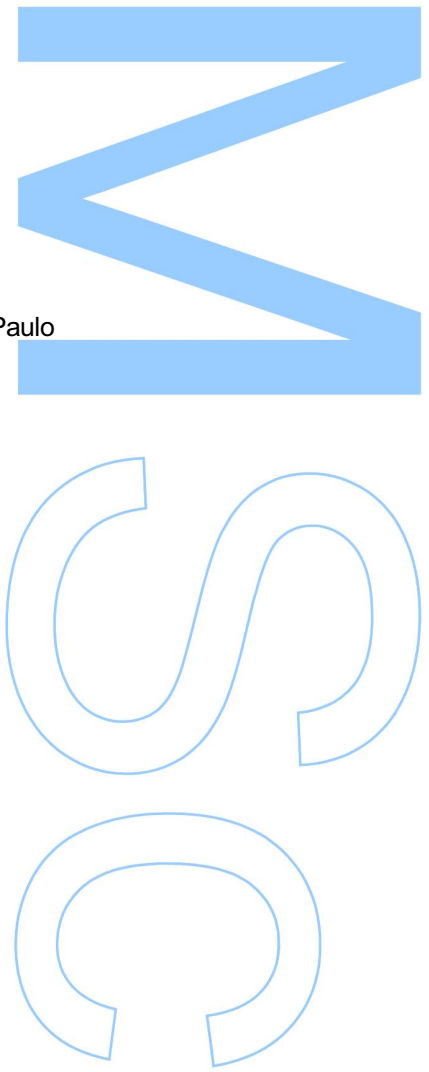
Extraction of a Bipolar Disorder associated genetic pattern

Luís Miguel Mendes Moreira
Mestrado Integrado em Engenharia de Redes e Sistemas Informáticos
Departamento de Ciência de Computadores
2018

Orientador
Inês Dutra, Professora Auxiliar, Faculdade de Ciências da Universidade do Porto
Coorientadores
Rodrigo Dias, Investigador, Faculdade de Medicina da Universidade de São Paulo
Camila Nascimento, Investigadora, Faculdade de Medicina da Universidade de São Paulo



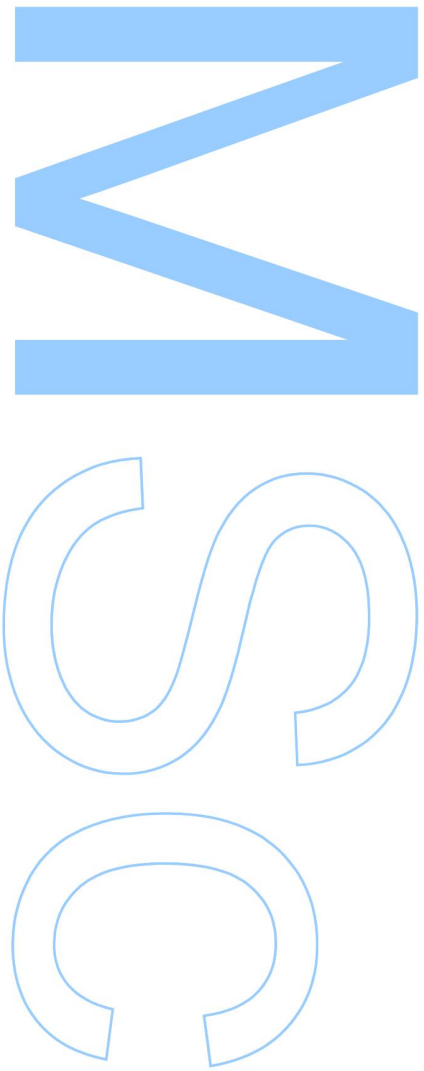
U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, _____/_____/_____



Dedico aos meus pais, ao meu irmão, à minha namorada e em especial à minha avó.

Acknowledgments

Aqui, e em português, agradeço à professora Doutora Inês de Castro Dutra por me ter orientado durante a realização e execução deste projeto: sempre disposta a ajudar a professora batalhou arduamente todos os momentos em que tudo parecia ficar mais difícil ou mesmo impossível de se resolver. Um grande e especial obrigado por me ter escolhido como seu orientando e executante deste projeto.

Quero deixar um especial obrigado ao Doutor Rodrigo Dias e Doutora Camila Nascimento por me terem co-orientado neste projeto que requer um enorme conhecimento biológico e genético, e não sendo esta a minha área de especialidade, vocês foram extremamente importantes no que diz repeito a ilucidar os meus conceitos teóricos e a indicar-me qual o melhor caminho a seguir. Um grande e especial obrigado a ambos.

Aos meus pais, José António e Maria de Fátima, e ao meu irmão, António José, agradeço todo o apoio dado desde que iniciei a minha caminhada no mundo académico: sempre me incentivaram a ir mais além e a cumprir todos os objetivos traçados por mim próprio para minha vida. Com muito carinho dedico-vos este projeto que marca o final de mais um dever cumprido.

À minha namorada, Rute Loureiro, agradeço uma, duas, mil vezes, por todo o apoio dado desde que iniciei este longo projeto. Tenho a certeza de que não seria capaz de o terminar sem te ter ao meu lado: foste extremamente importante, única e incansável em todos os momentos em que tudo parecia perdido. Tu estiveste lá, suportaste durante todo este tempo a minha ausência e excessivo tempo dedicado a trabalhar neste projeto. És tudo para mim, obrigado meu amor.

À minha avó deixo uma especial dedicatória: obrigado avó pois sei que sempre quiseste o melhor para mim e me deste força para ir em frente e concretizar mais um objetivo de vida e mesmo estando tão distante sempre me apoiaste em todos os momentos que mantivemos contacto. Um beijinho especial do seu neto Luís Miguel.

Luís Miguel Mendes Moreira
Setembro, 2018

Abstract

Bipolar Disorder is a widely known psychiatric disorder with an onset in adolescence or early adulthood. Also known as manic-depressive illness, bipolar disorder is one of the most common severe mental illnesses that triggers involuntary mood reactions affecting mainly the ability of dealing with daily tasks: it is indeed often to experience educational difficulties, job related problems, interpersonal difficulties, psychosocial dysfunction, marital issues, multiple suicide attempts, completed suicide and medication side effects [24] [50]. It has been estimated that there is a recurrence factor on 90% of the patients diagnosed with bipolar disorder during their lifetime [52] and this is mostly related with a faulty medication system induced by disease's comorbidity, where one or more psychiatric pathologies are found to be related with a primary disorder. Bipolar disorder is often unrecognized and misdiagnosed [59], hence more unexpected relapses fall into the need of being submitted to a new diagnostic process in order to find which medication should be prescribed to a certain patient, avoiding worst case scenarios where these start to have unexpected outcomes like suicidal behaviors.

The growth of *machine learning (ML)* opens a whole new world when trying to solve problems within this area. Databases containing patients' clinical and genetic information have been disclosed for researching purposes (as an example, *BDGene* [10] and *StepBD* [46]). A few studies in the literature have applied machine learning methods to classify patients according to their mood [2] and to predict relapses [47] whereas others approached bipolar related disorders such as schizophrenia [48]. In this work, we focus on patient's clinical-genetic data disclosed by the *Wellcome Trust Case Control Consortium (WTCCC)* [13]. Some works have studied patients genetic data using this same data set whilst aiming to find gene sequences associated with bipolar disorder [36]. These studies are limited to the application of a restricted set of genetic data whilst applying machine learning methods. In this work we would like to advance the area by (1) applying an unsupervised machine learning approach in order to confirm genetic patterns mentioned in the literature or discovering new associations, (2) revealing a possible gender-disease direct relation that matches the specific clinical-genetic features whilst highlighting subgroups similarities, and (3) discovering gene signatures of individual patients in order to reduce first cases misdiagnosis rate.

Keywords: Bipolar Disorder, Data Mining, Machine Learning, Unsupervised Learning, Chromosome, Genetic Variants, Genotype.

Contents

List of Tables

List of Figures

Acronyms

AKA Also Known As

AUC Area Under Curve

BD Bipolar Disorder

CPU Central Processing Unit

DCC Departamento de Ciência de Computadores

DNA Deoxyribonucleic Acid

DSM Diagnostic and Statistical Manual

FCUP Faculdade de Ciências da Universidade do Porto

GB GigaByte

GWAS Genome-Wide Association Study

IDE Integrated Development Environment

IOT Internet of Things

KDD Knowledge Discovery in Database

KB KiloByte

MAF Minimum Allele Frequency

MB MegaByte

ML Machine Learning

PGC Psychiatric Genomics Consortium

RAM Random-Access Memory

ROC Receiver operating characteristic

SNP Single-Nucleotide Polymorphism

SSH Secure Shell

SVM Support Vector Machine

SWAP Portion of a Hard Disk Drive

WSS Within-Cluster Sum of Square

WTCCC Wellcome Trust Case Control Consortium

Chapter 1

Introduction

Mental disorders have been around for a long time. For instance, *Aretaeus* of *Cappadocia* was a Greek physician and one of the first to detail symptoms in the medical field regarding mental illness during the 1st century. For many years his discoveries were left forgotten, but back then they were already giving evidence for the existence of a link between mania and depression.

"...in those periods of life with which much heat and blood are associated, persons are most given to mania, namely, those about puberty, young men, and such as possess general vigour." - Aretaeus of Cappadocia, ca. 150 A.D. [16]

During that period, religious beliefs had such a big influence on people's thinking - it was common to see people being executed for having some sort of mental disturbance that could not be explained by science, thus these were considered cursed. As science kept evolving, these tragedies no longer happened once people realized that there was a scientific explanation for these abnormal behaviors - Mental Disorders.

Nowadays these are categorized and properly detailed in several books, scientific magazines and research articles. In fact, The American Psychiatric Association have mentioned in one of their publications, *Diagnostic and Statistical Manual (DSM) of Mental Disorders* [3], that a mental disorder is a syndrome characterized by clinically significant disturbance in an individual's cognition, emotion regulation, or behavior that reflects a dysfunction in the psychological, biological or development processes underlying mental functioning. These are categorized by the DSM-5 model - This model uses five different categories/axis in order to better describe a mental disorder:

- Axis I - Clinical disorders;
- Axis II - Personality disorders and intellectual disabilities;
- Axis III - General medical conditions;
- Axis IV - Psychosocial and environmental problems;

- Axis V - Global assessment of functioning scale.

This work focus mainly on axis I regarding one of the most common mental disorder - Bipolar Disorder.

The World Health Organization has revealed that about 60 million people around the globe have been diagnosed with bipolar disorder and, considering the latest United Nations world population revision, numbers point out for the existence of approximately 7.6 billion people worldwide [41]. This means that bipolar disorder is affecting 0,78% of the current world population considering a limitless age range. It is a disorder with a lifetime prevalence of approximately between 1% and 2% with a high risk of recurrence manic and depressive episodes, elevated risk of suicide and a substantial high percentage of heritability as well, between 60% and 80%, but it is not all bad news: it is also highly treatable usually when recurring to the use of anti-psychotic medications and psychosocial therapy [26].

Bipolar disorder is highly associated with an elevated amount of costs per year. A study was realized upon the economic burden of this disease in the *United States of America* and an article was published on *Journal of Affective Disorders* [7]. With regard to Bipolar I Disorder the main results point out for an annual total cost of 202.1$ billion in 2015 corresponding to an average cost of 81,559$ per person. These results are directly related with the following outgoing:

- Caregiving (36%);
- Direct healthcare (21%);
- Unemployment (20%).

It has been concluded that it is indeed possible to perform a cost reduction by providing efficient treatments and also by having better health strategies, in order to optimize the usefulness of the health system and each country's resources [12].

In fact, we intend to improve current health system by making use of a bespoke machine learning model which will be applied on an European cohort of 1998 bipolar disorder samples. Such data was disclosed by WTCCC, upon agreement between both parties, and as far as we know only a few work has been made around bipolar disorder cases, whilst applying a machine learning model that aims for pattern extraction by finding clinical-genetic similarities between subgroups of bipolar disorder samples. Unsupervised Learning algorithms allow us to reveal such subgroups in which we can then apply our local analysis. For instance, we have used K-means algorithm which we found to be suitable for this task, with regards to *Knowledge Discovery from Data (KDD)*, as we intend to retrieve subgroups which were not identified in recent studies, by performing a clustering analysis.

In this work, beside highlighting relevant subgroups, we also extract knowledge out of the data set provided, that would not be possible to acquire without such technology. As a matter of

fact, we reveal potential indicators of a direct relation between multi-factor clinical features, in particular gender, and genetic variants present in someone's genetic code. Such work has never been accomplished, and we are most likely the first ones to do it when taking in consideration the WTCCC data set research studies history and the number of chromosomes that were simultaneously involved in our research. In addition, we have investigated several genes that have been already associated with presence of bipolar disorder: DISC1 in [35], ARPP21 in [49],GABRB1 in [45], ANKRD46 in [5], ANK3 in [27] [17], CACNA1C in [27] [51] [17], DUSP6 in [28], GRIN2B in [60] [40] [31] [34] and SYN3 in [39]. As a result, multiple clusters are retrieved in which each of these groups include samples' clinical features and genetic variants genotypes. Each genetic variant has been carefully analyzed by using a single table for each SNP so patterns could be extracted and knowledge retrieved accordingly.

In the next chapters we present *Fundamental Concepts* (Chapter 2) that we find to be essential throughout this project development; we present current *State of Art* (Chapter 3) that is directly and indirectly associated to this ML study; *The effect of genetics on bipolar disorder* (Chapter 4) are described prior to the actual experience as several factors were taken in consideration in order to complete the KDD assignment which is described in detail under *Methodology and Experiments* (Chapter 5); lastly, we present this project main results and highlights under *Conclusion* (Chapter 6) as well as *Future Work* (Chapter 7) that needs be carried on in order (1) to retrieve more information than the one extracted by ourselves and (2) to improve current implementation so we can achieve better performance and results accuracy.

Chapter 2

Fundamental Concepts

Mental Illness is a very common and treatable health problem that regards both humans' physical and psychological features. It affects the way you think of yourself and the world in general, as you can either start looking at it as if nothing really matters in this precise moment, or you can start thinking that every single moment ahead of you will be great and full of grandiosity and these might take over your ability to deal with your daily duties, which can affect both your life and the life of others.

Two main types of depression are mostly referred to in the literature as: Unipolar Disorder and Bipolar Disorder.

We often make the assumption that these are very distinguishable disorders, which they are not. In fact, fairly recent evidence suggests that both have their own properties but there are still many similar characteristics when it comes to describe one disorder and another. This factor becomes extremely important when we try to assess someone's clinical health condition in order to classify its mental illness, as we could misclassify it as unipolar disorder instead of bipolar disorder, and vice-versa. However, unipolar disorder episodes are mostly described as periods of depressive behaviors that are not followed by high or energized moments [55].

In the next section we will approach one of the most common types of depression: Bipolar Disorder.

We will start by presenting some important genetic and bipolar disorder concepts that were used during this project fulfillment, which include revealing risk factors, symptoms, treatments and clinical statistic results, and we will end it by describing some of the most important concepts of both data mining and machine learning.

2.1 Genetics

Genetic components it is what every single living organism is made up of and it is also what makes us all different and unique within the same ecosystem. It is essential to understand elementary genetic concepts when working with technologies in the field of genetics as lack of theoretical knowledge could lead to misleading results. So we will start our terminology by presenting the most elementary genetic concept: Cell.

Living organisms main structures are composed of cells in which these contain specific instructions coded inside the *deoxyribonucleic acid (DNA)*. DNA is a set of two complementary strands made up of four nucleotides, entitled Adenine, Cytosine, Thymine and Guanine, mostly referred as simply A, C, T and G, in which A matches T and C matches G, therefore these nucleotides conceive base pairs consisting on an individual genotype, hence a phenotype (set of characteristics) is established. As an example, lets consider two children's blood type inherited from their parents: for instance, if their father is a blood type AB individual and their mother is also a blood type AB individual, then both of them have a change of 25% of being a blood type A, 50% of being a blood type AB and 25% of being a blood type B, as presented in Table 2.1 - father's and mother's blood type regards table's top and left side respectively.

	A	**B**
A	A	AB
B	AB	B

Table 2.1: Blood type inheritance.

All the genetic information inside a cell's nucleus is packed into a bigger structured entitled chromosome in which is presented with a set of two, one inherited from our father and another one inherited from our mother. In total there are twenty three pairs of chromosomes, in which twenty two are autosomes and the last one is a sex chromosome (XX or XY).

These genetic structures are made of a variety of genes that contain specific and unique sequences present in our DNA, which encode instructions regarding proteins creation process. It has been estimated that the human genome contains between 20000 and 25000 genes [19]. Each of these genes are selectively unique, as an example, active genes in brain cells differ from genes in heart cells simply because these genes perform unique tasks within our organism, hence they produce unique proteins.

Most of the already known diseases have a genetic pattern associated to it. As it has been mentioned, inside each gene there is a specific DNA sequence that encodes certain instructions to be carried on by each of their proteins, that includes how well food and chemicals are metabolized, how well toxins are detoxified, and many others. A single mutation inside a gene

causes a malfunction within our organism, hence it might potentially reveal disease symptoms. These are associated with something so simple as having a single nucleotide change within a genetic sequence, whereas some disease may be revealed when a deletion or duplication occurs within a genetic sequence. This modified protein will still be able to function but with reduced capacity, however some proteins might even become completely disabled or reveal a new function different from its original purpose [19].

These gene variants are not always considered disease-related subjects, in fact, 99.9% of the human genome is identical among all individuals [30]. Therefore, gene variants its what makes us different and unique within the human population as they produce certain physical and psychological traits which differ from one individual to another within the same species population. Two forms of gene variants are:

- Mutations;

  In a general sense these are associated to diseases.

- Poymorphisms.

  Genetic variations that produce certain physical and psychological traits which result in having different individuals within the same species population.

Single Nucleotide Polymorphisms (SNPs) are genetic variants that regards one single nucleotide variation within a DNA sequence. It has been estimated that a single individual might carry millions of SNPs. This single variation may incur an increased disease risk, therefore could lead to a disease development in an individual that originally was not meant to it. Most of the SNPs have no effect on well-being or disease risk/development, however some have been helping scientists and researchers to reveal associations between SNPs inside certain genes and disease future development or response to specific drugs for disease treatment purposes.

There are still a few old diagnosis methods that rely only in the health professional diagnostic results, hence some of the diagnostics are not accurate and in some cases can incur in providing the wrong medication to a certain patient. However, technological advances in the filed of genetics have been crucial whenever it comes to create the possibility of performing early diagnostics, developing new treatments and reducing first cases misdiagnosis rate, where bipolar disorder, in particular, has been given a lot of attention.

In the next chapter we will present crucial concepts to better understand bipolar disorder as a whole within the mental illness field.

2.2 Bipolar Disorder

Formerly called manic depression, bipolar disorder is a mental disorder that causes extreme changes on our current mood, energy and activity levels. These mood changes range from extremely energized periods (also known as manic episodes) to very dark and low periods (also know as depression). It is not a character flaw or a sign of personal weakness, as anyone can experience mood changes, but a bipolar disorder depressive episode would hardly affect our ability as human beings of dealing with daily duties, such as going to work or even taking care of our family and ourselves. Although, these mood changes may vary from long periods of humor swings, down to short periods of prevalence, considering episodes that last a few hours, minutes or even seconds [42]. These episodes are associated with moments of a manic occurence, although scientific research mention unsual mood changes where people experience, simultaneously, energized and depressive moments within one single manic episode [21].

As like any other mental illness, also bipolar disorder has been classified into different types. *The National Institute of Mental Health* defines four different categories/types, however only the first two are primarily discussed in the scientific literature [42]:

- Bipolar I Disorder - describes manic episodes that last for at least 7 days, or either episodes that are so severe that the patient needs immediate hospital assistance;
- Bipolar II Disorder - describes episodes with a depressive pattern, as well as hypomanic episodes where abnormal levels of high energy might be spotted, however not the full episode occurs like mentioned in bipolar I disorder cases;
- Cyclothymic Disorder - describes multiple hypomanic episodes that last for at least two years;
- Other specified and unspecified bipolar-related disorders - this last category is used when it is not possible to fit someone's diagnostic within one of the first three categories. These are also treated as rare cases of bipolar disorder.

The most significant differences between these two types of bipolar disorder resides on the fact that a person with bipolar I disorder has manic episodes, while someone with bipolar II disorder has hypomanic episodes - its the severeness of an episode and its time of prevalence that differs from one to another.

As pictured in Figure 2.1, during a bipolar I episode that same person might experience moments of depression and mania. When one of these two moments reach very high/low values of moodiness (evidenced in Figure 2.1) usually that is the moment when people get into such a bad health state, hence He/She should be either treated immediately by a health professional or should take their medication, in a case where it has been already prescribed.

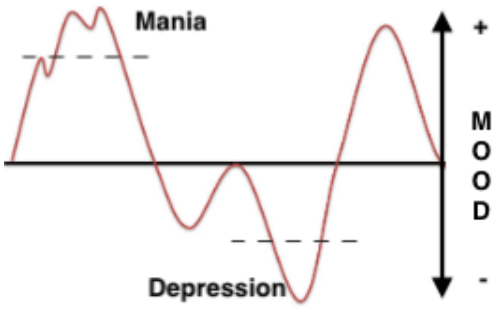


Figure 2.1: Bipolar I Disorder episode mood pattern.

In comparison with a bipolar II episode, pictured in Figure 2.2, people do experience relative patterns of moodiness, although these are less severe and prevail for a brief period of time. A patient that have been diagnosed with bipolar II disorder would experience hypomania moments, where it would feel higher levels of energy than the usual, or it would feel more confident and happy than the usual, but would never reach insane values as these are cases of bipolar I disorder. Even whilst not reaching a mania state this patient could still get into a depressive state and it could prevail longer in this same state considering previously described bipolar I disorder episode.

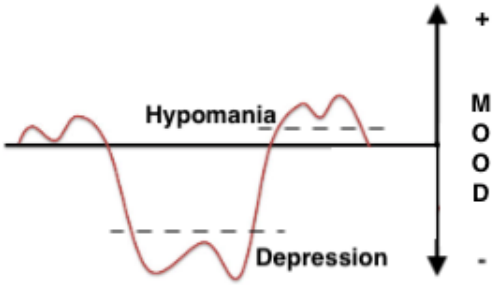


Figure 2.2: Bipolar II Disorder episode mood pattern.

