
PTCH1_SV	0	PubchemFP834	0.002487085	PubchemFP144	0.006774682
PTEN_SV	0	PubchemFP409	0.002560299	PubchemFP735	0.006795397
PubchemFP622	0	FLT3_CN	0.002562755	PubchemFP764	0.006799618
RB1_CN	0	CTNNB1_CN	0.002641575	PubchemFP381	0.006825116
RB1_SV	0	PubchemFP583	0.002646121	PubchemFP283	0.006922791
RET_CN	0	PubchemFP367	0.002784351	PubchemFP399	0.006942272
RUNX1_CN	0	PubchemFP745	0.002814031	PubchemFP611	0.006967814
SMARCB1_CN	0	PubchemFP130	0.002869127	PubchemFP167	0.007117065
SMO_CN	0	NRAS_CN	0.002878713	PubchemFP805	0.00712091
SMO_SV	0	PIK3CA_SV	0.002890981	PubchemFP725	0.007167242
STK11_CN	0	PubchemFP801	0.002997357	PubchemFP485	0.007173464
STK11_SV	0	mindS	0.003008127	PubchemFP829	0.007176077
SUFU_CN	0	NRAS_SV	0.003034636	PubchemFP473	0.007352073
TET2_CN	0	PubchemFP427	0.003070841	PubchemFP830	0.007412938
TSC1_CN	0	MYCL1_CN	0.003071887	PubchemFP155	0.00744949
TSC1_SV	0	PubchemFP517	0.003094504	PubchemFP770	0.00747575
TSC2_CN	0	KDR_CN	0.003099029	naaS	0.007477643
TSC2_SV	0	FAM123B_SV	0.00310296	PubchemFP553	0.007544244
VHL_CN	0	ALK_SV	0.003113308	PubchemFP248	0.00755422

Atributo	Ganho	Atributo	Ganho	Atributo	Ganho
PubchemFP670	0.007620654	PubchemFP199	0.01080706	PubchemFP655	0.0151577
PubchemFP493	0.007625456	PubchemFP150	0.010810865	PubchemFP386	0.015442962
nBr	0.007656566	PubchemFP752	0.010815389	PubchemFP368	0.015443104
PubchemFP606	0.007692154	PubchemFP574	0.010845122	PubchemFP397	0.015487133
PubchemFP17	0.00770355	PubchemFP330	0.010845326	PubchemFP300	0.015559258
PubchemFP365	0.007789901	PubchemFP687	0.010888045	PubchemFP6	0.015588482
PubchemFP629	0.007833234	PubchemFP630	0.010894655	PubchemFP403	0.015690484
PubchemFP400	0.007841775	PubchemFP431	0.011048262	PubchemFP412	0.015793855
PubchemFP809	0.007872896	PubchemFP498	0.011063131	PubchemFP547	0.015893805
PubchemFP779	0.007921632	PubchemFP694	0.011067684	PubchemFP645	0.015920768
PubchemFP762	0.007931407	PubchemFP515	0.011131948	PubchemFP206	0.015928831
PubchemFP477	0.008003971	PubchemFP276	0.011143804	PubchemFP778	0.016008987
nF6Ring	0.008036845	PubchemFP554	0.011155358	maxaasN	0.016063501
PubchemFP543	0.008043897	PubchemFP685	0.011270947	PubchemFP387	0.016095799
PubchemFP474	0.00808522	PubchemFP717	0.011349262	PubchemFP24	0.016254277
PubchemFP382	0.008161232	PubchemFP458	0.011353542	PubchemFP472	0.016499222
PubchemFP763	0.008221448	PubchemFP118	0.01140231	PubchemFP758	0.016548346
PubchemFP475	0.008235555	PubchemFP274	0.011456617	PubchemFP605	0.016561532
PubchemFP262	0.008338461	PubchemFP350	0.011477799	PubchemFP658	0.016637626
PubchemFP673	0.008358031	PubchemFP729	0.011530072	PubchemFP734	0.016662541
PubchemFP357	0.008362062	PubchemFP650	0.011688466	PubchemFP396	0.016729564
PubchemFP394	0.008419689	PubchemFP301	0.011695673	PubchemFP193	0.01674905
PubchemFP156	0.008427595	PubchemFP728	0.011717869	PubchemFP812	0.016769623
PubchemFP840	0.008429447	PubchemFP675	0.011817396	PubchemFP821	0.016772027
PubchemFP454	0.008447246	PubchemFP813	0.01182941	PubchemFP755	0.016809371
PubchemFP541	0.008482328	PubchemFP589	0.011904543	PubchemFP340	0.017042281
PubchemFP533	0.008489282	PubchemFP615	0.011983992	maxsssN	0.017050122
PubchemFP784	0.008503585	PubchemFP481	0.011996267	PubchemFP218	0.017095486
PubchemFP825	0.008587865	PubchemFP408	0.012013767	PubchemFP383	0.017125549
PubchemFP189	0.008614022	PubchemFP733	0.012062148	PubchemFP418	0.017253905
PubchemFP194	0.008671679	PubchemFP652	0.01213443	PubchemFP706	0.017277273
PubchemFP602	0.008681073	PubchemFP833	0.012143658	PubchemFP256	0.017384366
PubchemFP395	0.008800686	nl	0.012427609	PubchemFP612	0.017539028
PubchemFP624	0.008908353	nB	0.012437206	PubchemFP580	0.017574684
PubchemFP651	0.008931798	PubchemFP467	0.012573166	PubchemFP593	0.017682683
PubchemFP451	0.008948723	PubchemFP297	0.012661635	PubchemFP597	0.017810287
PubchemFP388	0.00908196	PubchemFP695	0.01270045	nT7Ring	0.017822072
PubchemFP149	0.009089167	PubchemFP740	0.012711101	PubchemFP507	0.01788419
PubchemFP780	0.009101617	PubchemFP860	0.012744298	PubchemFP482	0.017889384
PubchemFP803	0.009123536	PubchemFP500	0.012772561	PubchemFP715	0.017896441
PubchemFP398	0.009170907	PubchemFP195	0.012797437	maxsCl	0.018053451
PubchemFP721	0.009232767	PubchemFP682	0.012848001	PubchemFP749	0.018090698
PubchemFP469	0.009274744	PubchemFP644	0.012884553	PubchemFP182	0.018110849
PubchemFP716	0.009375735	PubchemFP232	0.012922693	PubchemFP34	0.018149008
PubchemFP347	0.009386134	PubchemFP476	0.012998323	maxdssC	0.018212252
maxdsN	0.009399798	PubchemFP499	0.013068659	PubchemFP219	0.018271372
PubchemFP810	0.009471003	PubchemFP756	0.01310925	PubchemFP759	0.018431271
PubchemFP747	0.009474311	PubchemFP626	0.01313405	nsssB	0.018460551
PubchemFP153	0.009521086	PubchemFP514	0.013142751	PubchemFP443	0.018494052
PubchemFP298	0.009544205	PubchemFP43	0.01316992	PubchemFP246	0.01860379
PubchemFP439	0.009562924	nsBr	0.013223509	PubchemFP437	0.018677485
PubchemFP835	0.009602129	PubchemFP471	0.013226985	PubchemFP442	0.018689639
PubchemFP147	0.009611737	PubchemFP152	0.013252529	PubchemFP613	0.019012656
PubchemFP349	0.009781389	PubchemFP345	0.013357944	PubchemFP586	0.019029139
PubchemFP625	0.009803184	PubchemFP815	0.013376824	PubchemFP647	0.019113582
PubchemFP751	0.009809016	PubchemFP534	0.013384921	PubchemFP683	0.019158785
PubchemFP260	0.009832573	PubchemFP861	0.013470935	PubchemFP713	0.019250572
PubchemFP653	0.009942845	PubchemFP190	0.013553969	PubchemFP531	0.019408616
PubchemFP814	0.009970475	PubchemFP506	0.01359637	n4Ring	0.01944603
PubchemFP422	0.009982288	PubchemFP750	0.013598783	PubchemFP800	0.019510264
PubchemFP792	0.009990528	PubchemFP669	0.013697766	PubchemFP623	0.019541892
PubchemFP742	0.010023753	PubchemFP148	0.013698313	maxsssO	0.019627627
PubchemFP818	0.010075344	PubchemFP565	0.013705655	PubchemFP592	0.019741713
PubchemFP518	0.01018376	PubchemFP214	0.013854419	PubchemFP822	0.019750939
PubchemFP616	0.01019978	PubchemFP536	0.013949578	PubchemFP374	0.020169061
PubchemFP604	0.010248936	PubchemFP714	0.01395397	PubchemFP241	0.020266127
PubchemFP327	0.01025253	PubchemFP826	0.014178603	PubchemFP364	0.020357362
PubchemFP44	0.010301862	PubchemFP233	0.014255958	PubchemFP757	0.020527195
PubchemFP772	0.010315915	PubchemFP761	0.014451178	PubchemFP464	0.020620442
PubchemFP509	0.01034302	PubchemFP46	0.014508203	PubchemFP737	0.020641411
PubchemFP791	0.010432183	PubchemFP453	0.014509907	PubchemFP700	0.020746274
PubchemFP328	0.010452858	PubchemFP544	0.014618972	nT4Ring	0.020800887
PubchemFP548	0.010605274	PubchemFP703	0.014697755	n7Ring	0.020842275
PubchemFP719	0.010609989	PubchemFP359	0.014760158	PubchemFP353	0.02084794
PubchemFP839	0.010630437	PubchemFP305	0.014783602	PubchemFP25	0.020872528
PubchemFP796	0.010662262	PubchemFP385	0.014885134	maxsF	0.020910678
PubchemFP391	0.010682445	PubchemFP363	0.014885743	PubchemFP820	0.020977977
nsI	0.010695088	PubchemFP358	0.01502028	PubchemFP38	0.021091619