The best way to avoid these raw moments, for those who were diagnosed with bipolar disorder, is to have family members, close friends and coworkers, that are completely aware of these patients' health situation and they must take action once they notice a certain abnormal behavior, such as to help out these patients on providing a quick access to a health institution or medication. Based in one single individual we could witness unusual and very likely new patterns. In fact, as mentioned before, one single patient could experience, simultaneously, rapid cycling and mixed state symptoms.

Patients that have been successfully diagnosed with bipolar disorder experience symptons that fall more often into the manic episodes kind, and its very likely that these might differ from one person to another [42]. With regard to axis I, whilst experiencing a manic episode these patients may reveal some of the following symptoms:

- Feel very energetic;
- Have increased activity levels;
- Have difficulties to fall asleep;
- Become irritable very easily;
- Start thinking that it is possible to do everything at once;
- Have an excessive willing of spending money.

In the other hand, these are the most common symptoms that a patient would experience during a depressive episode:

- Have a feeling of emptiness, loneliness;
- Have no energy to do anything;
- Have either difficulties to fall asleep or the will to sleep too much;
- Have concentration problems;
- Have no food habits so they may eat too much or nothing at all;
- Have periodical thoughts of death.

It is very important to preserve that during bipolar disorder early stages its symptoms are mistakenly associated with other factors, which are not necessarily related with this disease prevalence: at a first glance, a health professional would look to a depressive episode as related with alcohol or drug abuse, or maybe with an unsatisfactory performance level at work, which it could be true, but unfortunately there are many clinic cases that remain to be diagnosed due to a bad first interpretation of the real problem, hence a high misdiagnosis rate follows this disease clinical assessment.

There are many risk factors associated with this mental illness development. In fact, a *National Institute of Mental Health* research reveals that there is a set of causes that contribute for becoming a bipolar disorder carrier, rather than having a single reason for its progression [42]. These risk factors are also very complex as they usually vary from one person to another, hence achieving a bipolar disorder accurate diagnostic becomes even more harder than what already is. Therefore, we present a set of main categories that are highly related with a bipolar disorder episode recurrence:

- Brain structure and functioning;
  - Alcohol and/or drugs abuse will not cause a bipolar disorder episode, but these factors can lead to experience abnormal behaviors and unusual mood shifts. Since medication is used to prevent relapses the first thing that any health professional will demand to do is to perform a detox from any harmful substances so then a proper diagnostic could be performed. We also consider physical and psychological traumatic experiences as triggers of a bipolar disorder onset, such as sexual harassment/abuse, the death of a close person, the lost of a job position or a an important life opportunity.
- Genetics;
  - Specific genetic sequences present in human DNA are strong indicators of a possible bipolar disorder development/recurrence. As an example, several researchers pointed out CACNA1C as one gene that contain genetic variants which are highly related to this disease development/recurrence [18].
- Family history.
  - Having a direct relative that experienced bipolar disorder episodes in the past is also an important factor for any health professional diagnostic, since it may indicate that other family members might experience similar bipolar disorder episodes as this is a very heritable disease.

All these symptoms and risk factors lead us to the idea that a proper diagnostic needs to be done once any of these are visible in someone's behavior. A proper diagnostic significantly improves people's quality of life whether this disease is inhibit under a precise and correct treatment. As an initial step, to be taken by anyone who has the aforementioned symptoms, we must reach out to a psychiatrist. It is this health professional's job to provide the right health care by performing either a physical and/or mental assessment to achieve a more precise diagnostic. It is common to find people seeking for health professional assistance when they are already experiencing depressive symptoms or manic episodes, and in these cases it is very important to have a medical history based on this person's behavior. Unfortunately, bipolar disorder is very often primarily diagnosed as a mental-related disorder termed schizophrenia.

According to the National Institute of Mental Health [42], if left untreated, these symptoms tend to get worse and, having in consideration a long term scenario, it may lead to evolve from a slow cycling series of episodes to a rapid cycling series of episodes, where low and high stages tend to happen more often than the usual. Some patients may even reveal a critical behavior consisting of suicide attempts, drugs/alcohol abuse issues, broken relationships and many other disturbances.

A typical treatment usually consists in a combination of psychiatric medications and psychotherapy, and some form of treatment is usually needed throughout most of a person's adult life [37]. In fact, medication is one of the most important parts of a patient's treatment

as it helps on stabilizing moods and decrease the number of mental illness episodes. There are currently a few types of medication in world's health system, such as:

- Mood stabilizers - mostly used to treat manic or hypomanic episodes;
- Antidepressants - mostly used to treat depressive episodes;
- Antipsychotics - mostly used in addition to other medication.

Recovery strategies for this mental illness might reveal effectiveness flaws as never the same treatment can be applied to other patients: each treatment is unique as each person has their own characteristics that never work in the same way in comparison to others. In fact, one of the most important challenges rely on finding and maintaining a daily treatment for the same patient, as very often health professionals find the need of keeping one patient under different treatments along their lives. However, people have found themselves very comfortable when joining a support group that helps on clearing up their minds and keeping a healthy/stable environment.

The National Institute of Mental Health [42] mentions that there is no knowledge of bipolar disorder exact causes, however it is known that this is a highly treatable mental illness. It is important to build a healthy routine where these bipolar disorder patients found themselves controlled and in control of their current mood, as many can loose it and to start having relapses after thinking that no longer need their medication nor their treatments.

"Stay active, stay strong, stay positive." - Luis Moreira.

2.3 Data Mining and Machine Learning

We live in a world where vast amounts of data are collected, on a daily basis, in which Megabytes nor Gigabytes are enough to measure such amount of data - we often now see it in amounts of Zettabytes (1 x 10^{12} GB). This amount of data has being termed as *Big Data*. *Reuters* has recently revealed that the global growth of data in 2020 will be nearly 35 Zettabytes [4]. Due to this huge amount of produced data the term *Big Data* has been spread in a world wide scale.

Several companies make profit out of selling data, hence it has become one of the most profitable business model around the world, but having stored such a big amount of data without really making it useful it is pointless. There is no doubt that across several years data has been collected, for the better and for the worse, therefore we must consider that most of the cases, collecting it and making really useful out of it has helped us many more times than has been harmful, but it always should be an user decision to allow data collection. We are more than capable to perform important tasks over data, rather than just focus on profit, such as KDD that allow us to evolve and improve sensitive areas, some of which are the health area.

Machine learning is definitely one of the most spoken-about fields of Computer Science. Some examples show back in history that these same ideas were already in use: On our own history there is no better example of machine learning usage than *the Bombe* , *also known as (AKA) the Enigma machine*, built in 1939 by Alan Turin [57] who was (among other expertise) a mathematician and a computer scientist. During *World War II* this machine was designed to read and process encrypted German messages, hence a human readable message was produced upon completion of the decryption task. If we look at it in the same way that Alan did, this is nothing but a system that receives an input, processes it and generates an output. Although, there is an important detail to be mentioned: A machine learning model is considered valid if it generates most of the times the same output for different inputs as depicted in Figure 2.3.

Figure 2.3: Machine Learning schematic.

Almost 80 years have passed and we still look at machine learning tasks by using the

exact same logic but with even more complexity as many more algorithms were designed and implemented since that period of time.

Machine learning works alongside data mining, however these two fields have their own features and properties: it is often seen machine learning associated to model construction which involves the study of algorithms that can automatically perform knowledge extraction without human intervention, whereas data mining is carried out by a data scientist on a particular data set with a specific goal in mind. Usually, specific tasks like pattern recognition, that have been developed in machine learning, are in this person's focus and it is likely to be used along its work. Quite often, this data set is massive and it requires to go through a data mining process in order to reduce its complexity and easily become more understandable, hence a machine learning model can make use of a data set that follows the principle of a high quality data set: Accuracy, Completeness and Consistency [23].

Here we present some of the most used data mining algorithms [58] as well as a full review of their main features:

- C4.5;
  - It is one of the most used decision tree algorithms. It was developed between 1970 and 1980 by J. Ross Quinlan who is a machine learning researcher after its predecessor, ID3 (Iterative Dichotomiser 3). This algorithms has become a benchmark in a manner that other similar algorithms compare their performance against C4.5 performance.

    Both were independently created within the same period, although these two algorithms follow a similar approach with regards to learning decision trees from training tuples, as they carry out a greedy learning approach in a manner that decision trees are built in by adopting a top-down recursive divide-and-conquer strategy [23], in which a training set is recursively subdivided into smaller subsets of tuples.

- K-means;
  - It has been considered one of the simplest clustering analysis and partitioning algorithms that uses a centroid-based technique. It is this algorithm main goal to partition a data set D into k clusters $(C_1, C_2, ..., C_n)$ in which its elements are similar to each other within a cluster, but different between clusters.

    In order to carry out this partitioning task, first of all it defines the centroid of each cluster as the mean value of the points within it by randomly selecting k elements of data set D, therefore each of the remaining elements it will go through an evaluation process based on the Euclidean distance between the current element and the cluster mean. Then the algorithm itself will improve the within-cluster variation as it computes a new cluster mean using the elements previously assigned to it, hence all the following elements to be tested will then use this new cluster mean value. This algorithm iteration will end once the assignment task is stable, in other words, it will end once

a newly-formed cluster has the same properties as others that have been previously created [23].

- Support Vector Machine (*SVM*);
  - This is an algorithm that it was first presented on a 1992 paper, entitled *"A training algorithm for optimal margin classifiers"* and written by Vladimir Vapnik, Bernhard Boser and Isabelle Guyon [6], although there are references of previous work made on this algorithm prior to this paper issuing. SVM is an algorithm that uses nonlinear mapping in order to retrieve a higher dimension out of the original data, in which it searches for the linear optimal separating hyperplane that is able to separate our data set into two distinct classes with high accuracy [23].

    The pros and cons of this algorithm resides mainly on training time and high accuracy: this is an extremely complex algorithm that is able to model complex nonlinear decision boundaries but it requires a long period of training time. Despite this fact, it will perform flawless a modeling task with high accuracy, it is less propitious to data overfitting and it provides a compact description of the learned model.

    They can be used in numerical related tasks as well as for classification tasks: it has been seen in action in areas like handwritten digit recognition, object recognition and speech recognition.

- Apriori;
  - This is an algorithm known for its ability of finding frequent itemsets within a data set D that allows association rules discovery which highlist trending transactions within a data set D. It was proposed by R. Agrawal and R. Srikant in 1994 [1] for mining frequent itemsets for Boolean association rules and works as follows: first of all, we perform an initial full scan of the data set D in order to retrieve current items count and to use the ones that satisfy minimum support (usefulness of discovered rules) - this set will be denoted as L_1-itemsets (Level 1). Next, we use this current level to retrieve L_{k+1}-itemsets until no more frequent L_k-itemsets are found.

    The main disadvantage of this algorithm resides on the fact that it requires a full scan in order to get L_{k+1}-itemsets, however there is an important property that significantly improves this algorithm's efficiency by reducing its search space - *All nonempty subsets of a frequent itemset must also be frequent* [23]. This means that we also need to consider a *min_sup* variable that will be used to prune our L_k-itemsets in each iteration, in other words:

    $I \leftarrow k\text{-}itemset$;
    $if\ Count(I) \leq min\_sup\ then\ I\ is\ not\ frequent.$

We now present an example of this Algorithm application, depicted in Figure 2.4 and Figure 2.5:

TID	List of item_IDs
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3

Figure 2.4: Example of a data set D [23].

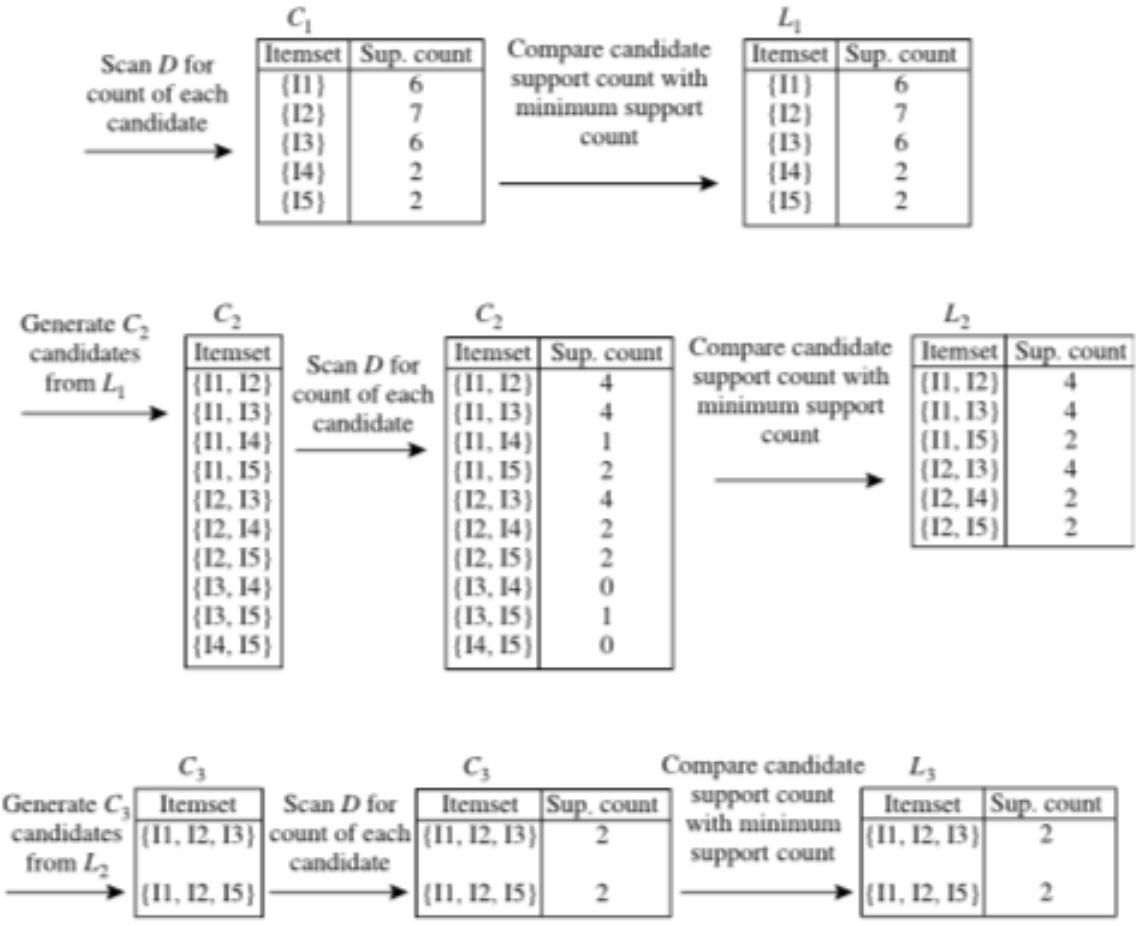


Figure 2.5: Example of Apriori algorithm application on a data set D [23].

Several variations have been applied to this algorithm in order to accomplish better execution efficiency, such as:

  * Hash-based technique;
  * Transaction reduction;
  * Partitioning;
  * Sampling;
  * Dynamic itemset counting.

- EM;

  – It is this probabilistic algorithm's main goal to find the maximum-likehood of a model parameters, whether our data has incomplete fields or some variables are hidden, such as *fuzzy* clusters that were not fully optimized. It was first explained and mentioned on a 1977 paper, entitled *"Maximum likelihood from incomplete data via the EM algorithm"* and written by Arthur Dumpster, Nan Laird, and Donald Rubin [14].

  With regards to previously mentioned clustering algorithm (*k-means*), a cluster will be retrieved after n iterations over our newly created cluster until no improvement can be obtained, but in some cases this algorithm will retrieve a *fuzzy* cluster.

  We now explain on how EM algorithms can improve our clustering methods, which it consists in two steps:

  1. Expectation step (E-step): Having in consideration current cluster centers, we perform an initial guess in order to retrieve a probabilistic distribution of our model. We expect to retrieve objects that belong to the closest cluster;
  2. Maximization step (M-step): During this step it is intended to maximize our cluster (reduce *fuzzy* data as much as we can), therefore this algorithm adjusts the center of each cluster so the sum of the distances from the objects assigned to this cluster and the new center is minimized [23].

  As expected, the E-step slowness is directly related with missing data ratio, thus it is this algorithm main flaw. Although, it has been applied in several ways within the statistic/probabilistic fields, such as:

  * Hidden Markov model assessment;
  * Gaussian mixture model assessment.

- PageRank;

  – This is an algorithm that was proposed by four computer scientists, that included Google's co-founder Larry Page in a 1997 paper, entitled "*The PageRank Citation Ranking: Bringing Order to the Web*" written by Lawrence Page, Sergey Brin, Rajeev Motwani and Terry Winograd.

  PageRank brought to the world wide web a feature that is still present in current search engines: is a method that will compute a page rank based on the graph of the

web [44], in other words, every time a user type on some random topic in a search engine, it will return a list of ranked, indexed and mined web pages according to that topic. In order to retrieve this ranked list, this algorithm perform as it follows (simplistic approach):

1. An initial probabilistic distribution is assigned to each page;
2. Each page and its outbound linked pages probabilities will be updated equally according to how often this pages are accessed.

Each PageRank value is calculated according to this formula (having in consideration three web pages depicted in Figure 2.6):

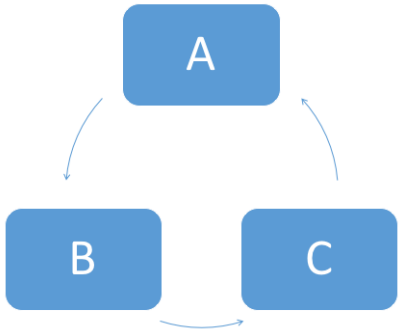


Figure 2.6: Example of a web that consists of 3 web pages.

$$PR(A) = 0.5 + 0.5 * P(C)/1$$
$$PR(B) = 0.5 + 0.5 * P(A)/1$$
$$PR(C) = 0.5 + 0.5 * P(B)/1$$

Some of PageRank algorithm pros and cons resides, in one hand, on the fact that this algorithm can perform measurements in a global scale and it is query independent, on the other hand, old web pages might get higher ranks.

- AdaBoost;
  - Adaptive Boosting (AdaBoost) is a boosting algorithm - it is it main purpose to boost our classification model accuracy in order to build a strong classifier $f(x)$ as a linear combination of weak classifiers $h(x)$, by using the following formula:

    $h_t(x) \leftarrow weak\ classifiers$
    $f(x) = \sum_{t=1} \alpha_t * h_t(x)$

As an example lets consider a data set D in which tuples are represented as it follows: $(X_1, y_1), (X_2, y_2), ..., (X_n, y_n)$, where X_n represents an actual tuple and y_n represents its class label. The first step consists in assigning to each training tuple an equal weight of $\frac{1}{n}$ in order to retrieve k classifiers we need to execute k iterations through the rest of the algorithm. For each k round, tuples will be sampled to form a training set D_k of size n, in which will be select according to their weight. These weights are adjusted according to how well these tuples were classified, in other words, if a tuple (X_n, y_n) is correctly classified then its weight will decrease, otherwise it would increase as this algorithm focus on all the misleading classifications.

In the very end, we might get an improved classifier, or a classifier that is less accurate than the original one, as the overfitting (refers to a model that trains too well) factor might affect the boosting task. There are other boosting algorithms, such as *Bagging*, that are less susceptible to overfitting. Overall, boosting algorithms tend to achieve better accuracy.

- k-Nearest Neighbors (*kNN*);

  – This is an algorithm that has been widely used in important areas, such as pattern recognition. It can be used for solving classification and regression problems: essentially having in consideration classification problems, this algorithm would output a discrete value, such as $k = 1$, hence a tuple would simply be assigned to that specific class, whereas for regression problems it outputs a continuous value, such as $k \geq 1$, which does not specify to which class should this test tuple be assigned to. This algorithm works as follows, with regards to a three tuples training set:

    1. Initially all training tuples are described by n attributes;
    2. Each tuple will then be spread out on a n-dimensional space, as depicted in Figure 2.7;

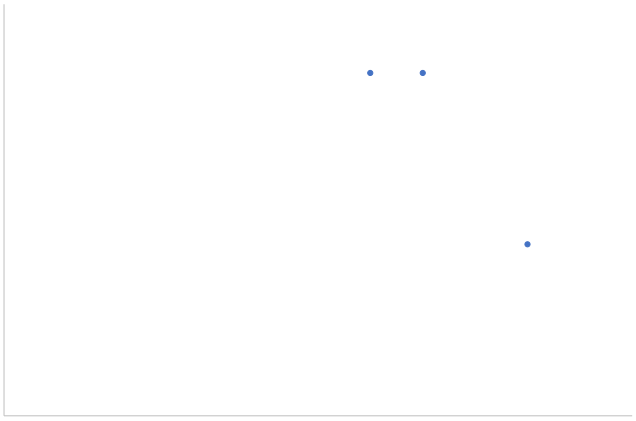


Figure 2.7: Example of a binary kNN classification training set.

3. Given an unknown test tuple, kNN will measure the distance between this new tuple and the current ones, by performing an Euclidean distance measurement between two points or tuples. As an example, lets say we have two tuples: $X_1 = (x_{11}, x_{12})$ and $X_2(x_{21}, x_{22})$. Then we could use the following formula:
$$dist(X_1, X_2) \ = \ \sqrt{\sum_{i=1} (x_{1i} - x_{2i})^2}$$
4. In the end, we should expect one single cluster or multiple clusters distinct from each other, as depicted in Figure 2.8 after receiving several unknown test tuples.

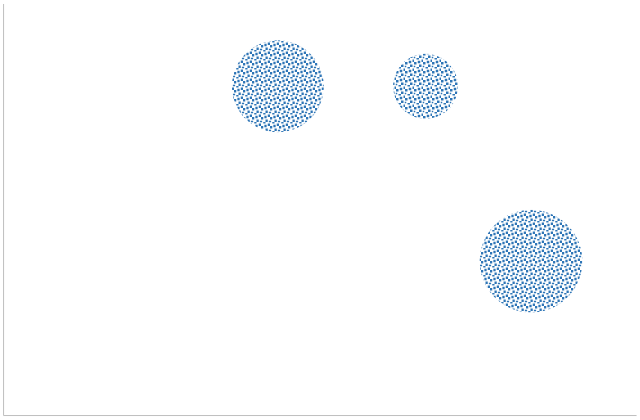


Figure 2.8: Example of a clustering visual output.

Unfortunately this algorithms suffers from a poor accuracy when irrelevant attributes are given to it. However, many improvements have been implemented in order to reduce the amount of noisy data, such as, *pruning*. Although, it is very important to select the proper distance metric, as Euclidean distance is not the only one available. As a matter of fact, many data scientists use Manhattan distance over Euclidean distance.

- Naive Bayes;
  - This algorithm belongs to the Bayesian classifiers category and is a simple version of a typical Bayesian classifier. Its main purpose is to estimate the probability that a given tuple belongs to a certain class by assuming that there is no effect of an attribute value from one class to another. This is also called: *class conditional independence*. In that sense, this version of the algorithm was entitled *naive*.

    As its name indicates, this is an algorithm that follows the Bayes' theorem: Lets say we have a tuple X, a class C and an hypothesis H, which designates the probability of having a tuple X that belongs to a specified class C. Thus, this theorem computes the probability of a tuple X belonging to a certain class C assuming that we already know the attribute description of X, and it is described by using the following mathematical formula:

$$P(H \mid X) = \frac{P(X \mid H).P(H)}{P(X)}$$

Where,

$P(X) \leftarrow Prior\ probability\ of\ X$;
$P(H) \leftarrow Prior\ probability\ of\ H$;
$P(X \mid H) \leftarrow Posterior\ Probability\ of\ having\ X\ given\ that\ H\ is\ true.$
$P(H \mid X) \leftarrow Posterior\ Probability\ of\ having\ H\ given\ that\ X\ is\ true.$

Overall, Bayesian classifiers have the minimum error rate in comparison to other classifiers [23]. However, having the class conditional independence factor it could create a disadvantage as that might be a wrongly made assumption.

- CART;
  - As previously described CART is also part of a set of decision trees algorithms. In comparison with C4.5 their main differences reside on the following facts:
    * CART always performs binary tests, whereas C4.5 allows two or more outcomes [58];
    * CART uses the Gini diversity index to rank tests, whereas C4.5 uses information-based criteria [58];
    * CART prunes trees using a cost-complexity model whose parameters are estimated by cross-validation [58];

    Results of a performance test between CART, C4.5 and Multi-Layer Perceptrons was published on a 1991 paper, entitled *"Comparison of three classification techniques, CART, C4.5 and Multi-Layer Perceptrons"*. Here data scientists used 256 data examples categorized into 4 classes:

    * class 1 - 37 data samples;
    * class 2 - 126 data samples;
    * class 3 - 84 data samples;
    * class 4 - 9 data samples.

    All 256 data samples were used as both training and testing samples. Here we present their results, depicted in Figure 2.9:

name	# of errors	accur %
cart	96	0.625
c4.5	105	0.59
mlp1	117	0.54
mlp2	47	0.82

Figure 2.9: Performance test results of CART, C4.5 and Multi-Layer Perceptrons algorithms [44].

CART seems that performed better than C4.5 back in 1991. However, after many years of research and development C4.5 has been improved and is currently used as a benchmark with regards to decision trees algorithms.

A machine learning model construction process goes through two main stages and once completed we should have a model/classifier that is able to predict a certain class label attribute for any new data income with reasonable accuracy:

- Learning stage;
- Classification stage;

2.3.1 Learning

During this stage we start to build a classifier or machine learning model by analyzing our current data set that is made up of tuples (or samples) and their associated class labels. A tuple, X, is represented by an n-dimensional attribute vector, $X = (x_1, x_2, x_3, ..., x_n)$ and each tuple X is assumed to belong to a predefined class, as determined by another database attribute, called the class label attribute. The class label attribute is discrete-valued and unordered. It is categorical (or nominal) in that each value is used as a category or class. The individual tuples making up the training set are referred to as training tuples and are randomly sampled from the database to be analyzed. In the context of classification, data tuples can be referred to as samples, examples, instances, data points, or objects [23].

From a high level perspective we classify learning algorithms as belonging into one of these two main categories: Supervised Learning and Unsupervised Learning.

2.3.1.1 Supervised Learning

This is a task where the class label attribute was acknowledged *a priori*. As an example, lets consider a data set that contains both lung cancer patients information and controls, such as:

$LUNG\_CANCER(name,\ age,\ sex,\ cancer\_type,\ age\_of\_diagnosis,\ lung\_cancer\_carrier)$
$LUNG\_CANCER(PatientA,\ 55,\ M,\ SCLC,\ 55,\ yes)$
$LUNG\_CANCER(PatientB,\ 30,\ F,\ ND,\ ND,\ no)$
$LUNG\_CANCER(PatientC,\ 60,\ M,\ NSCLC,\ 55,\ yes)$
$LUNG\_CANCER(PatientD,\ 75,\ M,\ NSCLC,\ 50,\ yes)$
$LUNG\_CANCER(PatientE,\ 20,\ M,\ ND,\ ND,\ no)$

- *SCLC*, stands for small cell lung cancers;

- *NSCLC*, stands for non-small cell lung cancers;
- M/F, stands for Male and Female respectively;
- ND, stands for Not Defined.

Our task is to build a machine learning model that can predictive lung cancer arise. If so, having in consideration this data set we have already acknowledged our class label attribute: *lung_cancer_carrier* (*yes* or *no*).

Our training data phase is ready and we can proceed by training our data set. We now present some examples of extracted classification rules:

$$IF\ age \geq 55\ AND\ sex = M\ THEN\ lung\_cancer\_carrier = yes$$
$$IF\ age \leq 30\ AND\ sex = F\ THEN\ lung\_cancer\_carrier = no$$

The main goal is to retrieve a function $y = f(X)$ which can predict the associated class label y for a given tuple X, thus we pretend to accomplish a function that can separate multiple data classes with enough accuracy. We often see in the literature this function to be represented in the form of classification rules, decision trees or mathematical formulas, as depicted in Figures 2.10, 2.11 and 2.12 respectively [23].

$$IF\ age >= 55\ AND\ sex = M\ THEN\ lung\_cancer\_carrier = yes$$
$$IF\ age <= 30\ AND\ sex = F\ THEN\ lung\_cancer\_carrier = no$$

Figure 2.10: Machine Learning - classification rules representation.

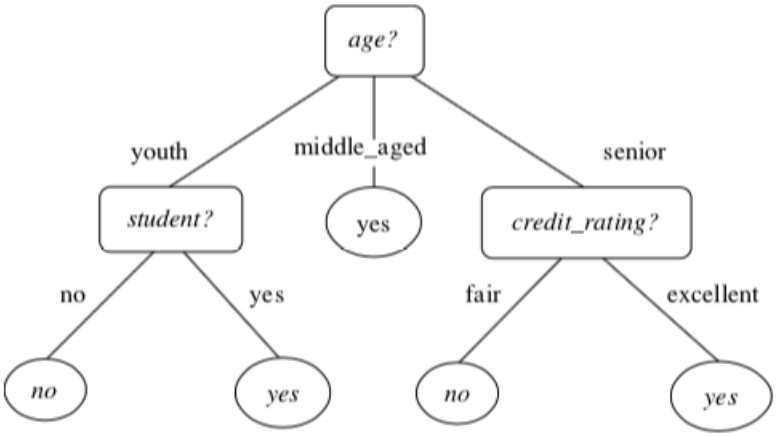


Figure 2.11: Machine Learning - decision tree representation [23].

$$y = \begin{cases} yes, & \text{if } age \geq 55 \wedge sex = M \\ no, & \text{if } age \leq 30 \wedge sex = F \end{cases}$$

Figure 2.12: Machine Learning - mathematical formula representation.

2.3.1.2 Unsupervised Learning

In contrast to supervised learning we often see unsupervised learning tasks in which the class label of each training tuple is not known, and the number or set of classes to be learned may not be known in advance. The main idea is that a computer can learn and extract patterns out of data whereas a human would not do be able to do it by itself - Artificial Intelligence.

As an example, looking at our previous data set, we could apply an unsupervised learning approach if our class label attribute was not known in advance, as shown in the following examples:

$LUNG\_CANCER(name,\ age,\ sex,\ cancer\_type,\ age\_of\_diagnosis)$
$LUNG\_CANCER(PatientA,\ 55,\ M,\ SCLC,\ 55)$
$LUNG\_CANCER(PatientB,\ 30,\ F,\ ND,\ ND)$
$LUNG\_CANCER(PatientC,\ 60,\ M,\ NSCLC,\ 55)$
$LUNG\_CANCER(PatientD,\ 75,\ M,\ NSCLC,\ 50)$
$LUNG\_CANCER(PatientE,\ 20,\ M,\ ND,\ ND)$

Once applied an unsupervised learning algorithm we could try to extract patterns out of this data which would lead us to the idea that some groups of rules are likely related to the fact of being/not being a lung cancer carrier. As an example we are going to apply one of the most used algorithms for unsupervised learning tasks: K-means clustering.

We will consider all our class labels as it follows:

- $sex \in \{1, 2\}$, for M and F respectively;
- $cancer\_type \in \{0, 1, 2\}$, for ND, SCLC and NSCLC respectively;
- $age\_of\_diagnosis \in \{0, age\_of\_diagnosis\}$, for ND, age_of_diagnosis respectively;

In result of applying k-means clustering on our data set we have achieved the following output:

$$PatientA, 55, 1, 1, 55, 1$$

$$PatientB, 30, 20, 0, 0$$

$$PatientC, 60, 1, 2, 55, 1$$

$$PatientD, 75, 1, 2, 50, 1$$

$$PatientE, 20, 10, 0, 0$$

As depicted in Figure 2.13 we can see that this algorithm successfully distinguished two main clusters (*Cluster id 1* and *Cluster id 2*) - k clusters value was set *a priori*. In comparison with our previous data set, where we present which of these tuples have either "yes" or "no" as a class label attribute, we can see that this algorithm assigned correctly these two Cluster ids without any previous knowledge (an open-source tool was used for demonstration purposes):

- 0, stands for "no";
- 1, stands for "yes".

Label	Vector	Cluster id	Cluster centroid
PatientA	55,1,1,55,1	1	63.33333333333333,1,1.6666666666666665,53.33333333333333,1
PatientB	30,2,0,0	0	25,1.5,0,0
PatientC	60,1,2,55,1	1	63.33333333333333,1,1.6666666666666665,53.33333333333333,1
PatientD	75,1,2,50,1	1	63.33333333333333,1,1.6666666666666665,53.33333333333333,1
PatientE	20,1,0,0	0	25,1.5,0,0

Figure 2.13: Machine Learning - K-means clustering.

2.3.2 Classification

Once completed the Learning stage we can now move forward to the Classification stage where test data will be used to compute our classifier accuracy: if a reasonable accurate machine learning model is retrieved we can then submit new data into our model (depicted in Figure 2.14) and testing it in real life situations, such as predicting the possibility of a bipolar disorder development on first diagnosis cases. In order to make this task possible an accuracy measurement needs to be done: there are several algorithms in which their main goal is to measure how well a machine learning model will perform when submitted to new data. We now present evaluation metrics used for accuracy assessment:

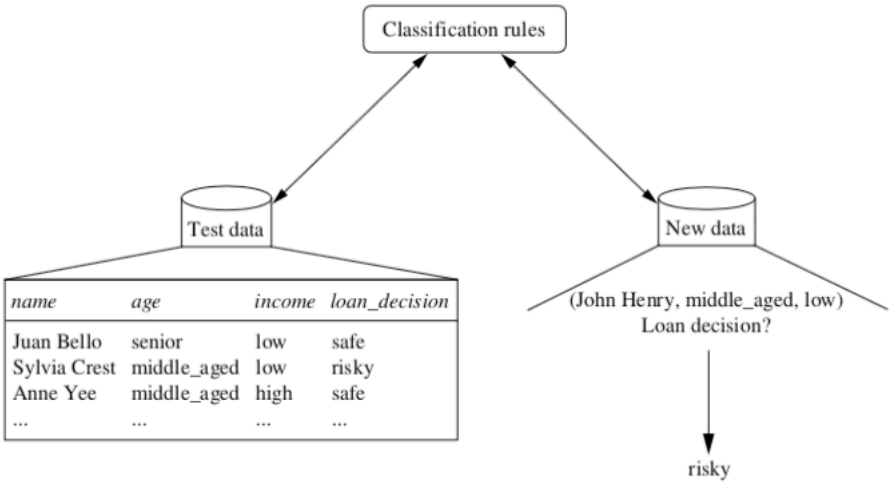


Figure 2.14: Machine Learning - classification stage [23].

- Evaluation metrics;
  - Having in consideration this subject the following evaluation metrics will now be presented: accuracy, sensitivity, specificity, precision, F_1 and F_β, depicted in Figure 2.15. Before jumping on to each measure details, there is an important terminology to be acknowledged which approaches positive tuples (tuples of the main class of interest) and negative tuples (all the remaining tuples). As an example, we may consider, in one hand, a class label $bipolar\_disorder\_development = yes$ as a positive tuple and, in the other hand, a class label $bipolar\_disorder\_development = no$ as a negative tuple.

Measure	Formula
accuracy, recognition rate	$\frac{TP+TN}{P+N}$
error rate, misclassification rate	$\frac{FP+FN}{P+N}$
sensitivity, true positive rate, recall	$\frac{TP}{P}$
specificity, true negative rate	$\frac{TN}{N}$
precision	$\frac{TP}{TP+FP}$
F, F_1, F-score, harmonic mean of precision and recall	$\frac{2 \times precision \times recall}{precision + recall}$
F_β, where β is a non-negative real number	$\frac{(1+\beta^2) \times precision \times recall}{\beta^2 \times precision + recall}$

Figure 2.15: Machine Learning - evaluation metrics[23].

In order to understand these metric's formulas we now describe each variable meaning:

* True positives (TP): It is addressed to positive tuples that were correctly classified by our model, in order words, our model assigns the class label $bipolar\_disorder\_development = yes$ to a tuple that should be classified as such;
* False positives (FP): It is addressed to negative tuples that were not correctly classified by our model, in order words, our model assigns the class label $bipolar\_disorder\_development = yes$ to a tuple that should be classified as $bipolar\_disorder\_development = no$;
* True negatives (TN): It is addressed to negative tuples that were correctly classified by our model, in order words, our model assigns the class label $bipolar\_disorder\_development = no$ to a tuple that should be classified as such;
* False negatives (FN): It is addressed to positive tuples that were not correctly classified by our model, in order words, our model assigns the class label $bipolar\_disorder\_development = no$ to a tuple that should be classified as $bipolar\_disorder\_development = yes$;

Once these are defined we may now build a confusion matrix - it can be viewed as a statistical tool that assess how well our model is capable of classifying different class tuples. A generic view is depicted on figure 2.16.

A confusion matrix is nothing less than a m by m table in which m represents the number of classes. Additional rows may be added, such as $Total$. As illustrated in Figure 2.16 this variable indicates the number of positive/negative tuples, represented by $P^{'}$ and $N^{'}$ respectively, in which $P^{'} = TP + FP$ and $N^{'} = FN + TN$. Hence, the total number of tuples is given by either $P^{'} + N^{'}$ or $P + N$.

Once we produce a confusion matrix out of our model results we shall now assess its reliability:

* accuracy: Also known as the percentage of tuples that were correctly classified by our classifier. We can retrieve this percentage by applying the following formula:

		Predicted class		
		yes	*no*	Total
Actual class	*yes*	*TP*	*FN*	*P*
	no	*FP*	*TN*	*N*
	Total	P'	N'	$P + N$

Figure 2.16: Machine Learning - confusion matrix[23].

$$accuracy = \frac{TP+TN}{P+N}$$

We can also retrieve our classifier's error rate out of the accuracy returned value by simply calculating $1 - accuracy(M)$, in which M represents our current model, or by applying the following formula:

$$error\ rate = \frac{FP+FN}{P+N}$$

Unfortunately *accuracy* by itself is not enough to assess a model's classification reliability: lets consider a model that has been trained to classify tuples of a medical data set as either $bipolar\_disorder = yes$ or $bipolar\_disorder = no$ - class imbalanced problem. If an *accuracy* of 93.46% is retrieved it might be misleading by the fact that this value is related to negative tuples instead of positive tuples. In such scenario the classifier could be correctly labeling only the tuples associated to $bipolar\_disorder = no$, instead of, $bipolar\_disorder = yes$. In order to avoid such misleading information we also should apply two extra evaluation metrics: *sensitivity* and $specificity$.

Once these two extra evaluation metrics are applied we should get the following accuracy formula:

$$accuracy = sensitivity\frac{P}{P+N} + specificity\frac{N}{P+N}$$

* sensitivity: having in consideration previously mentioned bipolar disorder class labels, this evaluation metric would tell us how well our model can classify the positive tuples $bipolar\_disorder = yes$, by applying the following formula:

$$sensitivity = \frac{TP}{P}$$

* specificity: Having in consideration previously mentioned bipolar disorder class labels, this evaluation metric would tell us how well our model can classify the negative tuples $bipolar\_disorder = no$, by applying the following formula:

$$specificity = \frac{TN}{N}$$

* precision: It is a measure of exactness that retrieves the percentage of tuples labeled as positive/negative that indeed are as such and it can be calculated by applying the following formula:

$$precision = \frac{TP}{TP+FP}$$

* recall: It is a measure of completeness that retrieves the percentage of positive/negative tuples that were classified as positive/negative respectively and it

can be calculated by applying the following formula:

$$recall = \frac{TP}{TP+FN}$$

Unfortunately, also precision and recall values may incur in error in a manner that when one increases the other decreases - inverse relationship. Lets consider, as an example, a scenario whilst getting precision/recall out of our bipolar disorder data set we retrieve a high value of precision, hence our model is more than capable of correctly labeling bipolar disorder tuples as such, but we may find a low recall value if it mislabels many other bipolar disorder tuples. Usually precision is used alongside recall, so we can compare precision values against fixed values of recall, and vice versa [23]. The bright side resides on the fact that it is possible to combine both measures into one single evaluation measure: F measure (also viewed as F_1 *score* or simply $Fscore$) and F_β.

* F and F_β measures: F measure uses the harmonic mean of precision and recall as it gives equal weight to both metrics, whereas F_β is a weighted measure, in a manner that assigns a β weight to recall that is β times much weighted than precision, so if we mention F_2 it simply means that recall weight is twice more than precision weight.
  These two percentages can be calculated by using the following formulas:

$$F = \frac{2 \times precision \times recall}{precision + recall}$$

$$F_\beta = \frac{(1+\beta^2) \times precision \times recall}{\beta^2 \times precision + recall}$$

All these metrics have their pros and cons and it goes as it follows: Accuracy is widely used and reliable when data classes are evenly and fairly distributed. For class imbalanced problems we may use other metrics, such as Sensitivity, Specificity, Precision, Recall (same as Sensitivity), F and F_β.
We shall now describe other evaluation metrics that focus their job on retrieving a more reliable accuracy evaluation.

- Holdout and random sampling;

  - These two methods approach the accuracy assessment task by using a splitting technique. However, holdout method was designed to be based on the splitting technique, whereas random sampling makes use of a subsampling technique. Both are now described individually:

    * Holdout: It makes use of the main idea behind splitting our data set into a training/test set as well as by using random sampling to estimate a mean accuracy, as it follows: first of all, it splits the given data into two sets - training/test set - usually on a $\frac{2}{3}$ and $\frac{1}{3}$ split ratio respectively. Then, it calculates the accuracy by using only the test set portion, hence the training set it will be used to derive the model. By using this method we approach the accuracy assessment with a pessimistic technique as only a portion of the initial data is used to derive the model.

* Random Sampling: It is a variation of the holdout method in which the holdout method is repeated k times. The overall accuracy is taken as the average of all iterations accuracy.

An illustration of this method is depicted in Figure 2.17.

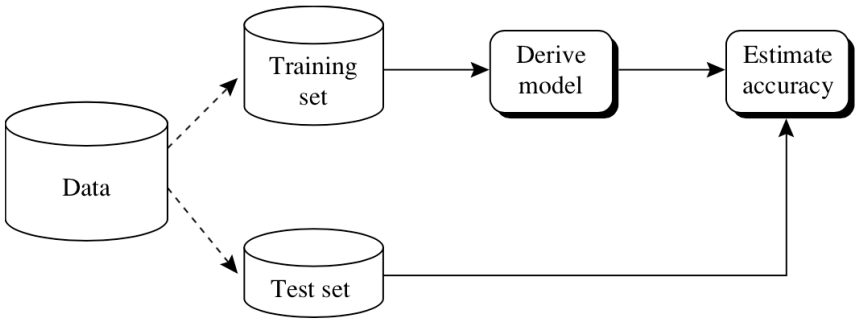


Figure 2.17: Machine Learning - holdout and random sampling [23].

- Cross-validation;
  - This method could be viewed as an improved version of the Holdout method where each subset is used k times as a test set instead of just once throughout the classification process.
    * k-fold cross-validation: Lets consider a data set D. The first step of this algorithm consists in dividing D into k subsets/folds in which all have approximately the same size.

      $$D = (D_1, D_2, D_3, \ ... \ , D_k).$$

      For each iteration a subset k is used as a test set and the remaining $k-1$ are used as training sets. By doing this, every single subset will be used exactly once as a test set, which it becomes one of the main advantages of this algorithm as it cares less with the way the data gets divided. However, this algorithm needs to run k times in order to retrieve all the desirable computations.
      After k iterations the accuracy is computed by having in consideration the overall number of correct classifications of all iterations divided by the total number of D tuple. Depicted in Figure 2.18 it is possible to see this algorithm behavior.

Figure 2.18: Machine Learning - K-fold cross-validation.

- Bootstrap.

  – This is an algorithm that was first introduced by Bradley Efron in 1993 [29]. Unlike other accuracy estimations, it samples training tuples uniformly with replacement, which means that during the sampling process a certain tuple X could be added to the training set more than one time.