Atributo	Ganho	Atributo	Ganho	Atributo	Ganho
PubchemFP659	0.021128155	PubchemFP535	0.029160639	nT11Ring	0.045788279
PubchemFP247	0.021149668	PubchemFP676	0.029215324	PubchemFP654	0.046165026
PubchemFP213	0.021276499	PubchemFP157	0.029216117	PubchemFP356	0.046433311
PubchemFP22	0.021297232	PubchemFP456	0.029295663	PubchemFP698	0.046483407
PubchemFP777	0.021598808	PubchemFP438	0.029627368	PubchemFP334	0.046525619
n9Ring	0.021729943	PubchemFP407	0.030166587	PubchemFP516	0.04672844
PubchemFP414	0.021863655	PubchemFP376	0.03017698	nF11Ring	0.046738701
PubchemFP160	0.021875424	PubchemFP538	0.031370048	PubchemFP257	0.047238616
PubchemFP351	0.021945312	nP	0.031522517	PubchemFP259	0.047605767
PubchemFP776	0.022041167	PubchemFP132	0.031572098	PubchemFP690	0.047897058
PubchemFP33	0.022059414	PubchemFP495	0.031702139	nsF	0.048481982
PubchemFP342	0.022143863	PubchemFP620	0.032172113	naasN	0.048668734
PubchemFP393	0.0222061	PubchemFP338	0.032535272	PubchemFP484	0.049197714
PubchemFP523	0.022230481	nF8Ring	0.032535369	nF	0.049236956
PubchemFP299	0.022257914	PubchemFP555	0.032642821	PubchemFP179	0.04926584
PubchemFP37	0.022522045	nAcid	0.03294384	PubchemFP601	0.049411917
PubchemFP459	0.022743495	PubchemFP704	0.032988977	PubchemFP709	0.049917443
PubchemFP671	0.022788009	ndsN	0.033251444	PubchemFP699	0.04997192
PubchemFP261	0.022798311	PubchemFP16	0.033268632	PubchemFP607	0.050177935
PubchemFP712	0.023205702	PubchemFP191	0.033515545	maxwHBa	0.050308904
PubchemFP591	0.023220835	PubchemFP798	0.033787738	PubchemFP710	0.050327503
PubchemFP293	0.023236649	PubchemFP258	0.034221218	PubchemFP430	0.050982441
VCH-3	0.023563762	PubchemFP696	0.034299916	PubchemFP664	0.051125124
PubchemFP824	0.023692329	PubchemFP549	0.034472065	PubchemFP708	0.051151277
PubchemFP494	0.023812917	PubchemFP19	0.034787038	PubchemFP674	0.051153315
PubchemFP661	0.023952105	PubchemFP411	0.034800059	PubchemFP379	0.051250246
PubchemFP445	0.02399402	PubchemFP632	0.034924907	PubchemFP614	0.052093944
PubchemFP362	0.023997757	PubchemFP657	0.034929346	PubchemFP564	0.052574048
PubchemFP287	0.024077453	PubchemFP373	0.035027631	PubchemFP333	0.053440635
PubchemFP797	0.024198305	ntsC	0.035278437	PubchemFP628	0.05348973
PubchemFP598	0.024242143	PubchemFP697	0.035635863	PubchemFP599	0.053631753
PubchemFP465	0.02424262	maxaasC	0.035766569	nF10Ring	0.053770754
PubchemFP566	0.024362125	PubchemFP573	0.035875999	nT10Ring	0.053904511
PubchemFP656	0.024577645	PubchemFP180	0.035902674	nT8Ring	0.054219811
PubchemFP550	0.024679246	PubchemFP447	0.036001961	PubchemFP633	0.054990517
PubchemFP186	0.024707036	PubchemFP530	0.036365073	PubchemFP596	0.055126284
nddssS	0.024756147	PubchemFP378	0.036450581	PubchemFP12	0.055420096
PubchemFP432	0.0248062	PubchemFP693	0.036565921	ssssN	0.055548158
PubchemFP15	0.024862264	PubchemFP619	0.036584224	PubchemFP678	0.056060419
PubchemFP563	0.024910913	PubchemFP255	0.036862283	PubchemFP560	0.05623277
PubchemFP457	0.024924735	PubchemFP545	0.037199031	PubchemFP600	0.05693464
PubchemFP420	0.024987976	PubchemFP585	0.037219647	maxdO	0.057128173
PubchemFP819	0.025011745	PubchemFP286	0.037589459	PubchemFP370	0.057213902
PubchemFP736	0.025051677	PubchemFP637	0.037684777	PubchemFP450	0.057408852
PubchemFP435	0.025114756	PubchemFP355	0.037734809	PubchemFP487	0.0574109
PubchemFP159	0.025196629	PubchemFP572	0.037928525	PubchemFP636	0.057424424
PubchemFP184	0.025223872	PubchemFP188	0.038058777	PubchemFP405	0.057743493
PubchemFP294	0.02525035	PubchemFP352	0.038141853	PubchemFP520	0.057832227
PubchemFP782	0.025305501	PubchemFP18	0.038472459	ETA_dAlpha_A	0.057999565
PubchemFP501	0.025526975	PubchemFP579	0.03910382	SCH-4	0.05831243
PubchemFP449	0.026215691	PubchemFP689	0.039150663	PubchemFP686	0.058341399
PubchemFP684	0.026351388	PubchemFP711	0.039179825	PubchemFP640	0.058638777
PubchemFP129	0.026367689	PubchemFP389	0.039494977	PubchemFP13	0.058761407
PubchemFP23	0.02644278	PubchemFP187	0.040117882	PubchemFP556	0.058876542
PubchemFP183	0.026518122	PubchemFP666	0.040411394	PubchemFP677	0.058916986
PubchemFP799	0.026646826	PubchemFP504	0.040628425	PubchemFP679	0.059050809
PubchemFP705	0.026719477	PubchemFP594	0.040807195	PubchemFP470	0.059145184
PubchemFP641	0.026734113	PubchemFP341	0.041268861	PubchemFP375	0.059283534
SCH-3	0.026771061	PubchemFP638	0.041713188	PubchemFP524	0.059424743
PubchemFP429	0.027018748	PubchemFP284	0.042198353	PubchemFP660	0.060054223
PubchemFP372	0.027054733	PubchemFP680	0.042598957	PubchemFP665	0.060434218
PubchemFP692	0.027072181	nCl	0.042606181	PubchemFP490	0.060850873
PubchemFP568	0.02717272	n8Ring	0.042614134	PubchemFP584	0.061077745
C1SP1	0.027661516	PubchemFP377	0.042638662	n5Ring	0.061169578
PubchemFP392	0.027787891	PubchemFP185	0.042660694	PubchemFP668	0.06156037
PubchemFP314	0.027818728	PubchemFP344	0.04290195	naaaC	0.061580322
PubchemFP419	0.027932356	PubchemFP621	0.042948054	PubchemFP603	0.061737647
PubchemFP192	0.028069112	PubchemFP227	0.043359117	maxaaN	0.061840709
PubchemFP425	0.02828568	PubchemFP441	0.04402603	PubchemFP618	0.062157369
PubchemFP181	0.02830844	PubchemFP582	0.044271043	nT5Ring	0.062514887
ssS	0.028411168	PubchemFP688	0.044388413	PubchemFP434	0.062833731
ndssSP	0.028479734	nG12Ring	0.044444753	PubchemFP552	0.062887228
nBondsT	0.028543017	PubchemFP701	0.044899769	PubchemFP578	0.063377633
PubchemFP30	0.028663698	PubchemFP11	0.044907443	PubchemFP371	0.064808001
PubchemFP577	0.028797787	nS	0.044950132	PubchemFP384	0.064902165
nTN	0.028819316	PubchemFP416	0.045030733	PubchemFP608	0.064916792
PubchemFP466	0.02884374	nsCl	0.045179787	PubchemFP634	0.064930364
PubchemFP228	0.02902661	PubchemFP332	0.045640127	PubchemFP570	0.065319796