  As an example lets consider a set of d tuples. Initially we sample or data set d times with replacement. Hence, it is expected to retrieve a bootstrap sample (training set) than contains repeated tuples. Therefore, all the remaining tuples that were not selected to be part of the training set will then build our test set and be part of it, hence the probability of a tuple not being chosen is $1 - \frac{1}{d}$, whereas the probability of being chosen is $\frac{1}{d}$. In a case where many tuples exist, then the probability of a tuple not being chosen tends to be approximately $e^{-1} = 0.368$ [29], in the other hand the probability of a tuple being chosen tends to be approximately $1 - e^{-1} = 0.632$ which these will then build up the training set (*.632 bootstrap*). Then the overall accuracy of our model is estimated by computing the following formula where $Acc(M_i)_{test-set}$ is associated to the model's accuracy retrieved from bootstrapping a sample i when applied to a test set, whereas $Acc(M_i)_{train-set}$ is associated to the model's accuracy retrieved from bootstrapping a sample i when applied to the original set of tuples.

$$Acc(M) = \frac{1}{k} \sum_{i=1}^{k} (0.632 \times Acc(M_i)_{test-set} + 0.368 \times Acc(M_i)_{train-set})$$

In the next chapter we review some of the scientific work around bipolar disorder that is directly and indirectly associated to our thesis theme.

Chapter 3

State of Art

Over the last years many more research studies have been made around the mental disorders scope: not just bipolar disorder in specific, but also other pathologies within the same diseases range and highly related with bipolar disorder, such as schizophrenia.

In the literature we have found some of the main research articles, regarding machine learning models used to improve patient's health care, whether they have been diagnosed with bipolar disorder or not, mentioning the ones where this disease hasn't been expressed yet. However, a certain patient might have the genetic information that indicates the possibility of a disease's near future development [25] or relapse tendency. We have faced some of the most wise methods that were used concerning this disease's main flaws, high misdiagnosis rate, in which we focus our study by finding patterns of interest between patients subgroups.

One of the main concerns whilst analyzing a data set resides on the fact that there are a lot of scattered data: some of them are not even valid and have been spread out across the scientific literature. For instance, [11] worked on a huge genetic database that reveals how bipolar disorder genetic information overlaps with other mental disorders, such as schizophrenia. They have mentioned that bipolar disorder is a common and severe psychiatric disorder characterized by cycles between bouts of mania and depression: it causes a significant impact on patient's lifestyle once it was discovered the high risk of suicide tendencies as one of these disorder's risk factors.

Therefore, we have also evidenced a particular interest in the ways and means used by some of these researchers concerning direct data collection of a bipolar disorder patient, in where the *Internet of Things (IOT)* field has a huge impact on improving current data acquisition mechanisms.

Whenever we think about data collection mechanisms, efficiency and reliability will always be two of the main concerns in reference to how well our collection system will perform. Over the past few years we have seen mobile platforms taking over most of modern machines' functionalities in order to produce the same output, but in this case by using a low-powered portable device. These mobile platforms have been used in many different fields, and mental illness is one of them - wearable devices are now capable of live capturing health data from an individual that has been

diagnosed with bipolar disorder, in a manner that allows researchers to get it and compute it without the need of waiting for this person's health status daily report - the majority of these reports actually came either from a health care professional or from the patient itself, filled in during a medical appointment. In this study [2], self-reported mood scores were recorded on a daily basis using a bespoke smartphone app - every day participants rated their humor on a 1 to 7 scale (1 meaning "*not at all*" and 7 meaning "*very much*") in many different categories, such as anxiety, elation, sadness, anger, irritability and energy, for a minimum of 2 months per participant, although 61 out of 130 have freely provided data for 12 more months. This is one great example of daily reporting without having the need of a regular patient's attendance on a medical appointment, in order to fill in a report, hence retrieving patient's health data.

A signature-based machine learning model was used to analyze data acquired from a clinical study, which explored daily reporting bipolar disorder carrier's mood, borderline personality disorder and healthy volunteers - this model was used to classify participants regarding their mood and then to predict their mood on the following day. A *supervised learning* approach was applied and executed by following these two tasks:

- Classification;
- Mood Prediction.

A total of 130 individuals were submitted into this machine learning approach: 48 participants were diagnosed with bipolar disorder, 31 were diagnosed with borderline personality and 51 participants were healthy.

To assess this model's classification performance they have done a *Receiver Operating Characteristic (ROC)* curve analysis by using accuracy, sensitivity, specificity and positive predictive value (PPV) as assessment metrics. The area under the ROC curve was used to assess this model's predictive quality by using different thresholds. For accuracy improvements they also applied a bootstrapping technique with an order value set to 2. The training step was performed by using each an every one of the available participants, thereupon the proportion of times period were calculated thus a participant could be classified as either belonging to the group bipolar disorder, borderline personality disorder or healthy.

The main goal was to correctly predict someone's mood on the following day rather than to predict the exact score, but even though in this article they used the following formula in order to assess the correctness of their score values: $|y - \widehat{y}| \leq 1$, where y is the correct score, assuming that it had predicted correctly, and $\widehat{y} \in \{1, ..., 6\}$. This specific signature-based model has proved to be effective when it comes to make predictions over 20 streams of observations (20 not consecutive days) regarding a participant's mood. In reference to our model's samples, it will be important to remember that we may retrieve better results when considering a small sampled-data rather than having large data samples.

However, there is one important factor that directly induces a misdiagnosis, which is widely-

known across the mental illness field, and that is heterogeneity. It is not easy to accurately diagnose someone's psychiatric disorder since there are several that evidence similar symptoms and treatments, and this is something that psychiatrists need to handle by themselves for every new patient admission. Machine learning models brought to us a whole new world that allows, not only to accurately identify someone's disorder, but also to reveal the required treatments for a certain patient, since different disease's biological origins might require a different health care routine.

In this next scientific study [48] we present a machine learning approach capable of exploring the heterogeneous nature of psychiatric disorders focusing on schizophrenia (in which both have similar symptoms, therefore are highly related).

To call what it seems an heterogeneity disorder we need to consider multiple factors that will have influence in the final verdict, thus some questions need to be answered:

- *Caucasian/Non-Caucasian*?
- *Male/Female* ?
- *With/Without comorbidities*?

These are just a few examples of the amount of factors that need to be gathered in order to corroborate our decision. A potential biomarker found in such a research must have a large effect size present in patients with matching sample characteristics. When a diagnostic is made most of them rely on a one-class or two-class separation model in order to discretize a very large group of patients into subgroups where each group stands for patients with specific features, such as, heterogeneity. The main goal relies on determining the number of clusters that optimally split these samples.

Despite heterogeneity they have also faced another relevant characteristic: homogeneity - the similarity of two or more prediction models, so basically here we can possibly find a two sampled-models in which both have the same set of features, and in such case the discriminative features list produced by each model can be compared, and the proportion of discriminative features shared by them is a qualitative indicator of homogeneity. In case of linear models they were able to define these set of features as weight vectors. In order to classify two models, as being more or less homogeneous, they used a function $f = cos(a)$ where a is an angle between two weight vectors: If $a = 0$, then we are facing two identical models, thus as a increases more we face two completely distinct models (heterogeneity).

The following techniques were used to tackle the heterogeneity problem:

- Linear separation;
- Quadratic Transformation + Linear separation;
- ANN (Artificial Neural Networks);

- HYDRA;
- Decision Tree;
- Normative Modeling.

To further map heterogeneity they compared the sub-typing and classification models against each other.

In this study they have pointed out that current diagnostic and prognostic models are not good enough for clinical applications - small-sampled models have 80% or higher accuracy in average, while large-sampled models show accuracies of about 70%. As a matter of fact, prediction models can be built to predict multiple responses considering different treatments, leading to targeted or personalized medicine.

This last point opens up a window in a matter that, clinical features are not enough for disorder's diagnostic and treatment, hence clinical and genetic features should be considered and used to perform an accurate diagnostic. In fact, when studying a bipolar disorder population, the use of clinical and genetic features throughout our study is essential as clinical factors but themselves are not enough to retrieve subgroups patterns of interest.

With regards to the genetic field most of the databases, which include people's genetic information, are either incomplete or have their info too scattered. In order to provide an enriched and improved genetic data aggregate to researchers, a new database has been built: *BDGene* - a genetic database that integrates multitype genetic factors of bipolar disorder from published genetic studies [10]. This database includes, not only the usual genetic details, but also more wider and deeper facts with reference to gene-to-gene relations, such as: SNPs, haplotype, genes, regions, and more specially, gene-to-gene interactions and pathways.

With reference to relations between these two mental disorders, *BDGene* has already enough findings which might become very handy during this thesis development, specially when analyzing genetic patterns observed in the *WTCCC* data set that require to be correlated against others people genetic information within the *BDGene* database.

This next study that we are about to present also used a genetic approach [36]. However, they focus their study in prediction tasks rather than finding patterns of interest between sub-groups of patients, as we intend to do.

Genome-Wide Association Study (GWAS) has been frequently used when it comes to perform specific tasks, such as prediction or search tasks - examples of predictions have been spotted within the health field, where they tried to predict disease risk in different individuals given their genetic sequence by using SNP's p-value rank, thus the top SNP associations were used as input for classification purposes. A typical GWAS requires a collection of genotypes from affected and healthy individuals, which researchers tried to find specific SNPs that frequently vary from one group to another.

It has been mentioned that this simple approach by itself is not very accurate and reliable, even though more studies have used the p-value as a trustful reference. Therefore, to avoid this prediction model inconvenient during this research they decided to use BootRank, which is a model that uses bootstrapping for accuracy improvement, hence a better and powerful prediction model might be achieved. Unfortunately results with better improvements were also the ones that had more heritability, possibly due to contributions from variants with low *minimum allele frequency (MAF)* - second most frequent allele value - so in these situations this model could become very unusable. Although this research has shown an overall success when it comes to predict disease risk from gene-related data, as it has been shown that BootRank was able to detect specific pathways that are associated with different diseases. As a matter of fact, these pathways are very important as they allow us to see which genes are mostly affected in order for someone to be diagnosed with bipolar disorder.

The data set used in this study consists in 3500 individuals where 2000 presented cases and 1500 were controls for 7 different diseases (Diabetes Type 1, Diabetes Type 2, Crohn's disease, Coronary Artery Disease, Bipolar Disorder, Rheumatoid Arthritis and Hypertension). Each disease data was randomly split into two separate sets: a training and a test set, by using a 5-fold cross-validation strategy. Furthermore, the minimum p-value was calculated for each SNP so bootstrapping could be applied in order to re-sample the training set, in order to produce a p-value based ranking for each sample and finally to aggregate all rankings into a final SNP ranking-based score by calculating the mean value of all p-values of a specific sample. It was verified that a smaller sample-size approach contributes for better results, as it has been referenced on previous studies mentioned in this chapter.

By using BootRank they were able to find two important enrichment pathways - representing our knowledge in reference to the molecular interaction, reaction and relation networks, namely:

- *Neuroactive ligand-receptor interaction*;
- *Propanoate metabolism.*

Once the SNP ranking was retrieved they tried to perform disease risk prediction. Researchers mentioned that this is an extremely difficult task that involves several steps, including training, modeling and testing. After selecting the top 1000 SNP's rankings a set of seven algorithms were used: *Random Forest*, *Regularized Logistic Regression*, *Support Vector Machine (SVM)*, *Naive Bayes*, *Robust Adaboost*, *Allele count* and *Log Odds. Area Under Curve (AUC)* was then used to assess the performance of such prediction model - results have shown improvements when combining multiple classification algorithms, but also have shown that for this specific research the Naive Bayes algorithm performed really well by producing an AUC value of 0.83.

Prediction tasks are always taken in consideration as almost impossible to accomplish. As a matter of fact, with regards to bipolar disorder, multiple factors contribute for this task overall success rate, such as biological pathways which radically change from one individual to another. Therefore, considering clinical and genetic features individually was never a good strategy, and

so new approaches had to be pursued in order to make progress in this disease research field and to try to figure out a way to reduce bipolar disorder main flaws, such as high misdiagnosis rate. For this reason, having an improved and accurate model built based on extracted patterns, that uses both clinical and genetic data, could be applied on first diagnostics cases that tend to be misleading.

In fact, recent studies have proven that there might be a clinical factor which could strongly help on reducing bipolar disorder cases misdiagnosis rate, and that is the human gender. For instance, [5] carried on a study which relies in genotype-by-sex direct interaction effects related to some of the most common mental illness in the modern era: schizophrenia and bipolar disorder.

They applied a method, whilst using a data set provided by *Psychiatric Genomics Consortium (PGC)*, which tries to reveal direct interactions between gender and disease risk increase or development. In fact, during their genome-wide analysis 18958 out of 65536 individuals were diagnosed as bipolar disorder patients, where 96.7% of this population are European individuals (which is extremely relevant for our research as our data set relies on patients data from the United Kingdom). Upon completion of their analysis, one specific genetic variant (rs80198067), present in ANKRD46 gene from chromosome 8, showed a strong relation between gender and genetic features. Also, ANKRD46 is a gene which encodes a protein that is highly expressed in frontal cerebral cortex, and this is an important factor to consider as gene CACNA1C, which is a very common bipolar disorder related gene, has also been associated to an odd frontal cerebral cortex morphology structure in patients with a certain genotype marker within that gene [56]. As a matter of fact, this study has shown that A-carrier group, which is the group of patients that carries genotype with risk allele A, showed significantly increased gray matter volume and reduced functional connectivity within a corticolimbic frontotemporal neural system [56]. Moreover, this last study was a genome-wide research that involved a population of 55 bipolar disorder European American patients. Once again, an European population is related to a bipolar disorder genome-wise study when having in consideration certain specific genetic features, such as genotype markers within genes.

It has been shown that there is a high possibility of cross-matching clinical and genetic data, in order to contribute to the current research status of the genetic field, which could lead to a breakthrough discovery if we prove that there is indeed a direct relation between human gender and SNPs genotype within certain genes that have been already related with the presence of bipolar disorder.

We have found one study which also uses data from WTCCC, with regards to bipolar disorder patients [38]. However, a gender-genotype relation was never revealed as it was only their intention to study SNPs combinations out of a variety of chromosome's data. Furthermore, there are only a few studies in which machine learning has been applied with regards to this mental disorder research field. For instance, in [33] a set of previously made research studies, in which machine learning techniques were used, have been reviewed and described alongside the algorithms used in those scientific studies. Also, they have concluded that despite the limitations that some studies have faced throughout their research, machine learning algorithms

can definitely be useful as they provide an abstraction of the problem in which multiple data types and sources can be simultaneously integrated within the analysis.

Based on current state of art, we have decided to carry on our research and to focus on revealing a gender-genotype relation, by using WTCCC bipolar disorder data set. Such study has never been done which we intend to take advantage of and to provide our scientific contribution whilst revealing clinical-genetic patterns of interest, and similarities between sub-groups of bipolar disorder patients by applying an unsupervised learning approach in several chromosome's data.

In the next chapter we will reveal specific details of our research that shows (1) which chromosome's genes will be involved in our research, (2) which genetic variants within genes we intend to carry on our data analysis and (3) how are we going to proceed whilst building our machine learning model.

Chapter 4

The Effect of Genetics in Bipolar Disorder

As years went by researchers needed to find a better way to identify this mental disorder in order to reduce as much as possible, not just the misdiagnosis rate, but also the total time spent per patient whenever a health professional tries to find the right treatment, hence patients could experience a significantly improved quality of life and treatment costs could be reduced. Having in consideration the estimated amount of people that have been diagnosed with this mental disorder, it has become even more relevant to the scientific research field what can be done to improve the whole health system, hence all these controversial factors might be eliminated. Clinical factors on their own are not enough to assess someone's health condition, as they can become false advertisers in a manner that people with similar clinical factors can experience completely different mental disorder symptoms, and by aiming the genetic approach of this disease we've ended up with the following thesis' theme: *Extraction of a Bipolar Disorder associated genetic pattern.*

There is an important human characteristic to be pointed out that has been proved to be highly related with mood disorders, and that is the human gender. There are several reasons that lead us to study this disorder's prevalence regarding the human gender: it could easily improve patients list management by giving priority to one of the genders.

An electronic systematic research has revealed that in cases of unipolar disorder depressive episodes are twice more frequently in women rather than in men [15]. Despite this fact, in cases of bipolar disorder it has been reported that there is no such association between this disorder and the human gender. In fact, further statistic results have shown that there is a quiet similar gender ratio regarding the prevalence of bipolar disorder, however facts point out to an increased risk of having mixed and rapid cycling episodes in women that have been already diagnosed with bipolar disorder type II. Regardless the gender it is extremely important to define, as better as we can, which human characteristics are mostly relevant for the study of bipolar disorder, such as age of onset - Regarding axis I, which is the axis that this disorder relies on, it is frequent to

become visible in adults between 20 and 30 years old [47] [15].

We are emphasizing here the unipolar disorder despite this thesis' theme mainly because when it comes to diagnosing, unfortunately many patients are misdiagnosed as having unipolar disorder, hence a bad treatment will be applied to that person in the first place - wrong medications and wrong therapy, or the lack of them. It is undeniable that bipolar disorder is one of the most common mental disorders, as mentioned before, but when it comes to treatment there is no such thing as a reliable plan ready to be triggered for each individual neither for each health institution that is admiting newly arrived patients, such as hospitals and health care institutions. Thereupon, there is an estimated delay of 5 to 10 years in average induced by a misleading diagnosis trial[26].

Researchers around the globe have already identified several genes that present direct and non-direct associations with bipolar disorder in which these are spread out across the human genome. Some of these associations reveal a promising interaction between specific genetic variants and human gender [32], specifically females. Having in consideration a slightly possible interaction between genetic variants within genes and human gender has caught our attention, hence we have started to focus on this subject in order to scientifically contribute to this discovery that could be quiet beneficial to the current world health system, with regards to bipolar disorder misdiagnosis rate. Indeed, it is a breakthrough discovery as throughout the years bipolar disorder has never been directly distinguishable by the human gender and other clinical factors, which turned out to be very important to consider both genetic and clinical factors, as together they can provide an enhanced medical diagnostic.

During this thesis' development we will be working with a dataset owned by the *Wellcome Trust Case Control Consortium*: *EGAD00000000003* (1998 samples of bipolar disorder cases) [8]. We've decided to work with case samples only, as other machine learning related studies within the psychiatric field have revealed more captivating results when applying unsupervised learning within a group in which individuals present similar characteristics [9], rather than comparing case samples and controls. Indeed, we have 1998 samples of bipolar disorder patients in which we intend to find particular subgroups that will support our research work.

We will build a machine learning model focusing on two essential goals: (1) extend other works in the literature by building other machine learning models for this data, and (2) to learn patterns that can be used for personalized health care which include a misdiagnosis rate reduce.

We've selected multiple chromosomes which contain genetic variants within certain genes that have been associated to bipolar disorder in previous and recent scientific researches. Some of these genetic variants are now presented along with their belonging genes:

- DISC1
  - It is a gene from chromosome 1 in which the following SNPs have been associated to the presence of bipolar disorder: rs203368, rs435136 [35];

- ARPP21
  - It is a gene from chromosome 3 in which the following SNPs have been associated to the presence of bipolar disorder: rs1523041 [49];
- GABRB1
  - It is a gene from chromosome 4 in which the following SNPs have been associated to the presence of bipolar disorder: rs7680321 [45];
- ANKRD46
  - It is a gene from chromosome 8 in which the following SNPs have been associated to the presence of bipolar disorder: rs203368, rs435136 [5];
- ANK3
  - It is a gene from chromosome 10 in which the following SNPs have been associated to the presence of bipolar disorder: rs10994336, rs9804190 [27] [17];
- CACNA1C
  - It is a gene from chromosome 12 in which the following SNPs have been associated to the presence of bipolar disorder: rs1006737, rs4765914, rs4765913 and rs2239063 [27] [51] [17];
- DUSP6
  - It is a gene from chromosome 12 in which the following SNPs have been associated to the presence of bipolar disorder: rs769700, rs704076, rs770087, rs808820 and rs2279574 [28];
- GRIN2B
  - It is a gene from chromosome 12 in which the following SNPs have been associated to the presence of bipolar disorder: rs1805502 and rs1805247 [60] [40] [31] [34];
- SYN3
  - It is a gene from chromosome 22 in which the following SNPs have been associated to the presence of bipolar disorder: rs9621532 [39].

Even though only a few genes were mentioned we loaded all genetic variants' information available regarding each chromosome that has been mentioned before, as more than just analyzing each set of genotypes within a certain gene, we want to make sure that we don't restrict our model to only a portion of each chromosome genetic data. Thus, we loaded genetic variants details of chromosomes 1, 3, 4, 8, 10, 12 and 22, which together generate a massive data set to be analyzed, hence cluster analysis can be performed simultaneously for all these genetic variants.

Our study relies on one specific type of machine learning: Unsupervised Learning.

We've chosen this method since it is the most appropriate for our study, as (1) our data set does not contain any target variable and (2) our main goal is to find patterns of interest within subgroups of patients that can support our theory - *Is there a plausible relation between gender and genetic variants regarding presence of bipolar disorder?*

We applied one of the most used algorithms in cluster analysis: the *k-means* clustering algorithm. Our data set was split up into two parts, in which $\frac{1}{3}$ of our data set samples was used to assess the optimal k clusters value, and we have then applied the k-means algorithm on the remaining $\frac{2}{3}$ of our data set. Even though this is an unsupervised learning method we've applied the train and test approach based on the optimal k clusters value, as we want to make sure that our machine learning model is as reliable and accurate as possible when applied to unseen data. In order to assess the quality of the clusters generated and choose the best k, we used two metrics: *elbow* and *silhouette*.

We will execute the k-means algorithm on our training data in which we will calculate the total *within-cluster sum of square (WSS)* for an integer k between one and ten:

$$\{k \in \Re \mid 1 \leq k \leq 10\}$$

Once all results are produced we shall plot the curve of WSS and look for a bend (knee) in this same plot, as represented in Figure 4.1, where usually it is considered the most appropriate (optimal) k value to be used throughout the clustering analysis. For instance, when looking to Figure 4.1 we would select $k = 2$ as our optimal k value.