Atributo	Ganho	Atributo	Ganho	Atributo	Ganho
PubchemFP595	0.0660843	nHBAcc2	0.257100265	ETA_Alpha	0.656451515
PubchemFP519	0.066245156	nHBAcc3	0.258151227	VP-2	0.656631
PubchemFP446	0.067663561	SC-6	0.264599222	ETA_BetaP_ns	0.656881945
VCH-4	0.067948995	ATSc3	0.266347969	ETA_Eta	0.657192381
C4SP3	0.069200161	nO	0.270379303	Kier2	0.660117695
PubchemFP346	0.070548376	MDEN-23	0.270829686	VCH-6	0.661172279
PubchemFP335	0.07169286	SCH-5	0.274426101	SP-0	0.661594751
naaN	0.072197766	MDEC-14	0.282788392	SP-2	0.661748781
PubchemFP178	0.072884722	SC-4	0.282840368	ETA_BetaP	0.662319894
PubchemFP440	0.07294748	MDEC-24	0.298334475	SC-3	0.663983387
nT12Ring	0.073236848	MDEN-22	0.307732889	MLFER_BO	0.66428073
maxHBa	0.073382513	MDEC-34	0.310564634	SP-6	0.664771536
nF12Ring	0.074602818	VC-6	0.331783164	ETA_Shape_Y	0.667226808
PubchemFP662	0.077681389	nAtomP	0.33662658	ETA_AlphaP	0.667543534
PubchemFP452	0.07772317	VC-4	0.34509896	hmax	0.667704609
PubchemFP339	0.077880699	MDEC-11	0.346469514	ETA_Eta_B	0.668272102
PubchemFP366	0.077918411	VCH-5	0.360203442	ETA_Eta_R	0.668440263
PubchemFP567	0.078772574	VAdjMat	0.374400933	VP-5	0.670658333
LipinskiFailures	0.079073711	nHeavyAtom	0.374530757	ATSm3	0.670712782
nT9Ring	0.07914918	nC	0.375664302	SCH-7	0.671327699
PubchemFP380	0.080067962	BCUTc-1l	0.380103468	ETA_EtaP_F	0.671342854
PubchemFP639	0.081447813	ETA_Shape_X	0.380745254	ATSm5	0.672264491
nF9Ring	0.081795054	nBondsS	0.382729746	ETA_Eta_L	0.672356008
PubchemFP537	0.082509138	BCUTc-1h	0.387119066	VCH-7	0.672447008
nBase	0.083881093	MLogP	0.387262775	SP-3	0.672473767
PubchemFP462	0.0849428	nH	0.392103685	XLogP	0.672834398
C1SP2	0.086995858	nBondsS3	0.400576885	VC-3	0.673225343
PubchemFP681	0.091359639	nBondsS2	0.400767598	ATSp1	0.673331481
PubchemFP337	0.092711659	MDEO-12	0.407169965	VP-0	0.673357633
PubchemFP406	0.092843856	nBonds2	0.414502812	ETA_Eta_B_RC	0.676191098
MDEN-11	0.093189258	nBonds	0.414529321	SP-7	0.676422377
PubchemFP21	0.095642146	ETA_BetaP_ns_d	0.414992244	MLFER_S	0.676497031
nN	0.101214482	ALogp2	0.426702458	MLFER_L	0.676816495
PubchemFP308	0.103722826	MDEO-11	0.466807016	hmin	0.677487602
PubchemFP20	0.105675223	ETA_Beta_ns	0.487906165	bpol	0.678662991
ETA_Beta_ns_d	0.105932588	ATSc4	0.501932398	VABC	0.678866399
C3SP2	0.111249365	MDEC-12	0.502534738	ATSp5	0.679089297
nAtomLAC	0.112232362	nAtom	0.514751173	ATSp4	0.679797955
MDEN-33	0.114487922	ETA_dBeta	0.518549926	SP-4	0.680577804
ndO	0.119671689	ETA_Epsilon_3	0.527369349	VP-3	0.680593471
nT6Ring	0.120617151	WPOL	0.532471948	ETA_EtaP	0.680797698
n6Ring	0.124749008	MDEC-13	0.535453334	MW	0.680901893
C3SP3	0.125460267	Zagreb	0.539590086	ATSp2	0.681066036
PubchemFP642	0.129798086	ETA_Beta	0.572409707	ALogP	0.681895854
PubchemFP581	0.130544248	Kier1	0.580280418	VPC-4	0.682019272
nssssC	0.130907287	WTPT-5	0.584176544	ETA_dEpsilon_B	0.684181539
nRing	0.138746546	ETA_Beta_s	0.584474105	ETA_Eta_F	0.684489145
nssO	0.143840806	ETA_BetaP_s	0.585917378	ETA_EtaP_F_L	0.684627969
nHBDon	0.149263866	fragC	0.58910314	BCUTp-1l	0.684770436
MDEN-13	0.151175043	WPATH	0.596255626	VP-4	0.684891681
naasC	0.155010269	MLFER_A	0.602898286	ETA_EtaP_B_RC	0.686611748
ndssC	0.155667382	WTPT-4	0.606374655	apol	0.688385892
nHBDon_Lipinski	0.159308571	WTPT-2	0.624239224	ETA_Epsilon_4	0.688868498
nBondsD2	0.163593132	MDEC-22	0.627593772	ETA_dEpsilon_A	0.689290265
nTG12Ring	0.164737229	SCH-6	0.631101016	ECCEN	0.689845391
nwHBa	0.165615063	MDEC-23	0.63335695	SPC-6	0.690414952
nFG12Ring	0.165869599	ATSp3	0.634685759	ETA_dEpsilon_C	0.691412001
MDEC-44	0.172952778	VC-5	0.636256081	SP-5	0.691649469
nFRing	0.17359885	ATSm1	0.638475497	McGowan_Volume	0.692064508
nAtomLC	0.181272237	SC-5	0.64262183	ETA_Psi_1	0.693665807
PetitjeanNumber	0.181753351	ETA_dAlpha_B	0.645399978	ETA_dPsi_A	0.693902315
nRotB	0.182924678	Kier3	0.646772059	WTPT-1	0.694905176
nHBa	0.184978959	ATSc5	0.647001579	ETA_Eta_R_L	0.695877793
C1SP3	0.187038171	VPC-6	0.647419231	AMR	0.6959801
nBondsD	0.203008612	TopoPSA	0.647532889	SP-1	0.696061066
BCUTw-1l	0.204045514	VP-7	0.648202943	BCUTp-1h	0.697268448
BCUTw-1h	0.2048198	SPC-5	0.648565129	ETA_Shape_P	0.698091245
nAromBond	0.205202002	ATSm4	0.648705237	ETA_EtaP_L	0.698101818
C2SP2	0.212512435	MDEC-33	0.64962073	SPC-4	0.705136449
naAromAtom	0.213802347	ETA_Epsilon_5	0.65251786	ETA_dBetaP	0.707761076
ATSc2	0.215529989	ETA_Epsilon_1	0.652796448	ETA_Eta_F_L	0.709144689
MDEN-12	0.219204687	ETA_Epsilon_2	0.653309451	MLFER_E	0.711961675
C2SP3	0.222792673	VPC-5	0.653476593	ETA_EtaP_B	0.7184825
nHBAcc_Lipinski	0.223426796	WTPT-3	0.653899223		
nHBAcc	0.235535068	ATSm2	0.653935447		
ATSc1	0.235787802	VP-1	0.654052713		
MDEO-22	0.236898037	VP-6	0.654835172		
nTRing	0.251741524	MLFER_BH	0.655310948		