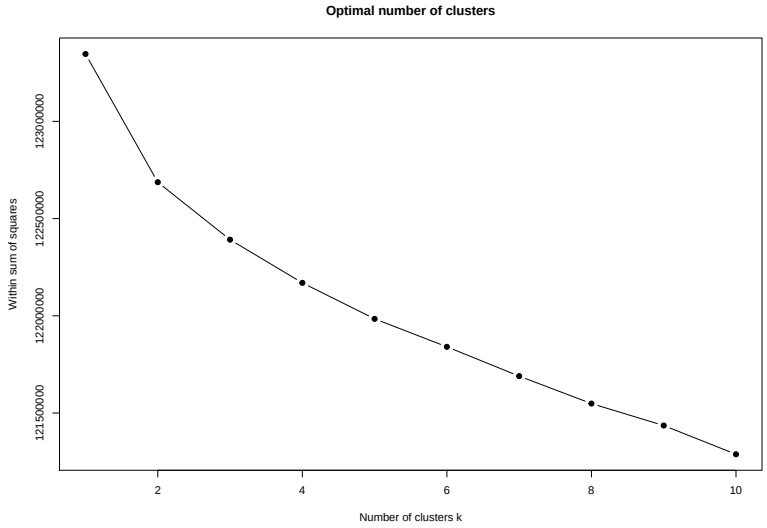


Figure 4.1: Machine Learning - Elbow method for optimal k assessment.

In order to assess Elbow's method results we shall use the Silhouette method as it will

measure the quality of our cluster by calculating all average silhouette coefficients for each k within the interval:

$$\{k \in \Re \mid 1 < k \leq 10\}$$

When applying these two methods on our discrete dataset we used the Euclidean distance as a distance metric parameter in which q and p are Cartesian coordinates of two points within the Euclidean space:

$$dist(p, q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$$

Upon assessment task completion we applied the k-means algorithm to the remaining $\frac{2}{3}$ of our data set, as previously mentioned, and we want to make sure that k-means always separate our data samples similarly by assigning a certain sample X_i to the same cluster whenever we run it, in which X represents our test data and i a specific sample. Hence, our cluster's information will remain identical throughout the local analysis task.

An analysis needs to be performed regarding each mentioned gene that has been associated to bipolar disorder interactions in previous studies - *Interpretation* task. We analyzed each and every single-nucleotide polymorphism (SNP's) genotype so we could than correlate genetic data and clinical data, such as age, region and gender, in order to extract as many patterns as we can which support the case where bipolar disorder can in fact be distinguishable taking into account a multi-factor scenario where specific genetic and clinical details are present in female individuals but not in male individuals. A genotype-by-sex interaction has been mentioned in other GWAS [5] but such analysis was never extended to other chromosomes and simultaneously when building a machine learning model that tries to correlate clinical and genetic data. A recent study examined combinations of SNPs [38] by using the exact same data set that we used from WTCCC, however it was never their goal to find any correlation between gender and bipolar disorder.

The process of extracting particular patterns (considering both genetic and clinical information) from our data samples may potentially provide an improvement to the current health care system: such implementation could potentially be applied directly on health institutions, by successfully performing a first medical diagnosis, regarding that someone's genetic and clinical data would positively correlate our model, hence it would become feasible to reduce the misdiagnosis rate, that has been shown to be too high during the last decades [53], and also the annual costs that countries have been dealing whilst maintaining their own health care system. Until now a patient would only have a proper diagnostic after a clear-cut episode of either mania or hypomania [20] and our machine learning approach would be able to avoid that scenario by assessing someone's genetic and clinical data before an episode has taken place.

Chapter 5

Methodology and Experiments

5.1 Environment Setup

Throughout this project development we have used R and Bash as our programming languages of choice. The reason we have chosen R relies on the fact that this programming language provides almost all the required tools/packages when it comes to data analysis and data manipulation: there are several packages available online which implements almost all features required for this project completion and these can easily be used for data manipulation when using specific data structures, such as data frames. As an IDE we have used RStudio which is the most indicated to be used regarding R projects development.

In order to fulfill this research goals we have used a powerful machine which is capable of handling Big Data analysis tasks such as ours. This machine's specs are listed as it follows:

- Operating System: *Fedora 20 3.19.8-100.fc20.x86_64*
- CPU: *4 × 6 core AMD Opteron 2400 MHz*
- RAM: *64 GigaBytes*
- SWAP: *66 GigaBytes*
- Disk Space Required: *40 GigaBytes*

Software specifications regarding this project development are listed as it follows:

- IDE: *RStudio Server Version 1.1.456*
- Programming Languages:
  - *R Version 3.2.0*
  - *Bash Version 4.2.53*

- R packages:
  - readr
  - ggplot2
  - dplyr
  - plyr
  - reshape2
  - RColorBrewer
  - foreach
  - doParallel
  - NbClust
  - cluster
  - beepr

Throughout this project development we have accessed this machine remotetly by using a web browser. As this machine had rstudio-server service running on its localhost we have established a port forwarding SSH tunnel in order to access this remote host's service (figure 5.2) from our local host as depicted in Figure 5.1:

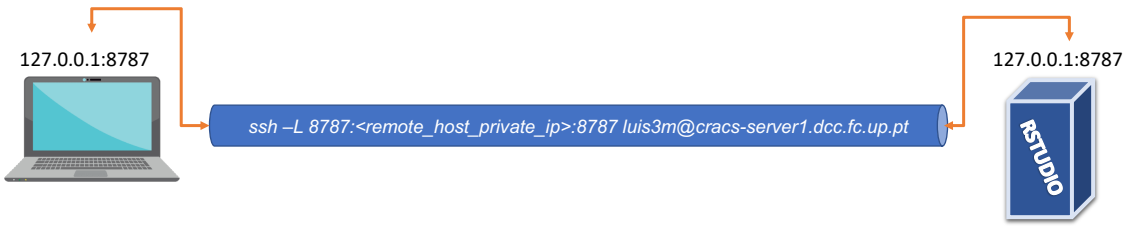


Figure 5.1: SSH tunnel between local and remote host.

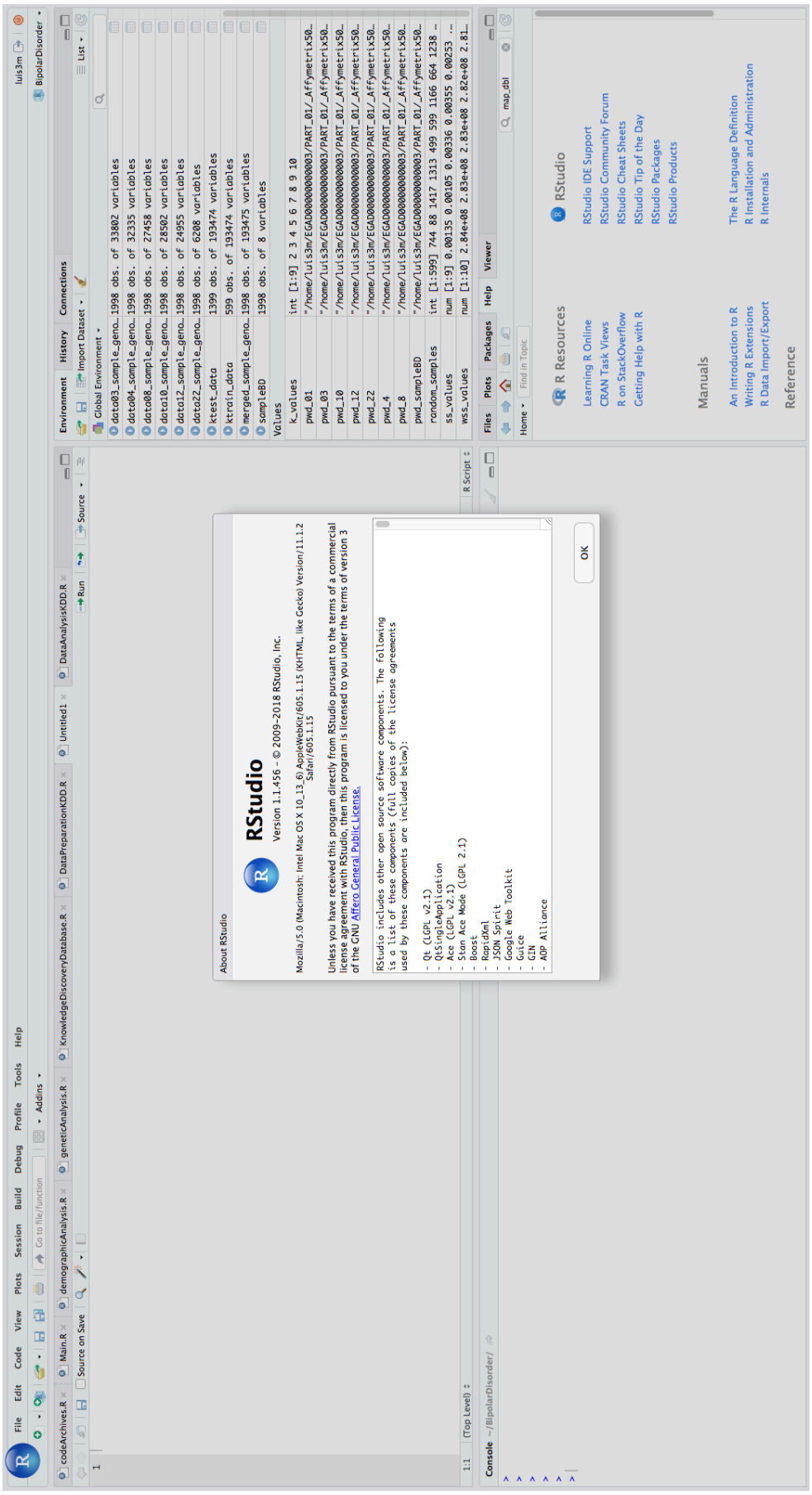


Figure 5.2: RStudio server.

5.2 Data Description

In order to perform our research we have used data of bipolar disorder cases which was disclosed by the Welcome Trust Case Control Consortium (WTCCC) after signing an agreement between the Consortium and our Department of Computer Science. The consortium aims for global disease treatments improvement by providing genetic data which will help researchers all around the globe to understand global diseases major causes that might be influenced by genetic factors. They have gathered genetic variants data from up to 500000 sites which have been categorized by their different phenotypes as described in Table 5.1.

Disease	**Co-Principal Applicants**	**Cohort Abbreviation**
Disease cohorts		
Type 1 diabetes	John Todd & David Clayton	T1D
Type 2 diabetes	Mark McCarthy & Andrew Hattersley	T2D
Crohn's disease	Miles Parkes & Chris Mathew	CD
Breast cancer	Michael Stratton & Nanzeen Rahmad	BC
Coronary heart disease	Alistair Hall & Nilesh Samani	CHD
Hypertension	Mark Caulfield & Martin Farrall	HT
Bipolar disorder	**Nick Craddock**	**BD**
Rheumatoid arthritis	Jane Worthington	RA
Multiple sclerosis	Alastair Compston	MS
Ankylosing spondylitis	Matthew Brown	AS
Autoimmune thyroid disease	Stephen Gough	ATD
Malaria	Dominic Kwiatkowski	ML
Tuberculosis	Adrian Hill, Melanie Newport & Giorgio Sirugo	TB
Control cohorts		
1958 Birth Cohort	Marcus Pembrey, David Strachan & Peter Shepherd	58C
UK Blood Service	Willem Ouwehand	UKBS

Table 5.1: WTCCC disease samples [8].

Our research relies in a specific WTCCC data set, as shown in Table 5.2, which uses the following technology:

- Affymetrix 500K SNP chip;

The Affymetrix 500K SNP chip yields approximately 4 *GigaBytes* per cohort, therefore all genotype data provided have been split according to chromosome and sorted according to SNP position. These data files used throughout the project are now presented in Table 5.3 which include both genetic variants genotype data and cohort's clinical data.

Access to quantile normalized signal data has been given, however we have not used it throughout this project development.

Dataset Accessionsort descending	Technology	Type	Samples	Description
EGAD00000000003	Affymetrix 500K	cases	1998	WTCCC1 project Bipolar Disorder (BD) samples

Table 5.2: WTCCC bipolar disorder cases data[8].

Technology	Data Files	Size
Affymetrix 500K	_Affymetrix500K_Chiamo_20070205fs1_Affx_20070205fs1_gt_BD_Chiamo_01.txt	636 MB
	_Affymetrix500K_Chiamo_20070205fs1_Affx_20070205fs1_gt_BD_Chiamo_02.txt	16 MB
	_Affymetrix500K_Chiamo_20070205fs1_Affx_20070205fs1_gt_BD_Chiamo_03.txt	620 MB
	_Affymetrix500K_Chiamo_20070205fs1_Affx_20070205fs1_gt_BD_Chiamo_04.txt	372 MB
	_Affymetrix500K_Chiamo_20070205fs1_Affx_20070205fs1_gt_BD_Chiamo_05.txt	428 MB
	_Affymetrix500K_Chiamo_20070205fs1_Affx_20070205fs1_gt_BD_Chiamo_06.txt	744 MB
	_Affymetrix500K_Chiamo_20070205fs1_Affx_20070205fs1_gt_BD_Chiamo_08.txt	387 MB
	_Affymetrix500K_Chiamo_20070205fs1_Affx_20070205fs1_gt_BD_Chiamo_09.txt	681 MB
	_Affymetrix500K_Chiamo_20070205fs1_Affx_20070205fs1_gt_BD_Chiamo_10.txt	925 MB
	_Affymetrix500K_Chiamo_20070205fs1_Affx_20070205fs1_gt_BD_Chiamo_11.txt	869 MB
	_Affymetrix500K_Chiamo_20070205fs1_Affx_20070205fs1_gt_BD_Chiamo_12.txt	953 MB
	_Affymetrix500K_Chiamo_20070205fs1_Affx_20070205fs1_gt_BD_Chiamo_13.txt	1.2 GB
	_Affymetrix500K_Chiamo_20070205fs1_Affx_20070205fs1_gt_BD_Chiamo_14.txt	1.5 GB
	_Affymetrix500K_Chiamo_20070205fs1_Affx_20070205fs1_gt_BD_Chiamo_15.txt	1.6 GB
	_Affymetrix500K_Chiamo_20070205fs1_Affx_20070205fs1_gt_BD_Chiamo_16.txt	1.7 GB
	_Affymetrix500K_Chiamo_20070205fs1_Affx_20070205fs1_gt_BD_Chiamo_17.txt	1.4 GB
	_Affymetrix500K_Chiamo_20070205fs1_Affx_20070205fs1_gt_BD_Chiamo_19.txt	1.7 GB
	_Affymetrix500K_Chiamo_20070205fs1_Affx_20070205fs1_gt_BD_Chiamo_20.txt	1.9 GB
	_Affymetrix500K_Chiamo_20070205fs1_Affx_20070205fs1_gt_BD_Chiamo_21.txt	1.9 GB
	_Affymetrix500K_Chiamo_20070205fs1_Affx_20070205fs1_gt_BD_Chiamo_22.txt	2.0 GB
	_Affymetrix500K_Chiamo_20070205fs1_Affx_20070205fs1_gt_BD_Chiamo_23.txt	2.0 GB
	_Affymetrix500K_Chiamo_20070205fs1_Affx_20070205fs1_gt_BD_Chiamo_24.txt	2.5 GB
	_Affymetrix500K_Chiamo_20070205fs1_Affx_20070205fs1_gt_BD_Chiamo_X.txt	2.4 GB
	_Affymetrix500K_Chiamo_20070205fs1_Affx_20070205fs1_sample_BD.txt	128 KB

Table 5.3: WTCCC bipolar disorder cases genetic and clinical data files.

All SNP's genotype data files have an integer or character as a suffix, which indicates the chromosome the file is associated with. Therefore, we present a single genetic data file for each homologous chromosome, in which X is related to the sexual chromosome genetic data. Furthermore, these data follow a specific text format as presented in Table 5.4 and table 5.5 respectively for genetic and clinical data.

With regards to genetic data, all mentioned files provide specific information, such as *SNP* number, *Sample* number associated to one unique subject, *Genotype* and Score in which the last one is related to a score value generated by the *CHIAMO* algorithm [8] during the genotype calling task. As a matter of fact, only *CHIAMO* generated data with a score value greater than 0.9 should be considered whilst analyzing this data set [8]. On the other hand, WTCCC has provide us some interesting clinical data regarding our samples, such as *Gender* which is set as 1 for males and 2 for females, *Region* and *Age of Recruitment* which is a single integer that defines a unique age interval, as an example, 3 defines subjects' age range between 30 and 39 years old. Unfortunately for us *Age of Onset* has been provided as *Unknown* regarding all individuals that

belong to the BD cohort.

SNP	SAMPLE	GENOTYPE	SCORE
rs10488368	*WTCCC65841*	*AG*	*1*
rs10488368	*WTCCC65569*	*AA*	*1*
rs10488368	*WTCCC65777*	*AA*	*1*
rs10488368	*WTCCC65795*	*AA*	*1*
rs10488368	*WTCCC65810*	*AA*	*1*
rs10488368	*WTCCC65823*	*AG*	*1*
rs10488368	*WTCCC65845*	*AG*	*1*
rs10488368	*WTCCC65571*	*AA*	*1*
rs10488368	*WTCCC65779*	*AA*	*1*
rs10488368	*WTCCC65797*	*AA*	*1*

Table 5.4: WTCCC genetic data description.

SAMPLE	GENDER	COHORT	SUPPLIER	PLATE	REGION	AGE_RECRUITMENT	AGE_ONSET
WTCCC65841	2	BD	PMHWW	11142A1	Southwestern	3	Unknown
WTCCC65569	2	BD	PMHWW	11142A2	Southwestern	5	Unknown
WTCCC65777	2	BD	PMHWW	11142A3	Northwestern	6	Unknown
WTCCC65795	1	BD	PMHWW	11142A4	Southern	3	Unknown
WTCCC65810	1	BD	PMHWW	11142A5	Wales	7	Unknown

Table 5.5: WTCCC clinical data description.

In order to better understand our population of bipolar disorder cases we have carried on a demographic analysis by using these individuals clinical data, as previously shown in Table 5.5.

Such analysis has been applied on top of the clinical data provided by WTCCC (*_Affymetrix500K_Chiamo_20070205fs1_Affx_20070205fs1_sample_BD.txt*) in regards to bipolar disorder cases of the *1958 British Birth Cohort* and it has been summarized in Table 5.6. We also present a chart for each clinical factor, *Gender*, *Age* and *Region* depicted in Figures 5.3, 5.4 and 5.5 respectively.

	Bipolar Disorder
Population	*1998*
1958 British Birth Cohort	
Gender	
Male	*37,59%*
Female	*62,41%*
Age	
10-19	*0,70%*
20-29	*9,76%*
30-39	*20,02%*
40-49	*28,78%*
50-59	*24,03%*
60-69	*13,86%*
70-79	*2,50%*
80-89	*0,35%*
Region	
East+West Ridings	*1,30%*
Eastern	*3,15%*
London	*6,66%*
Midlands	*23,77%*
North Midlands	*5,91%*
Northern	*8,81%*
Northwestern	*3,40%*
Scotland	*9,96%*
Southeastern	*4,75%*
Southern	*5,81%*
Southwestern	*5,76%*
Wales	*20,72%*

Table 5.6: WTCCC data demographic analysis.

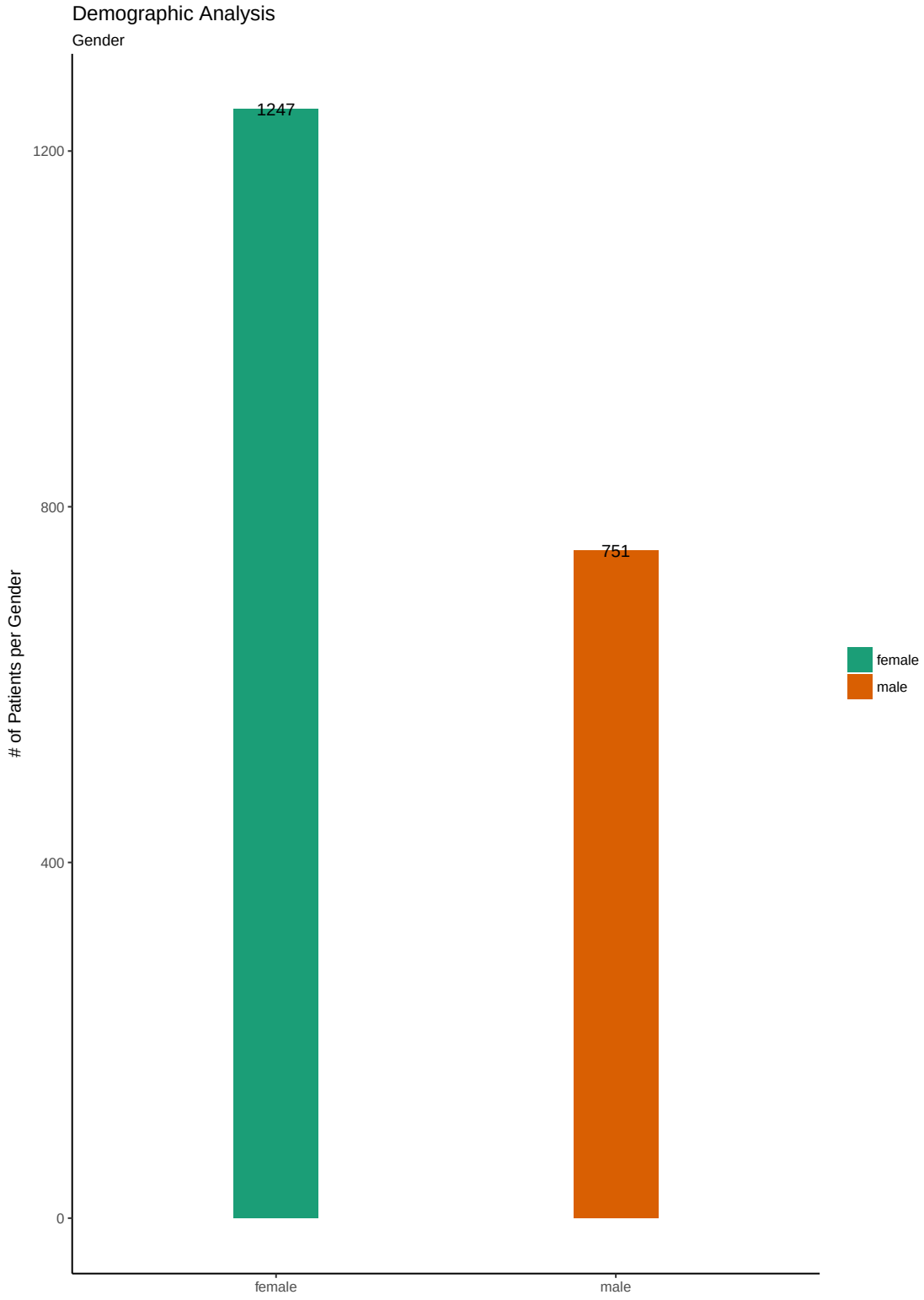


Figure 5.3: WTCCC demographic analysis - Gender.