B.2 Classificação

B.2.1 Ganho de informação mútua

Atributo	Ganho	Atributo	Ganho	Atributo	Ganho
ETA_Eta	0.467793523	apol	0.464160023	MDEO-11	0.313977806
VABC	0.467339402	McGowan_Volume	0.46412227	ETA_Epsilon_3	0.288855057
ATSc2	0.467304126	SPC-4	0.464082643	ETA_dBeta	0.287487101
ETA_dEpsilon_A	0.467257721	MW	0.463988752	Zagreb	0.287304761
ATSp2	0.46719847	VP-5	0.463875227	nAtom	0.281387734
ETA_Epsilon_4	0.467175199	ATSm3	0.46369856	ETA_Beta	0.280902143
ETA_EtaP_L	0.467111814	ATSc4	0.463436272	MDEO-12	0.275963661
SP-4	0.467041005	ATSc5	0.463419513	ETA_Beta_ns	0.266090463
VP-5	0.467022329	WPATH	0.463350782	ETA_Shape_X	0.25996127
SP-2	0.466926564	MLFER_BH	0.463279395	VC-4	0.227703186
ATSp2	0.469992454	VP-3	0.463249652	nBondsS2	0.227183041
SP-5	0.46848567	ETA_dEpsilon_B	0.463246558	VC-6	0.22336854
MLFER_BO	0.468405146	TopoPSA	0.463216557	nBondsS3	0.223214432
VPC-5	0.467442526	SP-7	0.463170517	MDEC-11	0.222418718
ETA_Eta_L	0.467396135	ETA_EtaP_L	0.463139245	ETA_BetaP_ns_d	0.216728332
BCUTc-1l	0.467265253	ETA_Eta_B_RC	0.463122375	nBonds2	0.216030487
Alogp2	0.467170343	MLFER_S	0.463050983	nBonds	0.215684887
ETA_EtaP_F	0.467134294	SC-3	0.463041909	MDEC-34	0.213526795
SPC-6	0.467063486	SP-6	0.462564132	MLogP	0.213072468
ETA_Eta	0.466911607	ETA_Shape_Y	0.462558419	VCH-5	0.21183003
VP-2	0.466871692	SCH-7	0.462362899	MDEC-24	0.210118024
ATSm5	0.466820855	hmin	0.462266863	nH	0.198798388
bpol	0.466691394	BCUTw-1h	0.461972757	SC-4	0.187686056
ATSp5	0.466644125	VP-6	0.461743038	VAdjMat	0.186450081
BCUTp-1l	0.466610326	VC-3	0.461486251	nBondsS	0.183951372
VPC-6	0.466570527	WTPT-3	0.46141129	nHeavyAtom	0.180307412
BCUTc-1h	0.466501632	ECCEN	0.46054219	nC	0.175715897
ATSp1	0.466407559	VCH-7	0.460525613	SC-6	0.173850434
SP-4	0.46639791	VP-7	0.460039665	MDEC-14	0.173303677
ETA_Shape_P	0.466385814	VC-5	0.459423388	SCH-5	0.16342928
SP-1	0.466360254	ETA_Psi_1	0.457515242	nHBAcc3	0.158875673
ATSp4	0.466248881	ETA_EtaP_F_L	0.457513804	MDEN-22	0.158281005
ATSp3	0.466127557	ETA_Epsilon_1	0.456210412	C2SP3	0.150876196
VPC-4	0.466097647	VCH-6	0.456030003	MDEN-23	0.142934365
WTPT-1	0.465945691	WTPT-4	0.455769479	MDEO-22	0.140943474
ETA_Eta_B	0.46558346	SP-0	0.455712121	nO	0.13309058
ETA_Eta_F	0.465581965	MDEC-33	0.455192214	nAtomP	0.132523386
AlogP	0.465581651	ETA_dPsi_A	0.454546282	nTRing	0.123332908
AMR	0.465571779	ATSm1	0.454326078	nHBAcc2	0.119842232
BCUTp-1h	0.465543884	ETA_Epsilon_5	0.453287124	MDEN-12	0.117561893
ETA_Eta_R_L	0.465517793	ETA_Epsilon_2	0.452170133	nHBAcc_Lipinski	0.112861711
SPC-5	0.465514758	ETA_EtaP_B_RC	0.45051922	nHBAcc	0.111674327
ETA_dEpsilon_A	0.4654897	ETA_Alpha	0.450430318	nTG12Ring	0.111617425
WTPT-2	0.465455073	ETA_dBetaP	0.445776103	nFG12Ring	0.111484682
ETA_dEpsilon_C	0.465370425	ETA_EtaP_B	0.443737232	PubchemFP642	0.109736363
SP-2	0.465236357	MDEC-23	0.436807965	PubchemFP581	0.10737009
VP-4	0.465177409	fragC	0.4360954	C2SP2	0.104474738
ATSm4	0.465016036	Kier2	0.435794461	nHBDOn	0.100462854
ATSc1	0.465013047	ETA_AlphaP	0.434928139	nHBa	0.098278784
ATSc3	0.464920034	MDEC-22	0.430875403	MDEC-44	0.095419379
XLogP	0.464873771	Kier3	0.429002377	C1SP3	0.095385193
VP-1	0.464862262	WTPT-5	0.426493982	nFRing	0.095221029
ETA_EtaP	0.464790254	SC-5	0.424754978	nBondsD	0.090191928
MLFER_L	0.464690638	ETA_BetaP_ns	0.422478847	nAtomLAC	0.085725719
SP-3	0.464654727	ETA_dAlpha_B	0.419327337	nssO	0.084730956
MLFER_E	0.46460454	ETA_BetaP	0.415090411	C3SP3	0.082186841
ETA_Eta_R	0.464578323	SCH-6	0.404774775	PubchemFP308	0.080013167
VABC	0.464565964	MLFER_A	0.403952254	PetitjeanNumber	0.079775315
ATSm2	0.464405119	MDEC-13	0.365643488	nHBDOn_Lipinski	0.079460655
ETA_Epsilon_4	0.464386498	MDEC-12	0.359330383	PubchemFP406	0.078604925
ATSc2	0.464383223	ETA_BetaP_s	0.355888591	PubchemFP567	0.078425813
ETA_Eta_F_L	0.464254458	Kier1	0.342596242	ndssC	0.076487024
hmax	0.464231591	BCUTw-1l	0.333008911	nAtomLC	0.075015234
VP-0	0.4642106	WPOL	0.322738549	nRotB	0.074672805