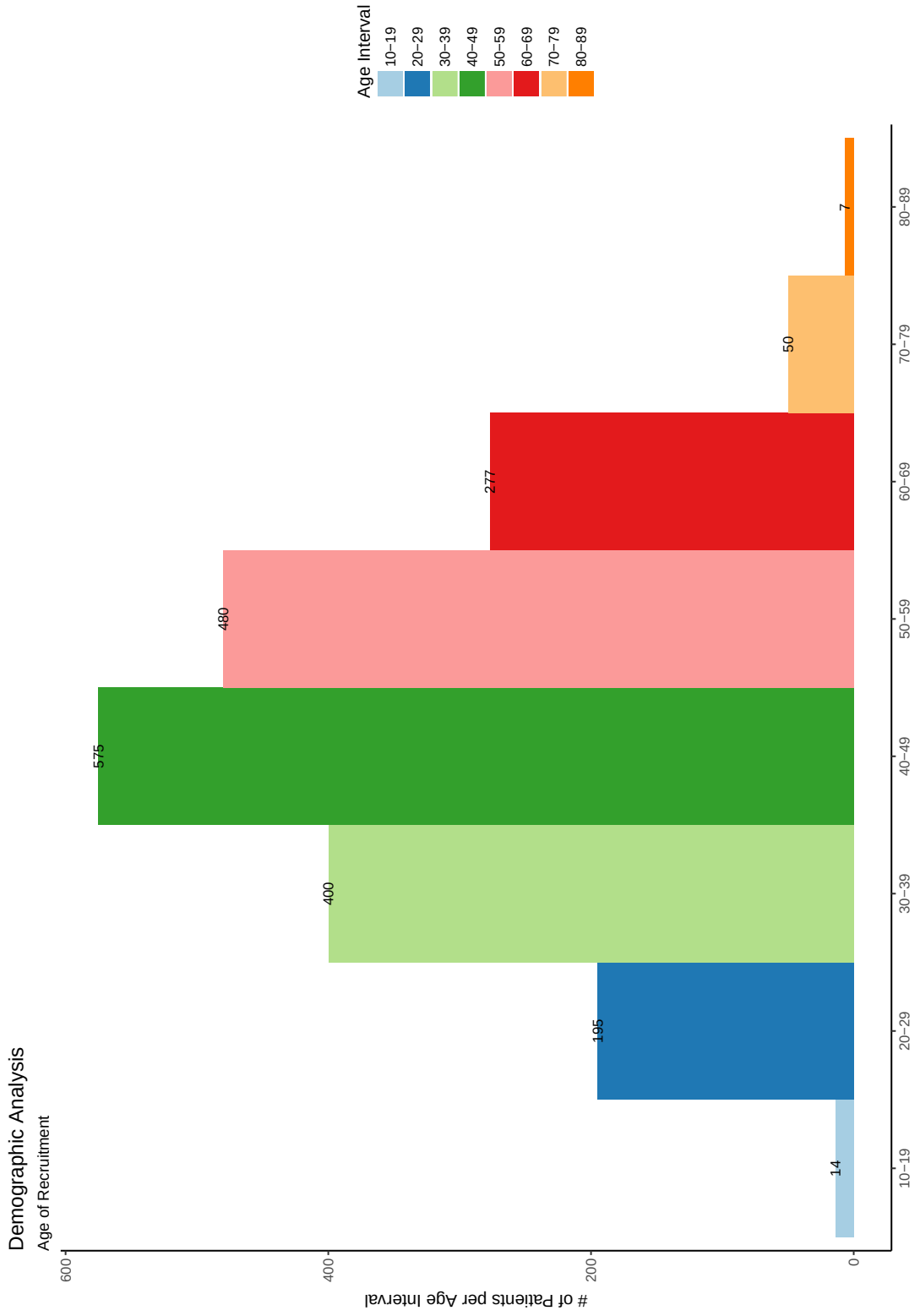


Figure 5.4: WTCCC demographic analysis - Age.

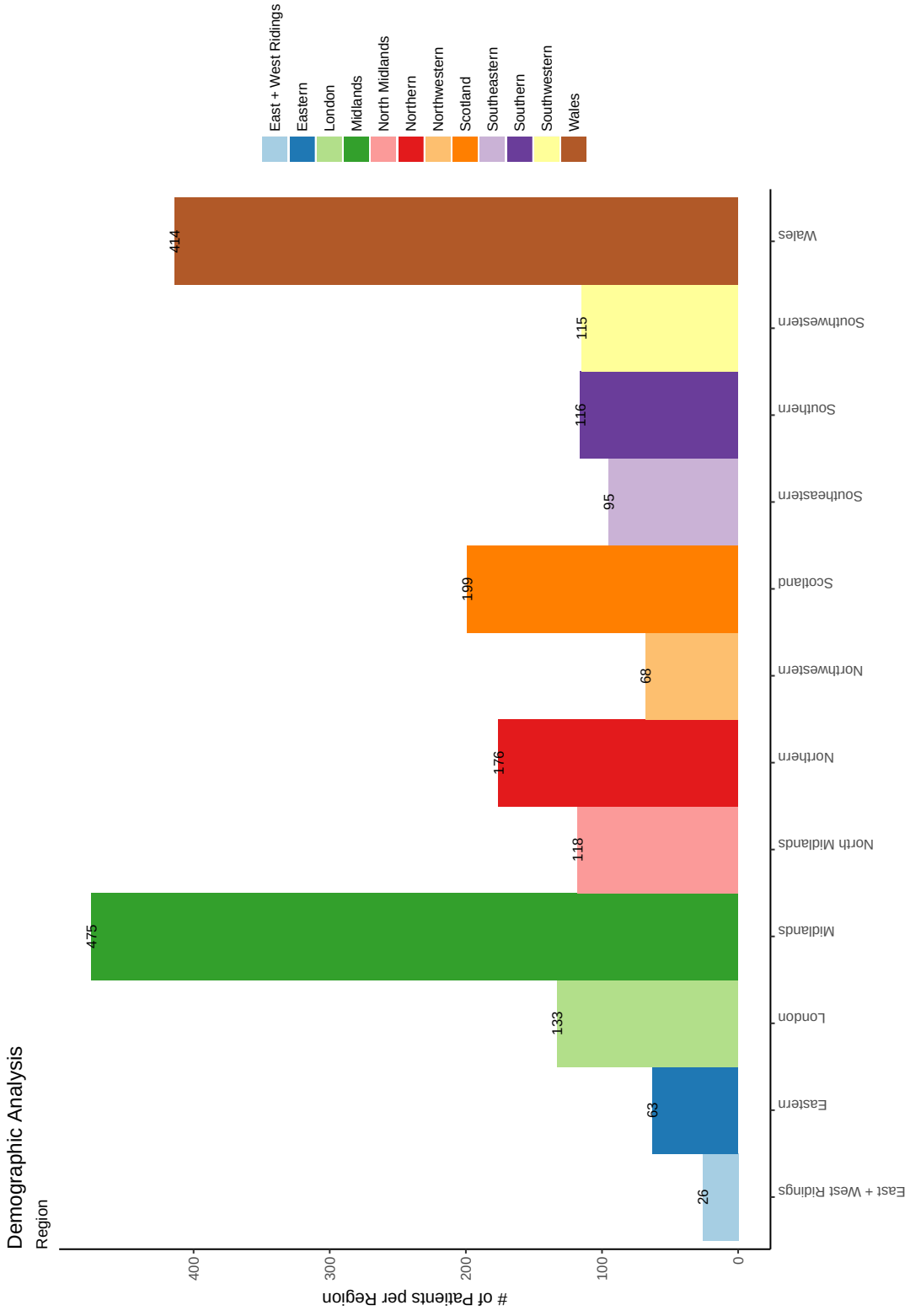


Figure 5.5: WTCCC demographic analysis - Region.

5.3 Data Preparation

As it has been mentioned in the previous chapter we have performed a clustering analysis based on genetic data from chromosomes 1, 3, 4, 8, 10, 12 and 22, in which their associated data files are now presented accordingly:

- *_Affymetrix500K_Chiamo_20070205fs1_Affx_20070205fs1_gt_BD_Chiamo_01.txt*;
- *_Affymetrix500K_Chiamo_20070205fs1_Affx_20070205fs1_gt_BD_Chiamo_03.txt*;
- *_Affymetrix500K_Chiamo_20070205fs1_Affx_20070205fs1_gt_BD_Chiamo_04.txt*;
- *_Affymetrix500K_Chiamo_20070205fs1_Affx_20070205fs1_gt_BD_Chiamo_08.txt*;
- *_Affymetrix500K_Chiamo_20070205fs1_Affx_20070205fs1_gt_BD_Chiamo_10.txt*;
- *_Affymetrix500K_Chiamo_20070205fs1_Affx_20070205fs1_gt_BD_Chiamo_12.txt*;
- *_Affymetrix500K_Chiamo_20070205fs1_Affx_20070205fs1_gt_BD_Chiamo_22.txt*.

Each and every of these files were loaded into a very common R data structure: *data frame*. These kinds of structures are loaded into memory allowing us to easily and quickly access our data, rather than having to read it from disk all the time and to create an IO performance bottleneck. Figure 5.6 shows how these chromosome data were loaded into RStudio and labeled accordingly.

```
# Chromossome 1 Data import
pwd_01 <- "/home/luis3m/EGAD00000000003/PART_01/_Affymetrix500K_Chiamo_20070205fs1_Affx_20070205fs1_gt_BD_Chiamo_01.txt"
if(!exists("data_01")) { data_01 <- read_tsv(file = pwd_01, col_names = F, progress = T, col_types = cols()) }
colnames(data_01) <- c("SNP", "SAMPLE", "GENOTYPE", "SCORE")

# Chromossome 3 Data import
pwd_03 <- "/home/luis3m/EGAD00000000003/PART_01/_Affymetrix500K_Chiamo_20070205fs1_Affx_20070205fs1_gt_BD_Chiamo_03.txt"
if(!exists("data_03")) { data_03 <- read_tsv(file = pwd_03, col_names = F, progress = T, col_types = cols()) }
colnames(data_03) <- c("SNP", "SAMPLE", "GENOTYPE", "SCORE")

# Chromossome 4 Data import
pwd_4 <- "/home/luis3m/EGAD00000000003/PART_01/_Affymetrix500K_Chiamo_20070205fs1_Affx_20070205fs1_gt_BD_Chiamo_04.txt"
if(!exists("data_04")) { data_04 <- read_tsv(file = pwd_4, col_names = F, progress = T, col_types = cols()) }
colnames(data_04) <- c("SNP", "SAMPLE", "GENOTYPE", "SCORE")

# Chromossome 8 Data import
pwd_8 <- "/home/luis3m/EGAD00000000003/PART_01/_Affymetrix500K_Chiamo_20070205fs1_Affx_20070205fs1_gt_BD_Chiamo_08.txt"
if(!exists("data_08")) { data_08 <- read_tsv(file = pwd_8, col_names = F, progress = T, col_types = cols()) }
colnames(data_08) <- c("SNP", "SAMPLE", "GENOTYPE", "SCORE")

# Chromossome 10 Data import
pwd_10 <- "/home/luis3m/EGAD00000000003/PART_01/_Affymetrix500K_Chiamo_20070205fs1_Affx_20070205fs1_gt_BD_Chiamo_10.txt"
if(!exists("data_10")) { data_10 <- read_tsv(file = pwd_10, col_names = F, progress = T, col_types = cols()) }
colnames(data_10) <- c("SNP", "SAMPLE", "GENOTYPE", "SCORE")

# Chromossome 12 Data import
pwd_12 <- "/home/luis3m/EGAD00000000003/PART_01/_Affymetrix500K_Chiamo_20070205fs1_Affx_20070205fs1_gt_BD_Chiamo_12.txt"
if(!exists("data_12")) { data_12 <- read_tsv(file = pwd_12, col_names = F, progress = T, col_types = cols()) }
colnames(data_12) <- c("SNP", "SAMPLE", "GENOTYPE", "SCORE")

# Chromossome 22 Data import
pwd_22 <- "/home/luis3m/EGAD00000000003/PART_01/_Affymetrix500K_Chiamo_20070205fs1_Affx_20070205fs1_gt_BD_Chiamo_22.txt"
if(!exists("data_22")) { data_22 <- read_tsv(file = pwd_22, col_names = F, progress = T, col_types = cols()) }
colnames(data_22) <- c("SNP", "SAMPLE", "GENOTYPE", "SCORE")
```

Figure 5.6: RStudio chromosomes data import.

Since we are interested in revealing potential relations, between genetic and clinical data, our first goal was to retrieve and organize each chromosome genetic data in a manner that we can have one data frame in RStudio per chromosome that contains all SNP's genotypes of each individual.

As these are massive files with thousands of SNPs per chromosome a trivial loop would certainly fail our performance tests, as more than accomplish our goals we intend to achieve an algorithm that is fast enough whilst mining our data, otherwise such big data analysis would take days to accomplish.

Having said that, we decided to make use of the *doParallel* and *foreach* R packages by implementing a parallel processing algorithm in which several slave threads are created out of the main R session process based on the number of CPU cores available for this task completion. As an example, we present in Figure 5.7 how we have retrieved all genotypes associated to each sample from chromosome 1. However, we also present in Appendix A.1, A.2, A.3, A.4, A.5 and A.6 the R code for the remaining chromosomes data loading process.

```
#####################################
# Chromossome: 1                    #
# Total SNPs: 40220                 #
#####################################

# Get Genotype per sample
if (!exists("data01_genotype_per_sample_parallel")) {
  # Set parallel cluster
  parallel_cluster <- makeCluster(detectCores()-1)
  registerDoParallel(parallel_cluster)
  data01_genotype_per_sample_parallel <-
    foreach (item = 1:length(samples), .verbose = T) %dopar% {
      data_01[data_01$SAMPLE == samples[item], "GENOTYPE"]
    }
  # Stop parallel cluster
  stopCluster(parallel_cluster)
  remove(parallel_cluster)
}

# Genotype per sample merged Dataframe
if(!exists("data01_sample_genotype_df")) {
  aux_ouput_df_pwd <- "/home/luis3m/DATA/data01_sample_genotype_df.txt"
  if(file.exists(aux_ouput_df_pwd)) {
    data01_sample_genotype_df <- read_tsv(file = aux_ouput_df_pwd, col_names = F, progress = T, col_types = cols())
    colnames(data01_sample_genotype_df) <- c("SAMPLES",unlist(unique(data_01["SNP"]), use.names = F))
    remove(aux_ouput_df_pwd)
  } else {
    cat("", file = aux_ouput_df_pwd)
    for (index in 1:length(data01_genotype_per_sample_parallel)) {
      aux_ouput_df <- data.frame(t(matrix(unlist(data01_genotype_per_sample_parallel[[index]]))))
      aux_ouput_df <- cbind(samples[index], aux_ouput_df)
      write_tsv(aux_ouput_df, path = aux_ouput_df_pwd, append = T, col_names = F)
    }
    # RUN categoricalTodiscrete.sh SCRIPT!
    beep(sound = 6)
    cat("DATA 01 - RUN categoricalTodiscrete.sh SCRIPT!\n")
    cat("Sleeping...\n")
    Sys.sleep(time = 297)
    beep(sound = 3)
    Sys.sleep(time = 3)
    cat("I'm back!\n")
    data01_sample_genotype_df <- read_tsv(file = aux_ouput_df_pwd, col_names = F, progress = T, col_types = cols())
    colnames(data01_sample_genotype_df) <- c("SAMPLES",unlist(unique(data_01["SNP"]), use.names = F))
    remove(aux_ouput_df)
    remove(aux_ouput_df_pwd)
  }
}
```

Figure 5.7: Retrieve SNP's genotypes per sample with parallel processing.

As we can see in Figure 5.7 there is one step in which this code will ask the user to run the *categoricalTodiscrete.sh* script which is a Bash script. We present in Appendix A.7 part of that script which treats chromosome 1 genotype data.

We decided to use Bash, as an auxiliary programming language, due to algorithm's performance concerns raised whilst using only RStudio and processing all these data frames using in memory storage. These data frames are computationally expensive whilst performing certain operations, such as converting every single qualitative value (e.g. "AA") to a quantitative value (e.g. 1), in a manner that it would slow down our algorithm's overall performance. Hence, we wrote every single data frame to an associated file and we used Bash to perform a simple and accurate replace, so we can retrieve the following list of files which, despite each sample ID, only contains discrete data:

- *data01_sample_genotype_df.txt*;
- *data03_sample_genotype_df.txt*;
- *data04_sample_genotype_df.txt*;
- *data08_sample_genotype_df.txt*;
- *data10_sample_genotype_df.txt*;
- *data12_sample_genotype_df.txt*;
- *data22_sample_genotype_df.txt.*

Every single genotype has been converted into an associated integer as it follows:

- "AA" $\rightarrow$ 1;
- "AC" $\rightarrow$ 2;
- "AG" $\rightarrow$ 3;
- "AT" $\rightarrow$ 4;
- "CC" $\rightarrow$ 5;
- "CG" $\rightarrow$ 6;
- "CT" $\rightarrow$ 7;
- "GG" $\rightarrow$ 8;
- "GT" $\rightarrow$ 9;
- "TT" $\rightarrow$ 10.

Upon completion of this task, we just need to read every single modified file and load them into new data frames, rather than operating on top of the previous ones, and we also label them accordingly.

Moving on, we now need to merge all our data files into one single genotype file that is going to be used during the clustering analysis. Once again, instead of using R to combine several data frames we decided to use Bash as we always aim for an overall algorithm performance improvement. In order to merge multiple data files we have used the *mergeData.sh* script presented in Appendix A.8 which retrieves one single massive data file entitled *merged_sample_genotype_df.txt*. As each and every of these data files had sample's IDs as their first column, the output of merging them will contain $n \times number\ of\ chromosomes$ sample's IDs columns which are nothing more than duplicated data on our merged data file, and we only intend to keep the first column with sample's IDs. Hence, once this massive data file has been loaded into RStudio we have carried on

with a clean up task, that removes duplicated data out of merged data file and we have labeled it accordingly: R code associated to these tasks is now presented in Figures 5.8, 5.9 and 5.10 respectively.

```
#########################################
# Chromossome: 1,3,4,8,10,12,22         #
# Total Samples: 1998                   #
# Total SNPs: 193474                    #
# Total Genotypes: 386,561,052          #
#########################################

# RUN mergeData.sh SCRIPT!
beep(sound = 6)
cat("MERGE DATA - RUN mergeData.sh SCRIPT!\n")
cat("Sleeping...\n")
Sys.sleep(time = 600)
ctr=0
while (!file.exists("/home/luis3m/DATA/merged_sample_genotype_df.txt")) {
  if(ctr == 4) {
    cat("Something is wrong. Please check your script!\n")
    stop()
    cat("File not found\n")
  }
  ctr=ctr+1
  cat("Wainting for it...\n")
  Sys.sleep(time = 30)
}
remove(ctr)
beep(sound = 3)
Sys.sleep(time = 3)
cat("I'm back!\n")
if(!exists("merged_sample_genotype_df")) {
  merged_sample_genotype_df <-
    read_tsv(
      file = "/home/luis3m/DATA/merged_sample_genotype_df.txt",
      col_names = F,
      progress = T,
      col_types = cols()
    )
}

beep(sound = 6)
Sys.sleep(time = 3)
```

Figure 5.8: Merge multiple genotype data files.

```
cat("Cleaning up data...\n")
# Remove unnecessary SAMPLES columns
aux_index <- dim(data01_sample_genotype_df)[2] + 1
merged_sample_genotype_df[aux_index] <- NULL

aux_index <-
  dim(data01_sample_genotype_df)[2] +
  dim(data03_sample_genotype_df)[2]
merged_sample_genotype_df[aux_index] <- NULL

aux_index <-
  dim(data01_sample_genotype_df)[2] +
  dim(data03_sample_genotype_df)[2] +
  dim(data04_sample_genotype_df)[2]
merged_sample_genotype_df[aux_index-1] <- NULL

aux_index <-
  dim(data01_sample_genotype_df)[2] +
  dim(data03_sample_genotype_df)[2] +
  dim(data04_sample_genotype_df)[2] +
  dim(data08_sample_genotype_df)[2]
merged_sample_genotype_df[aux_index] <- NULL

aux_index <-
  dim(data01_sample_genotype_df)[2] +
  dim(data03_sample_genotype_df)[2] +
  dim(data04_sample_genotype_df)[2] +
  dim(data08_sample_genotype_df)[2] +
  dim(data10_sample_genotype_df)[2]
merged_sample_genotype_df[aux_index] <- NULL

aux_index <-
  dim(data01_sample_genotype_df)[2] +
  dim(data03_sample_genotype_df)[2] +
  dim(data04_sample_genotype_df)[2] +
  dim(data08_sample_genotype_df)[2] +
  dim(data10_sample_genotype_df)[2] +
  dim(data12_sample_genotype_df)[2]
merged_sample_genotype_df[aux_index] <- NULL
```

Figure 5.9: Clean duplicated data from merged data file.

```
# Set merged_sample_genotype_df colnames
aux_colnames <- append(
    c("SAMPLES", unlist(unique(data_01["SNP"])), use.names = F),
    c(unlist(unique(data_03["SNP"]), use.names = F))
)
aux_colnames <- append(
  aux_colnames,
  c(unlist(unique(data_04["SNP"]), use.names = F))
)
aux_colnames <- append(
  aux_colnames,
  c(unlist(unique(data_08["SNP"]), use.names = F))
)
aux_colnames <- append(
  aux_colnames,
  c(unlist(unique(data_10["SNP"]), use.names = F))
)
aux_colnames <- append(
  aux_colnames,
  c(unlist(unique(data_12["SNP"]), use.names = F))
)
aux_colnames <- append(
  aux_colnames,
  c(unlist(unique(data_22["SNP"]), use.names = F))
)

col_names <- aux_colnames
colnames(merged_sample_genotype_df) <- col_names
remove(aux_colnames, col_names)

beep(sound = 3)
Sys.sleep(time = 3)
cat("DONE! Moving on...\n\n")
Sys.sleep(time = 10)
```

Figure 5.10: Label merged data file accordingly.