Atributo	Ganho	Atributo	Ganho	Atributo	Ganho
PubchemFP662	0.072061219	PubchemFP159	0.026055769	PubchemFP797	0.015648149
nssssC	0.070749999	maxaaasC	0.026005862	PubchemFP414	0.015608414
nT9Ring	0.069593403	PubchemFP568	0.025969224	PubchemFP420	0.015531094
PubchemFP20	0.068707908	PubchemFP419	0.025641775	PubchemFP11	0.01550209
nF9Ring	0.066961848	PubchemFP563	0.025389817	nCl	0.015487933
MDEN-13	0.065005288	PubchemFP379	0.025240312	PubchemFP715	0.015305562
PubchemFP339	0.064984782	PubchemFP676	0.024968778	PubchemFP456	0.015086119
nT12Ring	0.064940733	nssN	0.024704918	PubchemFP16	0.015013407
nRing	0.064589712	PubchemFP535	0.024668761	PubchemFP554	0.014866401
nF12Ring	0.064262424	PubchemFP416	0.024641012	PubchemFP620	0.014828936
nBondsD2	0.064106017	PubchemFP191	0.024422648	PubchemFP129	0.014801528
PubchemFP366	0.063366276	PubchemFP674	0.024391502	nsF	0.014775244
MDEN-33	0.062668786	PubchemFP757	0.02384938	PubchemFP591	0.014679621
PubchemFP639	0.062319552	PubchemFP457	0.02376814	PubchemFP693	0.014520168
PubchemFP335	0.060299253	nF10Ring	0.023232306	PubchemFP778	0.014422133
PubchemFP380	0.058282804	nT7Ring	0.023173603	PubchemFP352	0.01420681
PubchemFP452	0.058176384	PubchemFP822	0.023136282	PubchemFP218	0.013981241
PubchemFP346	0.056691048	PubchemFP441	0.022965069	ndssP	0.013929062
PubchemFP681	0.055757366	PubchemFP661	0.022936374	PubchemFP432	0.013806364
ndO	0.05453345	PubchemFP736	0.022871305	PubchemFP241	0.013789868
nAromBond	0.053124882	PubchemFP466	0.022764485	PubchemFP453	0.013620122
PubchemFP13	0.052563328	PubchemFP566	0.022753849	PubchemFP654	0.013534355
maxHBa	0.052401146	PubchemFP684	0.022724602	nN	0.013522299
PubchemFP462	0.051844872	PubchemFP824	0.022706991	PubchemFP687	0.013438111
nwHBa	0.051170641	nBase	0.022542495	PubchemFP796	0.013195305
PubchemFP21	0.049002753	PubchemFP596	0.022315612	maxsF	0.013151761
PubchemFP537	0.04745707	PubchemFP38	0.022215189	n4Ring	0.013104203
n5Ring	0.045970332	nsCl	0.022167145	PubchemFP418	0.012960284
nT5Ring	0.04551613	C1SP2	0.022059622	PubchemFP484	0.012951989
PubchemFP519	0.04549305	PubchemFP798	0.022057348	PubchemFP645	0.012742042
nT8Ring	0.044960955	PubchemFP429	0.021185732	PubchemFP545	0.012713774
naAromAtom	0.044871749	PubchemFP465	0.020967462	PubchemFP777	0.012557912
maxdO	0.04339108	PubchemFP614	0.020942475	PubchemFP534	0.012545268
nG12Ring	0.043381558	PubchemFP820	0.020676523	PubchemFP641	0.012517388
ETA_dAlpha_A	0.041959929	PubchemFP183	0.020465444	maxaaN	0.012514895
PubchemFP337	0.041825662	naasC	0.020357492	PubchemFP180	0.012353563
maxaaN	0.041458944	PubchemFP707	0.020308702	PubchemFP459	0.012346571
PubchemFP680	0.040732716	PubchemFP647	0.020242855	PubchemFP669	0.012339543
naasN	0.040716945	PubchemFP314	0.020142427	nAcid	0.012286821
PubchemFP440	0.039205001	maxsCl	0.019658817	nT4Ring	0.012263406
PubchemFP686	0.038840131	PubchemFP759	0.019354689	nF6Ring	0.012250407
PubchemFP638	0.038464337	nBondsT	0.019343994	PubchemFP334	0.01223555
nF8Ring	0.03807907	PubchemFP132	0.019093379	PubchemFP447	0.012193341
PubchemFP405	0.037911792	PubchemFP692	0.018992701	PubchemFP835	0.012120851
nF11Ring	0.037205498	PubchemFP594	0.018922464	PubchemFP246	0.011941711
PubchemFP338	0.036816141	PubchemFP704	0.018720021	PubchemFP378	0.011792299
VCH-4	0.036092903	PubchemFP286	0.018563647	PubchemFP624	0.011736474
PubchemFP375	0.035977509	PubchemFP193	0.018547148	nssS	0.011672778
PubchemFP450	0.03594605	PubchemFP18	0.018293975	PubchemFP634	0.011608004
PubchemFP487	0.03555055	PubchemFP619	0.018135527	PubchemFP156	0.01149062
LipinskiFailures	0.034702567	PubchemFP582	0.017977798	PubchemFP408	0.011470345
CASP3	0.03453145	PubchemFP573	0.017975334	PubchemFP752	0.011456681
nS	0.032540809	PubchemFP815	0.017959009	PubchemFP446	0.011429723
naaaC	0.03223845	nT6Ring	0.017882103	n9Ring	0.011320841
ETA_Beta_ns_d	0.032033602	PubchemFP353	0.017825834	PubchemFP411	0.011183647
SCH-4	0.031855666	PubchemFP228	0.017619553	maxdssC	0.011161864
PubchemFP637	0.031837426	PubchemFP293	0.017477921	PubchemFP481	0.011070877
PubchemFP157	0.031833912	PubchemFP407	0.017444317	EGFR_SV	0.011027453
nT11Ring	0.031649349	PubchemFP776	0.017377712	PubchemFP782	0.010899328
maxwHBa	0.031493204	PubchemFP219	0.017343962	PubchemFP190	0.010767263
PubchemFP579	0.030876307	ntN	0.017289756	PubchemFP37	0.010654573
PubchemFP341	0.030092162	PubchemFP632	0.017257532	PubchemFP772	0.010575247
PubchemFP799	0.02990647	PubchemFP33	0.017171309	PubchemFP668	0.010533673
PubchemFP227	0.02944496	PubchemFP666	0.017130959	PubchemFP393	0.010415381
PubchemFP701	0.02888536	PubchemFP332	0.017073369	PubchemFP43	0.010151167
PubchemFP689	0.028478214	PubchemFP697	0.017061516	PubchemFP356	0.010148082
PubchemFP690	0.027393736	PubchemFP659	0.016882275	PubchemFP184	0.010143848
C3SP2	0.027283825	PubchemFP698	0.016880579	PubchemFP454	0.010142773
C1SP1	0.026766246	PubchemFP706	0.016727123	PubchemFP504	0.010115749
n6Ring	0.026665953	PubchemFP425	0.01670292	PubchemFP826	0.010014373
PubchemFP699	0.026482735	PubchemFP160	0.016699579	PubchemFP598	0.009961527
PubchemFP560	0.026375786	PubchemFP22	0.016327207	PubchemFP810	0.009832925
n8Ring	0.026341997	nddssS	0.016228558	PubchemFP749	0.009718988
MDEN-11	0.026310104	PubchemFP800	0.01607414	nsBr	0.009714277
PubchemFP213	0.026284424	PubchemFP531	0.016040975	PubchemFP327	0.00970113
ntsC	0.026235964	PubchemFP383	0.015906527	PubchemFP328	0.009674202
PubchemFP636	0.026169421	PubchemFP373	0.015738116	PubchemFP330	0.009670565
n7Ring	0.026155789	nT10Ring	0.015671203	PubchemFP779	0.009606793