In the next section we will present our clustering analysis as well as every single detail of it.

5.4 Data Analysis

5.4.1 Results

As it has been described in previous sections we started our data analysis by looking into our sample's clinical data in order to better understand our population so we could find answers to the following questions:

- *How should we approach this clustering analysis?*
- *What demographic data has been disclosed?*
- *In what way will this be helpful regarding our research ?*

A demographic analysis has been done and presented in previous sections, however we will reveal our R code that was able to manage this task, depicted in Figures 5.11, 5.12, 5.13 and 5.14.

```r
# HOW MANY SAMPLES ?
samples_count <- function() {
  print("HOW MANY SAMPLES ?")
  length_samples <- length(unique(sampleBD$SAMPLE))
  #cat(paste0("Total: ",length_samples))
  samples_total_count <- data.frame(length_samples)
  colnames(samples_total_count) <- c("TOTAL")
  return(samples_total_count)
}
samples_count_df <- samples_count()
```

Figure 5.11: R code used to perform a demographic analysis I.

```r
# HOW MANY SAMPLE OF EACH GENDER (male = 1, female = 2)?
samples_gender_count <- function() {
  print("HOW MANY SAMPLES OF EACH GENDER ?")
  gender_type <- c("male","female")
  gender_ctr <- c(sum(sampleBD$GENDER==1),sum(sampleBD$GENDER==2))
  gender_count <- data.frame(gender_type,gender_ctr)
  colnames(gender_count) <- c("GENDER", "COUNT")
  return(gender_count)
}
samples_gender_count_df <- samples_gender_count()
# Plot Gender count
ggplot(samples_gender_count_df, aes(samples_gender_count_df$GENDER, samples_gender_count_df$COUNT, fill = samples_gender_count_df$GENDER, label = samples_gender_count_df$COUNT))+
  geom_bar(stat="identity", position = "dodge", width = 0.25)+
  geom_text(size = 4, position = position_dodge(width = 0))+
  labs(title = "Demographic Analysis", subtitle = paste0("Gender"), fill= "")+
  xlab(label = "")+
  ylab(label = "# of Patients per Gender")+
  scale_fill_brewer(palette = "Dark2")+theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
                                             panel.background = element_blank(), axis.line = element_line(colour = "black"))
```

Figure 5.12: R code used to perform a demographic analysis II.

```r
# FROM WHICH REGION ?
samples_region_count <- function() {
  print("HOW MANY PATIENTS PER REGIONS ?")
  regions <- unique(sampleBD$REGION)
  regions_count_aux <- c()
  for (item in regions) {
    #cat(paste0(region,": ",sum(sampleBD$REGION==region),"\n"))
    regions_count_aux <- append(regions_count_aux, values = sum(sampleBD$REGION==item))
  }
  #regions_count <- as.data.frame(cbind(regions,regions_count_aux))
  regions_count <- data.frame(regions,regions_count_aux)
  colnames(regions_count) <- c("REGION", "COUNT")
  return(regions_count)
}
samples_region_count_df <- samples_region_count()
# Plot Region count
ggplot(samples_region_count_df, aes(samples_region_count_df$REGION, samples_region_count_df$COUNT, fill = samples_region_count_df$REGION, label = samples_region_count_df$COUNT))+
  geom_bar(stat="identity", position = "dodge", width = 1)+
  geom_text(size = 3, position = position_dodge(0.5))+
  labs(title = "Demographic Analysis", subtitle = paste0("Region"), fill= "")+
  xlab(label = "")+
  ylab(label = "# of Patients per Region")+
  scale_fill_brewer(palette = "Paired")+theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
                                              panel.background = element_blank(), axis.line = element_line(colour = "black"))+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

Figure 5.13: R code used to perform a demographic analysis III.

```r
# AVERAGE AGE_RECRUITMENT ?
samples_age_recruitment <- function() {
  print("HOW MANY PATIENTS WITHIN AN AGE INTERVAL ?")
  ages <- as.character(sort(unique(sampleBD$AGE_RECRUITMENT)))
  ages_count_aux <- c()
  for (index in c(1:length(ages))) {
    ages_count_aux <- append(ages_count_aux, values = sum(sampleBD$AGE_RECRUITMENT==ages[index]))
    ages[index] <- paste0(ages[index],"0","-",ages[index],"9")
  }
  age_recruitment <- data.frame(ages,ages_count_aux)
  colnames(age_recruitment) <- c("AGE_RECRUITMENT", "COUNT")
  return(age_recruitment)
}
samples_age_recruitment_df <- samples_age_recruitment()
# Plot Age_Recruitment count - BarPlot
ggplot(samples_age_recruitment_df, aes(samples_age_recruitment_df$AGE_RECRUITMENT, samples_age_recruitment_df$COUNT, fill = samples_age_recruitment_df$AGE_RECRUITMENT, label = samples_age_recruitment_df$COUNT))+
  geom_bar(stat="identity", position = "dodge", width = 1)+
  geom_text(size = 3, position = position_dodge(0.5))+
  labs(title = "Demographic Analysis", subtitle = paste0("Age of Recruitment"), fill= "Age Interval")+
  xlab(label = "")+
  ylab(label = "# of Patients per Age Interval")+
  scale_fill_brewer(palette = "Paired")+theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
                                              panel.background = element_blank(), axis.line = element_line(colour = "black"))+
  theme(axis.text.x = element_text(angle = 0, hjust = 0.5))

# Age_Recruitment statistical summary
summary(sampleBD$AGE_RECRUITMENT)
```

Figure 5.14: R code used to perform a demographic analysis IV.

Upon this task completion, and after having our genetic data ready for clustering analysis, we started our analysis by assessing k-means optimal number of clusters. As our main goal relies on finding subgroups of patients and then to analyze their clinical and genetic data, we needed to make sure that we have used the optimal k value whilst running the algorithm. Hence, we have used two methods:

- Elbow method;
- Silhouette method.

As we were going after a massive data clustering and k-means has shown to be quiet sensitive to the k clusters value initially chosen [43], we needed to make sure that we have picked the best value in order to avoid misleading information regarding our subgroups local analysis. Therefore, these two methods were used throughout this project development.

Even though an unsupervised approach has been used, we followed the idea of splitting our data set into a train and test data set (commonly seen whilst using a supervised approach) as we wanted to guarantee that our machine learning model would be suitable when applied in unseen data. Hence, we have carried along an optimal k value assessment task in $\frac{1}{3}$ of our data (randomly chosen) and we have performed a clustering analysis on the remaining data.

Here we present the R code used to perform all these tasks:

1. First of all we have removed data frame's first column as it contains all sample's IDs unnecessary to carry on this analysis (depicted in Figure 5.15);

```
# Extract genotype data from merged df
cluster_data <- merged_sample_genotype_df[, 2:length(merged_sample_genotype_df)]
```

Figure 5.15: R code used to remove sample's IDs out of the data frame.

2. We had to make sure that no missing values were present in our data frame (depicted in Figure 5.16);

```
> sum(is.na(cluster_data))
[1] 0
```

Figure 5.16: R code used to test missing data in *cluster_data* data frame.

3. We have carried on the data split task which has extracted $\frac{1}{3}$ of the *cluster_data* (labeled

as *ktrain_data*) data frame in order to use it as train data, therefore $\frac{2}{3}$ of the *cluster_data* (labeled as *ktest_data*) data frame were used as test data (depicted in Figure 5.17). ;

```
# Assess optimal k clusters value
# Method: "silhouette", "wss"
cat("K CLUSTERS OPTIMAL VALUE ASSESSMENT\n")
set.seed(101)
random_samples <- sample.int(n = nrow(cluster_data), size = floor(nrow(cluster_data)*0.3), replace = F)
ktrain_data <- cluster_data[random_samples, ]
ktest_data  <- cluster_data[-random_samples, ]
```

Figure 5.17: R code used to split up *cluster_data* data frame.

4. The Elbow method was used to assess optimal k clusters value (depicted in Figure 5.18) in which we have set $n = 100$ as our initial centroid configuration so the algorithm should attempt 100 times to calculate the best centroid and to choose the best one before starting to calculate all within-cluster sum of squares. Therefore, we have calculate all within-cluster sum of squares for k values between 1 and 10, as previously explained. We also present the Elbow chart that has been retrieved after running this test (depicted in Figure 5.19);

```
cat("WSS METHOD\n")
k_values <- 1:10
parallel_cluster <- makeCluster(detectCores()-1, type = "PSOCK")
registerDoParallel(parallel_cluster)
wss_values <- foreach(k = k_values, .combine = 'c', .verbose = T) %dopar% {
  kmeans(ktrain_data, k, nstart = 100)$tot.withinss
}
stopCluster(parallel_cluster)
remove(parallel_cluster)
plot(
  seq_along(wss_values),
  wss_values,
  type = "b",
  pch = 9,
  main = "Optimal number of clusters",
  xlab = "Number of k clusters",
  ylab = "Within sum of squares"
)
cat("Sleeping...\n")
Sys.sleep(time = 30)
cat("\n")
```

Figure 5.18: R code used to implement a parallel approach of the Elbow method.

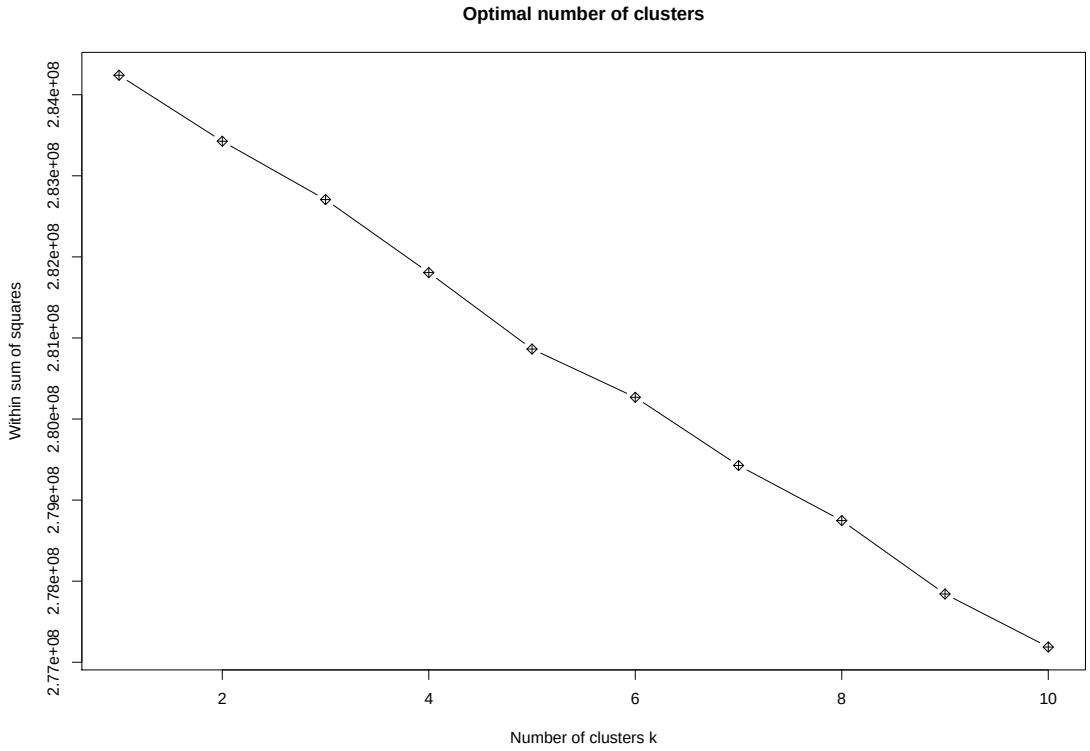


Figure 5.19: Elbow method chart results.

5. A similar approach has been carried on but this time by using the Silhouette method: it was used to assess optimal k clusters value (depicted in Figure 5.20) in which we have also set $n = 100$ as our initial centroid configuration. Therefore, we have calculate all average silhouette width values for k values between 2 and 10, as previously explained (we have not tested $k = 1$ as the average silhouette width of one single cluster is 0). We also present the Silhouette chart that has been retrieved after running this test (depicted in Figure 5.21);

```
cat("SILHOUETTE METHOD\n")
k_values <- 2:10
parallel_cluster <- makeCluster(detectCores()-1, type = "PSOCK")
registerDoParallel(parallel_cluster)
ss_values <- foreach(k = k_values, .combine = 'c', .packages = 'cluster', .verbose = T) %dopar% {
  kmeans_res <- kmeans(ktrain_data, k, nstart = 100)
  ss <- silhouette(kmeans_res$cluster, dist(ktrain_data, method = "euclidean"))
  mean(ss[, 3])
}
stopCluster(parallel_cluster)
remove(parallel_cluster)
plot(
  seq_along(ss_values),
  ss_values,
  type = "b",
  pch = 9,
  main = "Optimal number of clusters",
  xlab = "Number of k clusters",
  ylab = "Average silhouette width"
)
abline(v = which.max(ss_values), lty = 2)
cat("Sleeping...\n")
Sys.sleep(time = 30)
cat("\n")
```

Figure 5.20: R code used to implement a parallel approach of the Silhouette method.

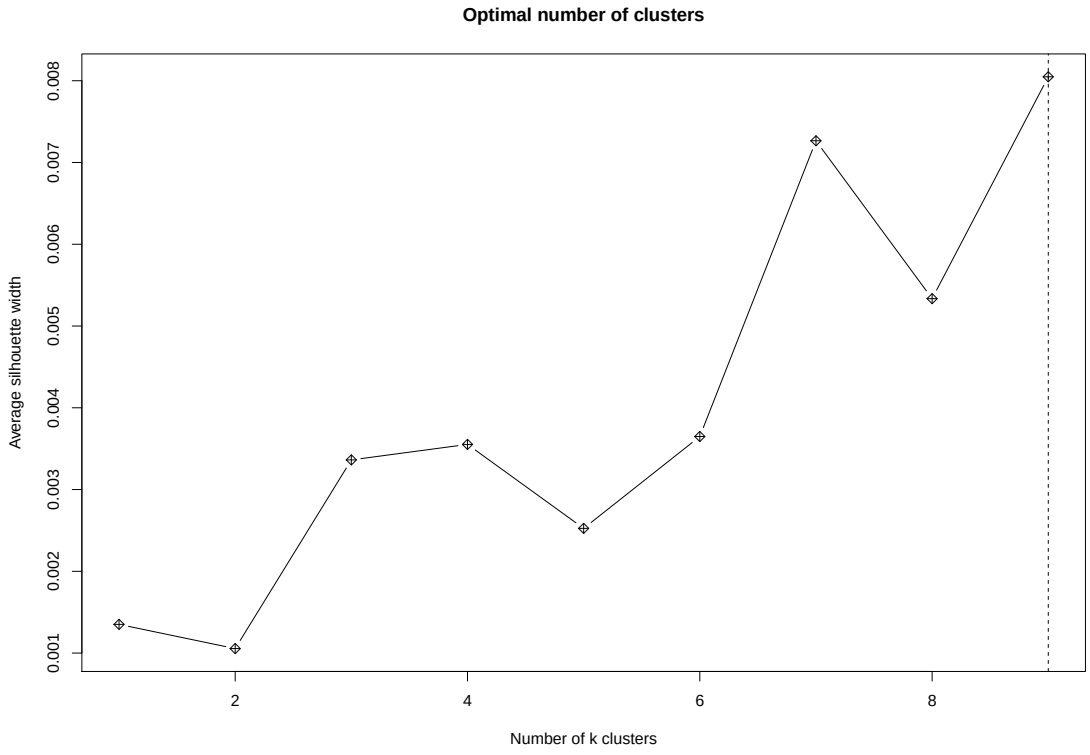


Figure 5.21: Silhouette method chart results.

Once these results were analyzed there was no clear k value to be set during our clustering analysis. In fact, both Elbow and Silhouette method reveal that our data does not present

clear clusters to which samples could be assigned to. However, as we are interested in finding subgroups of bipolar disorder patients out of our *ktest_data* data frame, regarding their genetic variants, we have correlated both methods' charts in order to select what we have considered to be the best k value: both have shown that $k = 7$ could be a suitable value to start our analysis. The Elbow method was not clear enough but by looking into the Silhouette method results we were able to see that $k = 7$ was our second best result. In fact, Elbow also reveals a slight bend (knee) when $k = 7$. Therefore, we have started our cluster analysis by using k-means algorithm and setting the k clusters value to 7 as depicted in Figure 5.21.

```
# Apply Kmeans clustering
start_exec <- Sys.time()
cat("APPLYING KMEANS CLUSTERING\n")
kmeans_data <- kmeans(ktest_data, centers = 7, nstart = 100, trace = T)
cat("\n")
cat("DONE! Please start data analysis...")
end_exec <- Sys.time()
cat("Execution time: ",end_exec-start_exec)
```

Figure 5.22: R code used to apply the k-means algorithm.

This part of the cluster analysis took more than two days to complete, as we have used a single thread process (as presented in Table 5.7) due to lack of resources spotted whilst handling our initial parallel approach that tried to execute this algorithm *ncore* times and retrieve the best outcome out of those k-means results. Unfortunately it requires more memory than the one we had available.

Execution start time	**Execution end time**
2018-09-08 09:42:04	*2018-09-10 15:13:59*

Table 5.7: K-means execution time.

Here we present our clusters specifications, regarding total samples that have been associated to one of the 7 available groups. Also, we present the number of samples within each group (depicted in Figure 5.23), as well as the R code used to perform such analysis that has required to retrieve the original samples of each data set, upon a restart of our RStudio server (depicted in Figure 5.24):

- Total samples in *ktrain_data*: 599;
- Total samples in *ktest_data*: 1399;
- Total samples per cluster:
  - cluster 1: 409;
  - cluster 2: 560;
  - cluster 3: 6;
  - cluster 4: 396;
  - cluster 5: 2;
  - cluster 6: 24;
  - cluster 7: 2.

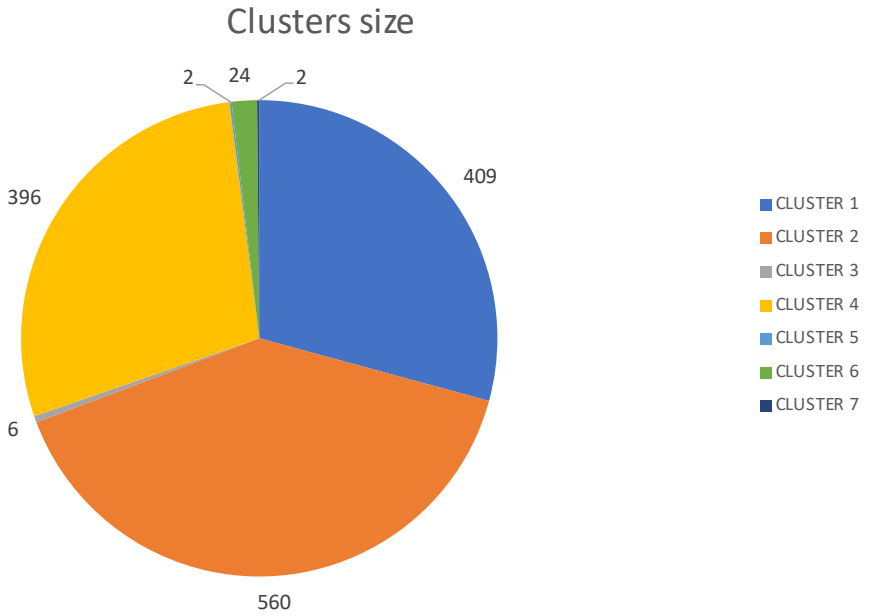


Figure 5.23: K-means clusters size chart.

```
# Retrieve samples of each data set (train/test data)
kmeans_clusters <- kmeans_data$cluster
ktest_samples_id <- with(join.keys(ktest_data, cluster_data), match(x, y))
ktest_samples <- (sampleBD$SAMPLE[ktest_samples_id])
ktrain_samples <- (sampleBD$SAMPLE[-ktest_samples_id])

# Assign each sample to one cluster according to kmeans results
length(ktest_samples) == length(kmeans_clusters)
# Create 7 aux arrays/data frames (associated to each cluster)
cluster_1_samples <-
  cluster_2_samples <-
  cluster_3_samples <-
  cluster_4_samples <-
  cluster_5_samples <-
  cluster_6_samples <-
  cluster_7_samples <- c()
for (index in 1:length(kmeans_clusters)) {
  if (kmeans_clusters[index] == 1) {
    cluster_1_samples <- append(cluster_1_samples, ktest_samples[index])
  } else if (kmeans_clusters[index] == 2) {
    cluster_2_samples <- append(cluster_2_samples, ktest_samples[index])
  } else if (kmeans_clusters[index] == 3) {
    cluster_3_samples <- append(cluster_3_samples, ktest_samples[index])
  } else if (kmeans_clusters[index] == 4) {
    cluster_4_samples <- append(cluster_4_samples, ktest_samples[index])
  } else if (kmeans_clusters[index] == 5) {
    cluster_5_samples <- append(cluster_5_samples, ktest_samples[index])
  } else if (kmeans_clusters[index] == 6) {
    cluster_6_samples <- append(cluster_6_samples, ktest_samples[index])
  } else if (kmeans_clusters[index] == 7) {
    cluster_7_samples <- append(cluster_7_samples, ktest_samples[index])
  } else {
    cat("Something went wrong!", "\n", "Please check your data.")
  }
}
cluster_1 <- data.frame("SAMPLES" = cluster_1_samples)
cluster_2 <- data.frame("SAMPLES" = cluster_2_samples)
cluster_3 <- data.frame("SAMPLES" = cluster_3_samples)
cluster_4 <- data.frame("SAMPLES" = cluster_4_samples)
cluster_5 <- data.frame("SAMPLES" = cluster_5_samples)
cluster_6 <- data.frame("SAMPLES" = cluster_6_samples)
cluster_7 <- data.frame("SAMPLES" = cluster_7_samples)
remove(cluster_1_samples, cluster_2_samples,
       cluster_3_samples, cluster_4_samples,
       cluster_5_samples, cluster_6_samples,
       cluster_7_samples)
```

Figure 5.24: R code used to assign samples to each associated cluster data frame.

Subsequent to the cluster size analysis we have generated a plot for each cluster regarding the clinical factor gender. We intend to correlated data that reveals a plausible relation between gender and genetic data, hence we first present how many samples of each gender were assigned to each cluster (depicted in Appendix A.9, A.10, A.11, A.12, A.13, A.14 and A.15) which is summarized this information in Table 5.8.

Cluster	Gender	Count
1	*Male*	152
	Female	257
2	*Male*	202
	Female	358
3	*Male*	1
	Female	5
4	*Male*	146
	Female	250
5	*Male*	2
	Female	0
6	*Male*	9
	Female	15
7	*Male*	0
	Female	2

Table 5.8: K-means clusters gender ratio.

Furthermore, we have started our analysis by filling up our cluster's data frames with relevant clinical data, as depicted in Figure 5.25 for cluster 1, also in Appendix A.16, A.17, A.18, A.19, A.20 and A.21 we present our R code for the remaining clusters.