Atributo	Ganho	Atributo	Ganho	Atributo	Ganho
PubchemFP30	0.009578441	PubchemFP549	0.007251543	PubchemFP494	0.00491487
PubchemFP523	0.0095334	PubchemFP742	0.007210474	PubchemFP497	0.004883333
ndsN	0.009482482	PubchemFP713	0.007205336	PubchemFP284	0.004879932
PubchemFP360	0.009441208	PubchemFP673	0.007189053	PubchemFP580	0.004879454
PubchemFP438	0.009430315	PubchemFP585	0.007184256	NPM1_SV	0.004871184
nP	0.009429209	PubchemFP469	0.007177732	PubchemFP475	0.004829946
PubchemFP538	0.009338806	PubchemFP299	0.007160471	ns5B	0.004803358
PubchemFP809	0.009317285	PubchemFP711	0.0071449	MYC_CN	0.004782404
PubchemFP347	0.009243562	PubchemFP15	0.007138673	CDKN2A_SV	0.00475091
PubchemFP792	0.00922956	PubchemFP483	0.007114837	PubchemFP206	0.004743287
VCH-3	0.009224525	IDH1_CN	0.006994729	PubchemFP44	0.004740576
PubchemFP712	0.00913666	PubchemFP650	0.006965663	PubchemFP733	0.004691461
PubchemFP703	0.009113622	PubchemFP678	0.006907972	PubchemFP423	0.004559957
PubchemFP345	0.009037552	PubchemFP597	0.006854444	PubchemFP389	0.004544337
PubchemFP397	0.008995653	PubchemFP258	0.006817341	PubchemFP740	0.004430302
PubchemFP399	0.008989435	PubchemFP780	0.006671852	PubchemFP548	0.004385018
PubchemFP9	0.008976659	PubchemFP747	0.006662481	PubchemFP605	0.004362108
PubchemFP610	0.008958193	PubchemFP607	0.006637321	PubchemFP467	0.004313403
PubchemFP374	0.008895481	nB	0.006615769	PubchemFP262	0.004290295
PubchemFP825	0.008826211	PubchemFP695	0.006503001	MIS	0.004271344
PubchemFP167	0.008821901	PubchemFP550	0.006485842	HRAS_CN	0.004205916
PubchemFP25	0.008820551	PubchemFP688	0.006437315	WT1_CN	0.004204729
PubchemFP232	0.008802694	PubchemFP387	0.006430764	PubchemFP186	0.004184592
PubchemFP189	0.008758058	PubchemFP342	0.006401953	PubchemFP385	0.004170548
PubchemFP34	0.008745674	PubchemFP474	0.006399139	SCH-3	0.004148378
PubchemFP751	0.00869102	PubchemFP608	0.006394997	CYLD_SV	0.004131918
PubchemFP514	0.008687833	PubchemFP553	0.006365833	PIK3R1_CN	0.004131663
WT1_SV	0.00863162	PubchemFP657	0.006350314	PubchemFP499	0.00411777
nl	0.008628417	PubchemFP622	0.006329809	PubchemFP19	0.00411522
PubchemFP233	0.008565191	PubchemFP766	0.006308167	PubchemFP199	0.004103242
PubchemFP12	0.008543961	PubchemFP257	0.006297781	PubchemFP679	0.0040376
PubchemFP192	0.008482152	maxaaaC	0.006280615	PubchemFP515	0.004035011
PubchemFP840	0.008431088	mindS	0.006208715	SETD2_SV	0.004029604
PubchemFP829	0.008378391	PubchemFP660	0.006163788	PubchemFP178	0.003949434
PubchemFP482	0.008347197	PubchemFP508	0.006121522	PubchemFP358	0.003943499
PubchemFP182	0.008331823	PubchemFP371	0.006091244	PubchemFP470	0.003912922
PubchemFP700	0.008311084	NTRK3_CN	0.006072544	MSH6_SV	0.003877763
PubchemFP294	0.008307957	PubchemFP477	0.006041355	PubchemFP363	0.003877283
PubchemFP616	0.008288082	nBr	0.006011803	PubchemFP439	0.003863225
naaN	0.008279667	PubchemFP390	0.005982429	PubchemFP729	0.003835678
PubchemFP362	0.008269725	PubchemFP145	0.00598004	SMARCA4_SV	0.003833839
CDKN2C_SV	0.008255256	PubchemFP801	0.005904139	PubchemFP555	0.003790997
PubchemFP442	0.008249901	FLT3_SV	0.005891369	PubchemFP808	0.003775476
nsI	0.008185489	maxssO	0.00588491	PubchemFP437	0.003741836
PubchemFP501	0.008184783	PubchemFP788	0.005809698	PubchemFP628	0.003732761
PubchemFP612	0.008125395	MDM2_CN	0.005791569	PubchemFP578	0.003708426
PubchemFP458	0.00811205	PubchemFP259	0.005782056	PubchemFP350	0.003696476
PubchemFP670	0.008090056	PubchemFP819	0.005767103	PubchemFP656	0.003671082
PubchemFP556	0.008081131	PubchemFP252	0.005752	PubchemFP376	0.003577665
PubchemFP297	0.00795585	PubchemFP595	0.005734352	SMO_CN	0.003560066
PubchemFP671	0.007899776	PubchemFP382	0.005704815	PubchemFP305	0.003546625
PubchemFP247	0.007887444	BRAF_CN	0.005694637	PubchemFP449	0.003502543
PubchemFP344	0.007864496	PubchemFP301	0.005670337	PubchemFP694	0.00348896
PubchemFP665	0.007834536	PubchemFP355	0.005641252	PubchemFP274	0.003454494
nF	0.007831179	PubchemFP152	0.005615784	KIT_SV	0.003420186
PubchemFP814	0.007807082	PubchemFP451	0.0055977	PubchemFP812	0.003390954
PubchemFP518	0.007766171	PubchemFP600	0.005529525	PubchemFP276	0.003389034
maxdsN	0.007730485	PubchemFP434	0.00550828	PubchemFP589	0.003362455
PubchemFP734	0.007729425	PubchemFP460	0.005492124	PubchemFP601	0.003345025
PubchemFP473	0.007714893	PubchemFP737	0.005458436	PubchemFP396	0.003329685
PubchemFP621	0.007695665	SdS	0.005451027	PubchemFP381	0.003273204
PubchemFP861	0.007651248	PubchemFP592	0.005381124	PubchemFP188	0.003262351
PubchemFP443	0.007620722	PubchemFP625	0.00536745	NF1_CN	0.003244971
PubchemFP500	0.007603805	PubchemFP46	0.005359588	PubchemFP767	0.003236241
PubchemFP509	0.007523921	BRCA1_SV	0.005355303	BRCA1_CN	0.003212035
PubchemFP403	0.007490864	PubchemFP839	0.005313195	PubchemFP708	0.003192524
PubchemFP489	0.007455359	PubchemFP195	0.005279665	PubchemFP570	0.003184531
n3Ring	0.007388504	PubchemFP493	0.005252828	PubchemFP185	0.003150691
PubchemFP653	0.007382245	PubchemFP762	0.005238956	PubchemFP721	0.003126979
MAP2K4_SV	0.007360846	PubchemFP611	0.005192758	TSC1_CN	0.003097248
PubchemFP705	0.007350374	PubchemFP575	0.005122655	PubchemFP146	0.003081169
PubchemFP626	0.007322717	PubchemFP164	0.005119158	TP53_SV	0.003075384
NRAS_SV	0.007297199	PubchemFP640	0.005116723	PubchemFP430	0.003008647
PubchemFP536	0.007293603	naaS	0.005074692	PubchemFP118	0.003007934
maxsssN	0.007283279	PubchemFP435	0.005008716	PubchemFP368	0.002969599
PubchemFP155	0.007268356	PubchemFP359	0.004995123	PubchemFP603	0.002933319
PubchemFP214	0.007264483	PubchemFP806	0.004946828	SUFU_SV	0.00292616
PubchemFP675	0.007263642	PubchemFP377	0.004922361	PubchemFP417	0.002903249

Atributo	Ganho	Atributo	Ganho	Atributo	Ganho
PubchemFP683	0.002900425	PubchemFP564	0.001494979	BRCA2_SV	0
PubchemFP644	0.002890058	PubchemFP604	0.001491633	CCND2_CN	0
C2SP1	0.002883141	PubchemFP285	0.001463932	CDH1_CN	0
PubchemFP565	0.002882376	PubchemFP834	0.001409325	CDH1_SV	0
PubchemFP490	0.002814163	PubchemFP388	0.00135826	CDK6_CN	0
PubchemFP609	0.002793197	PubchemFP651	0.001340916	CDKN2A_CN	0
APC_CN	0.002791515	PubchemFP543	0.001329009	EGFR_CN	0
PubchemFP200	0.002784485	PubchemFP412	0.001327221	EP300_CN	0
PubchemFP716	0.002770704	PubchemFP372	0.001321701	ERBB2_SV	0
MET_SV	0.002740788	FBXW7_SV	0.001316581	EZH2_CN	0
PubchemFP386	0.00272569	PubchemFP664	0.001309374	EZH2_SV	0
PubchemFP658	0.002702663	PubchemFP194	0.001304346	FAM123B_CN	0
PubchemFP116	0.002696941	PubchemFP602	0.001276879	FAM123B_SV	0
PubchemFP685	0.00268161	PubchemFP421	0.001272027	FBXW7_CN	0
CDKN2a(p14)_SV	0.002669946	PubchemFP115	0.001263995	FGFR3_CN	0
PubchemFP486	0.002650363	PubchemFP629	0.001182597	FGFR3_SV	0
PubchemFP367	0.00263672	PubchemFP655	0.001170977	FLCN_CN	0
CCND3_CN	0.002627725	PubchemFP187	0.001159472	FLCN_SV	0
PubchemFP572	0.002625835	TP53_CN	0.001105554	HRAS_SV	0
PIK3CA_SV	0.00250578	PubchemFP735	0.001054328	JAK2_SV	0
PubchemFP427	0.002502352	PubchemFP784	0.00102312	KDM5C_SV	0
PubchemFP860	0.00246307	DICER1_CN	0.001020103	KDM6A_CN	0
PubchemFP813	0.002395831	PubchemFP719	0.001008018	KIT_CN	0
PubchemFP498	0.002393522	TSC2_CN	0.00100003	KRAS_CN	0
PubchemFP606	0.002392029	PubchemFP714	0.000997438	MAP2K4_CN	0
PubchemFP710	0.00238891	PIK3CA_CN	0.000988776	MET_CN	0
CDKN2C_CN	0.002372286	PubchemFP260	0.000984128	MLH1_CN	0
PubchemFP633	0.002347145	AKT2_CN	0.000981777	MLH1_SV	0
PubchemFP696	0.002346732	NTRK3_SV	0.000924577	MLLT3_CN	0
PubchemFP717	0.002333334	CCND1_CN	0.000885695	MSH2_CN	0
PubchemFP618	0.002310154	PTEN_CN	0.000846938	MSH2_SV	0
PubchemFP400	0.002304099	PTEN_SV	0.000821292	MSH6_CN	0
PubchemFP394	0.002298825	PALB2_CN	0.000786209	MYCL1_CN	0
PubchemFP583	0.002292846	PubchemFP652	0.000779046	MYCN_CN	0
PDGFRA_CN	0.002268694	PubchemFP485	0.000775775	MYC_SV	0
CTNNB1_CN	0.002254669	PubchemFP702	0.000758691	NF2_CN	0
n10Ring	0.00224131	PubchemFP530	0.000686239	NF2_SV	0
PubchemFP764	0.002238011	DICER1_SV	0.000675534	NOTCH1_CN	0
GNAS_SV	0.002229275	naaO	0.000671341	NOTCH1_SV	0
PubchemFP569	0.002188881	PubchemFP541	0.000661026	NPM1_CN	0
PubchemFP520	0.002171855	PubchemFP551	0.000656891	PALB2_SV	0
PubchemFP821	0.002156296	PubchemFP471	0.00064068	PDGFRA_SV	0
PubchemFP586	0.002131622	PubchemFP516	0.000630558	PHOX2B_CN	0
NRAS_CN	0.002111336	PubchemFP505	0.000622369	PTCH1_CN	0
PubchemFP755	0.00210676	GNAS_CN	0.000576206	PTCH1_SV	0
KDR_SV	0.002105543	ndS	0.000564631	PubchemFP10	0
PubchemFP763	0.002084243	PubchemFP593	0.000560571	PubchemFP117	0
PubchemFP148	0.002081077	KDM6A_SV	0.000521305	PubchemFP14	0
PubchemFP150	0.002054922	FGFR2_CN	0.000516089	PubchemFP143	0
PubchemFP298	0.002048102	PubchemFP283	0.000499316	PubchemFP144	0
PubchemFP547	0.00202449	IDH1_SV	0.000486352	PubchemFP17	0
PubchemFP533	0.00199573	NF1_SV	0.000431112	PubchemFP179	0
PubchemFP287	0.001986383	RB1_SV	0.000411428	PubchemFP23	0
PubchemFP577	0.001983049	PubchemFP738	0.000398368	PubchemFP24	0
PubchemFP395	0.001979969	PubchemFP336	0.000379642	PubchemFP248	0
ERBB2_CN	0.001958132	PubchemFP461	0.000375476	PubchemFP255	0
PubchemFP6	0.001955748	CDK4_CN	0.000358538	PubchemFP256	0
PubchemFP147	0.001937786	CYLD_CN	0.00032541	PubchemFP261	0
PubchemFP623	0.001914326	FLT3_CN	0.000314002	PubchemFP300	0
PubchemFP750	0.001907929	PubchemFP351	0.000307401	PubchemFP333	0
PubchemFP803	0.001858656	PubchemFP431	0.000245403	PubchemFP340	0
PubchemFP495	0.00184978	APC_SV	0.000218771	PubchemFP349	0
PubchemFP613	0.001823012	KDR_CN	0.000217668	PubchemFP357	0
CDKN2a(p14)_CN	0.001791283	PubchemFP682	0.000212674	PubchemFP364	0
KDM5C_CN	0.001780343	JAK2_CN	0.000208095	PubchemFP365	0
PubchemFP709	0.00177845	PubchemFP149	0.000176909	PubchemFP392	0
PIK3R1_SV	0.001749033	PubchemFP756	0.000144731	PubchemFP398	0
RUNX1_SV	0.001744271	CTNNB1_SV	0.000141747	PubchemFP404	0
PubchemFP370	0.001731389	VHL_CN	0.000123553	PubchemFP409	0
PubchemFP130	0.001720874	PubchemFP391	0.000106998	PubchemFP413	0
KRAS_SV	0.00171509	PubchemFP677	0.000102797	PubchemFP422	0
ALK_SV	0.001619923	PubchemFP544	8.56E-05	PubchemFP428	0
STK11_CN	0.001586931	RUNX1_CN	5.77E-05	PubchemFP436	0
SUFU_CN	0.001563521	PubchemFP480	5.50E-05	PubchemFP445	0
PHOX2B_SV	0.001539636	PubchemFP181	5.32E-05	PubchemFP448	0
PubchemFP153	0.00153091	ALK_CN	0	PubchemFP464	0
PubchemFP384	0.00152722	BRAF_SV	0	PubchemFP472	0
PubchemFP805	0.001497575	BRCA2_CN	0	PubchemFP476	0

Atributo	Ganho	Atributo	Ganho	Atributo	Ganho
PubchemFP506	0	PubchemFP745	0	SMARCA4_CN	0
PubchemFP507	0	PubchemFP746	0	SMARCB1_CN	0
PubchemFP517	0	PubchemFP758	0	SMO_SV	0
PubchemFP524	0	PubchemFP761	0	SOCS1_CN	0
PubchemFP552	0	PubchemFP770	0	SOCS1_SV	0
PubchemFP558	0	PubchemFP771	0	STK11_SV	0
PubchemFP574	0	PubchemFP785	0	TET2_CN	0
PubchemFP584	0	PubchemFP791	0	TET2_SV	0
PubchemFP588	0	PubchemFP818	0	TSC1_SV	0
PubchemFP599	0	PubchemFP827	0	TSC2_SV	0
PubchemFP615	0	PubchemFP830	0	VHL_SV	0
PubchemFP630	0	PubchemFP833	0		
PubchemFP648	0	RB1_CN	0		
PubchemFP691	0	RET_CN	0		
PubchemFP722	0	RET_SV	0		
PubchemFP725	0	SETD2_CN	0		
PubchemFP728	0	SMAD4_CN	0		
PubchemFP743	0	SMAD4_SV	0		

Anexo C

Conjunto de dados completos

C.1 Regressão

1D 2D		Non Stringent	Stringent	Non Stringent Blind Test	Stringent Blind Test
RF	RMSE	0.85 +/- 0.028	0.89 +/- 0.029	0.95 +/- 0.003	0.86 +/- 0.004
	RP	0.84 +/- 0.010	0.82 +/- 0.011	0.79 +/- 0.001	0.83 +/- 0.002
	R2	0.70 +/- 0.017	0.67 +/- 0.018	0.62 +/- 0.002	0.69 +/- 0.002
SVM	RMSE	0.94 +/- 0.040	0.96 +/- 0.034	0.98 +/- 0.002	0.94 +/- 0.002
	RP	0.80 +/- 0.010	0.79 +/- 0.010	0.77 +/- 0.001	0.80 +/- 0.001
	R2	0.64 +/- 0.017	0.62 +/- 0.015	0.59 +/- 0.002	0.63 +/- 0.002
XGBoost	RMSE	0.83 +/- 0.035	0.86 +/- 0.032	0.92 +/- 0.008	0.84 +/- 0.005
	RP	0.85 +/- 0.010	0.83 +/- 0.011	0.80 +/- 0.003	0.84 +/- 0.002
	R2	0.71 +/- 0.017	0.69 +/- 0.019	0.64 +/- 0.005	0.71 +/- 0.004
ANN	RMSE	0.87 +/- 0.037	0.90 +/- 0.047	0.93 +/- 0.021	0.87 +/- 0.033
	RP	0.83 +/- 0.012	0.82 +/- 0.019	0.79 +/- 0.009	0.83 +/- 0.014
	R2	0.69 +/- 0.020	0.67 +/- 0.031	0.63 +/- 0.015	0.69 +/- 0.022

Tabela C.1: Resultados da regressão para os descritores moleculares 1D e 2D

1D 2D FP		Non Stringent	Stringent	Non Stringent Blind Test	Stringent Blind Test
RF	RMSE	0.85 +/- 0.029	0.88 +/- 0.033	0.95 +/- 0.003	0.88 +/- 0.003
	RP	0.84 +/- 0.029	0.82 +/- 0.011	0.79 +/- 0.001	0.83 +/- 0.001
	R2	0.70 +/- 0.018	0.68 +/- 0.018	0.62 +/- 0.002	0.69 +/- 0.002
SVM	RMSE	0.84 +/- 0.033	0.85 +/- 0.037	0.88 +/- 0.002	0.82 +/- 0.002
	RP	0.84 +/- 0.010	0.83 +/- 0.012	0.82 +/- 0.001	0.84 +/- 0.001
	R2	0.71 +/- 0.017	0.70 +/- 0.019	0.67 +/- 0.001	0.70 +/- 0.001
XGBoost	RMSE	0.83 +/- 0.035	0.84 +/- 0.032	0.92 +/- 0.003	0.86 +/- 0.005
	RP	0.84 +/- 0.010	0.84 +/- 0.009	0.80 +/- 0.001	0.84 +/- 0.002
	R2	0.71 +/- 0.017	0.70 +/- 0.015	0.64 +/- 0.002	0.70 +/- 0.004
ANN	RMSE	0.85 +/- 0.040	0.87 +/- 0.047	0.91 +/- 0.017	0.88 +/- 0.008
	RP	0.84 +/- 0.013	0.83 +/- 0.015	0.81 +/- 0.007	0.83 +/- 0.003
	R2	0.70 +/- 0.021	0.68 +/- 0.024	0.65 +/- 0.011	0.69 +/- 0.006

Tabela C.2: Resultados da regressão para os descritores moleculares 1D e 2D e FP

C.2 Classificação

1D 2D		Non Stringent	Stringent	Non Stringent Blind Test	Stringent Blind Test
RF	Accuracy	0.8910	0.8770	0.8645	0.8863
	Precision	0.8875	0.8745	0.8519	0.8804
	Recall	0.8954	0.8813	0.8808	0.8897
	f1-score	0.8914	0.8779	0.8660	0.8852
SVM	Accuracy	0.8436	0.8364	0.8357	0.8420
	Precision	0.8979	0.9001	0.8869	0.8974
	Recall	0.7758	0.7583	0.7687	0.7670
	f1-score	0.8310	0.8217	0.8214	0.8259
XGBoost	Accuracy	0.9027	0.8950	0.8721	0.8979
	Precision	0.9026	0.8944	0.8821	0.8943
	Recall	0.9026	0.8969	0.8771	0.8991
	f1-score	0.9026	0.8956	0.8770	0.8967
ANN	Accuracy	0.8866	0.8789	0.8629	0.8752
	Precision	0.8878	0.8804	0.8586	0.8774
	Recall	0.8854	0.8783	0.8676	0.8686
	f1-score	0.8847	0.8792	0.8629	0.8727

Tabela C.3: Resultados da classificação para os descritores moleculares 1D e 2D

1D 2D FP		Non Stringent	Stringent	Non Stringent Blind Test	Stringent Blind Test
RF	Accuracy	0.8923	0.8850	0.8639	0.8843
	Precision	0.8905	0.8845	0.8517	0.9012
	Recall	0.8945	0.8824	0.8794	0.8697
	f1-score	0.8925	0.8834	0.8653	0.8851
SVM	Accuracy	0.8956	0.8911	0.8860	0.8939
	Precision	0.8874	0.8861	0.8737	0.9048
	Recall	0.9062	0.8948	0.9013	0.8868
	f1-score	0.8965	0.8901	0.8871	0.8954
XGBoost	Accuracy	0.9059	0.8974	0.8756	0.8936
	Precision	0.9064	0.9023	0.8720	0.9140
	Recall	0.9052	0.8883	0.8790	0.8747
	f1-score	0.9058	0.8953	0.8755	0.8939
ANN	Accuracy	0.8968	0.8902	0.8705	0.8821
	Precision	0.8968	0.8982	0.8653	0.9055
	Recall	0.8969	0.8770	0.8761	0.8598
	f1-score	0.8968	0.8874	0.8706	0.8820

Tabela C.4: Resultados da classificação para os descritores moleculares 1D e 2D e FP