```
# Clinical Data Analysis per cluster
# CLUSTER 1
aux_samples_region <- c()
for (patient in cluster_1$SAMPLES) {
  aux_samples_region <- append(aux_samples_region,sampleBD[sampleBD$SAMPLE==patient,"REGION"])
}
cluster_1$REGION <- unlist(aux_samples_region, recursive = T, use.names = F)
remove(aux_samples_region)

aux_samples_age <- c()
for (patient in cluster_1$SAMPLES) {
  aux_samples_age <- append(aux_samples_age,sampleBD[sampleBD$SAMPLE==patient,"AGE_RECRUITMENT"])
}
cluster_1$AGE_RECRUITMENT <- unlist(aux_samples_age, recursive = T, use.names = F)
remove(aux_samples_age)

aux_samples_gender <- c()
for (patient in cluster_1$SAMPLES) {
  aux_samples_gender <- append(aux_samples_gender,sampleBD[sampleBD$SAMPLE==patient,"GENDER"])
}
cluster_1$GENDER <- unlist(aux_samples_gender, recursive = T, use.names = F)
remove(aux_samples_gender)
```

Figure 5.25: R code used to retrieve samples' clinical data of each cluster.

We shall now retrieve each gene's SNPs genotype data per sample inside every cluster according to the correct chromosome. Furthermore, we must find every SNP in our chromosomes data, otherwise it means that it is not present in WTCCC genetic data and we need to discard it.

Regarding the ones that are available we have decided to present and compare the five most countable samples between clusters, in a case where the number of samples within a cluster is greater than five.

- DISC1 (chromosome 1):
  - rs203368
  - rs435136

  After querying chromosome 1 data set we have found that there is no data available regarding these two genetic variants of gene DISC1, as depicted in Figure 5.26. Hence, we will skip it and analyze the remaining genes.

```
> sum(data_01$SNP=="rs203368")
[1] 0
> sum(data_01$SNP=="rs435136")
[1] 0
```

Figure 5.26: R code used to query gene DISC1 chromosome 1 data.

- ARPP21 (chromosome 3):
  - rs1523041

  After querying chromosome 3 data set we have found that there is indeed data available regarding this genetic variant of gene ARPP21, as depicted in Figure 5.27 (as expected it matches the number of samples). Therefore, we will start our genetic analysis by (1) retrieving each samples' genotypes regarding this SNP and (2) by appending its results to our clusters' data frames so a pattern can be extracted out of our clusters data frames. In order to retrieve such data we have used the following R code presented in Appendix A.22 (we have used parallel computation as several queries were made to each chromosome data).

```
> sum(data_03$SNP=="rs1523041")
[1] 1998
```

Figure 5.27: R code used to query gene ARPP21 chromosome 3 data.

Once results were achieved (presented in Table 5.9) we intend to study the presence of one specific allele (risk allele), *C*, as it has been associated with possible interaction between sex and diagnostic [49].

	Region	Age of Recruitment	Gender	rs1523041	Count
Cluster 1	*Wales*	*4*	*2*	*CC*	*10*
	Wales	*4*	*2*	*CG*	*9*
	Scotland	*5*	*2*	*CG*	*8*
	Northern	*4*	*2*	*CG*	*7*
	Northern	*5*	*2*	*CC*	*7*
Cluster 2	*Midlands*	*5*	*2*	*CG*	*14*
	Midlands	*5*	*2*	*CC*	*12*
	Midlands	*3*	*2*	*CG*	*11*
	Midlands	*4*	*2*	*CG*	*11*
	Wales	*4*	*1*	*CG*	*11*
Cluster 3	*North Midlands*	*4*	*2*	*CC*	*1*
	North Midlands	*1*	*2*	*CG*	*1*
	Scotland	*6*	*2*	*CG*	*1*
	Scotland	*3*	*1*	*GG*	*1*
	Scotland	*3*	*2*	*GG*	*1*
Cluster 4	*Midlands*	*5*	*2*	*CG*	*12*
	Midlands	*4*	*2*	*CG*	*10*
	Midlands	*4*	*2*	*CC*	*8*
	Midlands	*5*	*1*	*CC*	*7*
	Midlands	*3*	*2*	*CG*	*7*
Cluster 5	*Northern*	*6*	*1*	*CC*	*2*
Cluster 6	*Midlands*	*4*	*2*	*CG*	*4*
	Wales	*3*	*1*	*CC*	*2*
	Midlands	*3*	*1*	*CG*	*2*
	Southern	*2*	*2*	*CG*	*2*
	Wales	*2*	*2*	*CG*	*2*
Cluster 7	*Northwestern*	*3*	*2*	*GG*	*1*
	Southwestern	*3*	*2*	*GG*	*1*

Table 5.9: ARPP21 gene clustering analysis.

- GABRB1 (chromosome 4):
  - rs7680321

As similar to previous approach we have started our analysis by making sure that there is data available regarding this GABRB1 genetic variant, as depicted in Figure 5.28.

Upon this task completion we present our results in Table 5.10 after following previous procedure. Furthermore, we focus our research on allele (risk allele), *C*, as also this variation has been associated with presence of bipolar disorder [10] [22].

```
> sum(data_04$SNP=="rs7680321")
[1] 1998
```

Figure 5.28: R code used to query gene GABRB1 chromosome 4 data.

	Region	**Age of Recruitment**	**Gender**	**rs7680321**	**Count**
Cluster 1	*Northern*	*5*	*2*	*TT*	*17*
	Wales	*4*	*2*	*TT*	*16*
	Wales	*4*	*1*	*TT*	*11*
	Scotland	*5*	*2*	*TT*	*10*
	Wales	*3*	*2*	*TT*	*10*
Cluster 2	*Midlands*	*5*	*2*	*TT*	*27*
	Midlands	*4*	*2*	*TT*	*22*
	Midlands	*3*	*2*	*TT*	*20*
	Wales	*4*	*2*	*TT*	*20*
	Wales	*4*	*1*	*TT*	*16*
Cluster 3	*North Midlands*	*1*	*2*	*TT*	*1*
	North Midlands	*4*	*2*	*TT*	*1*
	Scotland	*3*	*1*	*TT*	*1*
	Scotland	*3*	*2*	*TT*	*1*
	Scotland	*5*	*2*	*CT*	*1*
Cluster 4	*Midlands*	*4*	*2*	*TT*	*20*
	Midlands	*5*	*2*	*TT*	*16*
	Midlands	*5*	*1*	*TT*	*12*
	Midlands	*3*	*2*	*TT*	*11*
	Wales	*4*	*1*	*TT*	*11*
Cluster 5	*Northern*	*6*	*1*	*TT*	*2*
Cluster 6	*Midlands*	*4*	*2*	*TT*	*4*
	Midlands	*3*	*2*	*TT*	*3*
	Midlands	*3*	*1*	*TT*	*2*
	Midlands	*4*	*1*	*TT*	*2*
	Southern	*2*	*2*	*TT*	*2*
Cluster 7	*Northwestern*	*3*	*2*	*TT*	*1*
	Southwestern	*3*	*2*	*TT*	*1*

Table 5.10: GABRB1 gene clustering analysis.

- ANKRD46 (chromosome 8):
  - rs80198067

  After querying chromosome 8 data set we have found that there is no data available regarding this genetic variant of gene ANKRD46, as depicted in Figure 5.29. Hence, we will skip it and analyze remaining genes.

```
> sum(data_08$SNP=="rs80198067")
[1] 0
```

Figure 5.29: R code used to query gene ANKRD46 chromosome 8 data.

- ANK3 (chromosome 10):
  - rs10994336
  - rs9804190

  After querying chromosome 10 data we have received a negative result, as no data is available regarding these two SNPs of gene ANK3, as depicted in Figure 5.30. Hence, we will skip it and analyze remaining genes.

```
> sum(data_10$SNP=="rs10994336")
[1] 0
> sum(data_10$SNP=="rs9804190")
[1] 0
```

Figure 5.30: R code used to query gene ANK3 chromosome 10 data.

- CACNA1C (chromosome 12):
  - rs1006737
  - rs4765914
  - rs4765913
  - rs2239063

  After querying this chromosome data we have confirmed that there is data available regarding CACNA1C genetic variants, but only for SNPs rs1006737 and rs4765914, as depicted in Figure 5.31.

Upon this task completion we present our results in Tables 5.11 and 5.12 respectively after following previous procedure. Furthermore, we focus our research on alleles (risk alleles), A and T, respectively for SNPs rs1006737 and rs4765914.

```
> sum(data_12$SNP=="rs1006737")
[1] 1998
> sum(data_12$SNP=="rs4765914")
[1] 1998
> sum(data_12$SNP=="rs4765913")
[1] 0
> sum(data_12$SNP=="rs2239063")
[1] 0
```

Figure 5.31: R code used to query gene CACNA1C chromosome 12 data.

	Region	Age of Recruitment	Gender	rs1006737	Count
Cluster 1	*Northern*	*5*	*2*	*AG*	*12*
	Wales	*4*	*2*	*AG*	*10*
	Northern	*4*	*2*	*AG*	*7*
	Scotland	*4*	*2*	*GG*	*7*
	Wales	*4*	*2*	*GG*	*7*
Cluster 2	*Midlands*	*5*	*2*	*GG*	*17*
	Wales	*4*	*2*	*AG*	*15*
	Midlands	*3*	*2*	*AG*	*14*
	Midlands	*4*	*2*	*AG*	*14*
	Midlands	*4*	*2*	*GG*	*11*
Cluster 3	*North Midlands*	*1*	*2*	*GG*	*1*
	North Midlands	*4*	*2*	*AA*	*1*
	Scotland	*3*	*1*	*AG*	*1*
	Scotland	*3*	*2*	*AG*	*1*
	Scotland	*5*	*2*	*AG*	*1*
Cluster 4	*Midlands*	*4*	*2*	*GG*	*13*
	Midlands	*6*	*2*	*AG*	*10*
	Midlands	*5*	*2*	*GG*	*9*
	Midlands	*5*	*1*	*AG*	*8*
	Midlands	*5*	*2*	*AG*	*8*
Cluster 5	*Northern*	*6*	*1*	*GG*	*2*
Cluster 6	*Midlands*	*4*	*2*	*AG*	*4*
	Wales	*3*	*1*	*GG*	*4*
	Midlands	*3*	*2*	*GG*	*3*
	Midlands	*3*	*1*	*GG*	*2*
	Midlands	*4*	*1*	*GG*	*2*
Cluster 7	*Northwestern*	*3*	*2*	*AG*	*1*
	Southwestern	*3*	*2*	*AG*	*1*

Table 5.11: CACNA1C gene clustering analysis I.

	Region	Age of Recruitment	Gender	rs4765914	Count
Cluster 1	*Wales*	*4*	*2*	*CT*	*12*
	Scotland	*5*	*2*	*CC*	*10*
	Wales	*4*	*2*	*CC*	*10*
	Wales	*5*	*2*	*CC*	*9*
	Northern	*5*	*2*	*CC*	*8*
Cluster 2	*Midlands*	*5*	*2*	*CC*	*20*
	Midlands	*4*	*2*	*CC*	*16*
	Wales	*4*	*2*	*CT*	*14*
	Midlands	*3*	*2*	*CC*	*11*
	Midlands	*3*	*2*	*CT*	*11*
Cluster 3	*North Midlands*	*1*	*2*	*CC*	*1*
	North Midlands	*4*	*2*	*CC*	*1*
	Scotland	*3*	*1*	*CC*	*1*
	Scotland	*3*	*2*	*CT*	*1*
	Scotland	*5*	*2*	*TT*	*1*
Cluster 4	*Midlands*	*4*	*2*	*CC*	*14*
	Midlands	*5*	*2*	*CC*	*12*
	Midlands	*5*	*1*	*CC*	*11*
	Wales	*4*	*1*	*CC*	*10*
	Midlands	*6*	*2*	*CC*	*9*
Cluster 5	*Northern*	*6*	*1*	*CC*	*2*
Cluster 6	*Midlands*	*4*	*2*	*CT*	*4*
	Wales	*3*	*1*	*CC*	*4*
	Midlands	*3*	*2*	*CC*	*3*
	Midlands	*3*	*1*	*CC*	*2*
	Midlands	*4*	*1*	*CT*	*2*
Cluster 7	*Northwestern*	*3*	*2*	*CT*	*1*
	Southwestern	*3*	*2*	*CT*	*1*

Table 5.12: CACNA1C gene clustering analysis II.

- DUSP6 (chromosome 12):
  - rs769700
  - rs704076
  - rs770087
  - rs808820
  - rs2279574

As similar to other SNPs we have queried chromosome 12 data and we have received

a negative result, as no data is available regarding these five SNPs of gene DUSP6, as depicted in Figure 5.32. Hence, we will skip it and analyze remaining genes.

```
> sum(data_12$SNP=="rs769700")
[1] 0
> sum(data_12$SNP=="rs704076")
[1] 0
> sum(data_12$SNP=="rs770087")
[1] 0
> sum(data_12$SNP=="rs808820")
[1] 0
> sum(data_12$SNP=="rs2279574")
[1] 0
```

Figure 5.32: R code used to query gene DUSP6 chromosome 12 data.

- GRIN2B (chromosome 12):
  - rs1805502
  - rs1805247
  - rs7301328

  Again and similar to other SNPs we have queried chromosome 12 data and we have received a negative result, as no data is available regarding these three SNPs of gene GRIN2B, as depicted in Figure 5.33. Hence, we will skip it and analyze remaining genes.

```
> sum(data_12$SNP=="rs1805502")
[1] 0
> sum(data_12$SNP=="rs1805247")
[1] 0
> sum(data_12$SNP=="rs7301328")
[1] 0
```

Figure 5.33: R code used to query gene GRIN2B chromosome 12 data.

- SYN3 (chromosome 22):
  - rs9621532

  After querying this chromosome data we have confirmed that there is data available regarding SYN3 genetic variant rs9621532, as depicted in Figure 5.32.

  Upon this task completion we present our results in Table 5.13 after following previous procedure. Furthermore, we focus our research on allele (risk allele), *A*.

```
> sum(data_22$SNP=="rs9621532")
[1] 1998
```

Figure 5.34: R code used to query gene SYN3 chromosome 22 data.

	Region	Age of Recruitment	Gender	rs9621532	Count
Cluster 1	*Wales*	*4*	*2*	*AA*	*21*
	Northern	*5*	*2*	*AA*	*15*
	Scotland	*5*	*2*	*AA*	*11*
	Wales	*4*	*1*	*AA*	*11*
	Wales	*5*	*2*	*AA*	*11*
Cluster 2	*Midlands*	*5*	*2*	*AA*	*23*
	Wales	*4*	*2*	*AA*	*23*
	Midlands	*4*	*2*	*AA*	*22*
	Midlands	*3*	*2*	*AA*	*18*
	Wales	*4*	*1*	*AA*	*18*
Cluster 3	*North Midlands*	*1*	*2*	*AA*	*1*
	North Midlands	*4*	*2*	*AC*	*1*
	Scotland	*3*	*1*	*AC*	*1*
	Scotland	*3*	*2*	*AA*	*1*
	Scotland	*5*	*2*	*CC*	*1*
Cluster 4	*Midlands*	*5*	*2*	*AA*	*19*
	Midlands	*4*	*2*	*AA*	*16*
	Midlands	*5*	*1*	*AA*	*15*
	Midlands	*3*	*2*	*AA*	*11*
	Wales	*4*	*1*	*AA*	*11*
Cluster 5	*Northern*	*6*	*1*	*AA*	*2*
Cluster 6	*Midlands*	*4*	*2*	*AA*	*4*
	Midlands	*3*	*2*	*AA*	*3*
	Midlands	*3*	*1*	*AA*	*2*
	Midlands	*4*	*1*	*AA*	*2*
	Southern	*2*	*2*	*AA*	*2*
Cluster 7	*Northwestern*	*3*	*2*	*AC*	*1*
	Southwestern	*3*	*2*	*AC*	*1*

Table 5.13: SYN3 gene clustering analysis.

5.4.2 Discussion

We shall now cross-validate each SNPs genotypes data with available clinical features, in order to answer the following question: *Is there a plausible relation between gender and genetic variants regarding presence of bipolar disorder?*

Before discussing our results, we must highlight that all samples from our data set were considered *a priori* to be correctly diagnosed as bipolar disorder patients. Hence, an accurate diagnostic has been made in the first place.

We have found many particular patterns, regarding these samples clinical-genetic data, and we described them in the following list:

1. Even before going into detail regarding each clinical-genetic relation, there are two clusters that caught our attention - cluster 5 and cluster 7. These two clusters have the exact same size (as presented in Table 5.8), and each cluster samples' details are now described:

   - *Cluster 5*: 2 male individuals between 60 and 69 years old from Northern England;
   - *Cluster 7*: 2 female individuals between 30 and 39 years old from Southwestern England and Northwestern England respectively.

   We went across all our results and we could identify one specific pattern: analyzing samples of both of these two subgroups, more than having the same age range within clusters, these individuals maintain the exact same genotype across different genes (ARPP21, GABRB1, CACNA1C and SYN3) as shown in Table 5.14 and 5.15 respectively, which suggests that samples of each subgroup might have a common ancestor and be related to each other.

Samples	Region	Age of Recruitment	Gender	rs1523041	rs7680321	rs1006737	rs4765914	rs9621532
WTCCC65779	*Northern*	*6*	*1*	*CC*	*TT*	*GG*	*CC*	*AA*
WTCCC65407	*Northern*	*6*	*1*	*CC*	*TT*	*GG*	*CC*	*AA*

Table 5.14: Patterns extraction I.

Samples	Region	Age of Recruitment	Gender	rs1523041	rs7680321	rs1006737	rs4765914	rs9621532
WTCCC65841	*Southwestern*	*3*	*2*	*GG*	*TT*	*AG*	*CT*	*AC*
WTCCC65548	*Northwestern*	*3*	*2*	*GG*	*TT*	*AG*	*CT*	*AC*

Table 5.15: Patterns extraction II.

2. With regards to ARPP21 gene (shown in Table 5.9) we were not able to find any particular pattern of interest for our theory, as both men and women have present in rs1523041 SNP genotype the risk allele *C*. However, we can still see a trending in which women are always on top of each cluster. Also, age range shown in *Age of Recruitment* feature for all clusters shows a trending of bipolar disorder patients being between 40 and 59 years old, which

matches the average age where people are diagnosed with bipolar disorder, specially women [54].

3. With regards to GABRB1 gene (shown in Table 5.10) we were a bit surprised that almost all our samples had TT as rs7680321 SNP genotype. However, it was an expected outcome, as last research references show as link between this SNP and bipolar disorder whilst having in consideration a Han Chinese population [45]. Therefore, we would not expect the same outcome from our European cohort, yet there is one sample with the risk allele C (shown in Table 5.16), which potentially reveals the existence of a Chinese ancestral within this sample's family history. This is very plausible as the United Kingdom, made up of England, Scotland, Wales and Northern Ireland, is one of the most multicultural regions in the world.

Samples	Region	Age of Recruitment	Gender	rs1523041	rs7680321	rs1006737	rs4765914	rs9621532
WTCCC171229	*Scotland*	*5*	*2*	*GG*	*CT*	*AG*	*TT*	*CC*

Table 5.16: Patterns extraction III.

4. With regards to CACNA1C we found a really strong indicator of SNP-by-Sex interaction whilst analyzing rs1006737 SNP genotype (shown in Table 5.11) and having in consideration *Northern* region, only present in clusters 1 and 5.

   - *cluster 1*: most individuals from Northern England have the risk allele A and are women;
   - *cluster 5*: most individuals from Northern England do not have the risk allele A and are men.

   Even though not all individuals have been presented in Table 5.11, we have just shown a slight evidence of a SNP-by-Sex interaction regarding SNP rs1006737 whilst taking into account most countable samples according to our clinical-genetic features in which the majority of them are in fact women.

   There is also a similar pattern with regards to SNP rs4765914 (shown in Table 5.12) that embraces the SNP-by-Sex interaction theory: considering the risk allele T, across all subgroups and regions, it is clear that the majority of these patients which have the risk allele are women. Again, another strong evidence of a possible relation between gender and genetic variants.

5. With regards to SYN3 we were not able to retrieve a specific pattern that shows a direct relation between gender and bipolar disorder prevalence, having in consideration risk allele T of SNP rs9621532 (sown in Table 5.13). However, we are still capable of identifying most of these samples as women but there is no genetic connection that allows us to extract a pattern out of this data.

Throughout this project development we have been limited by innumerable factors that have directly affected our research outcome, such as:

- Data disclosed by WTCCC it is not rich enough, regarding sample's clinical features, and so performing an unsupervised learning task, without having that many data available, has become a huge challenge since the very beginning. Ideally, we should know a lot more about these samples' clinical features, such as age of on set, possibility of existing common ancestors within this data and main symptoms experienced by each sample during the period in which they were diagnosed as bipolar disorder patients.

- Use of k-means has its own down points as the outcome completely relies in the number of k clusters that was established *a priori*, and unfortunately there was never a clear k value to be set in the first place. As a matter of fact, our k assessment has revealed the possibility of using other k clusters values which could potentially reveal different outcomes.

- Even though we have chosen $k = 7$, as optimal k clusters value, we could indeed had a better outcome by using our initial parallel approach, which would run the k-means algorithm $ncores$ times, hence the best result would be the one with lower WSS value. However, due to lack of our server resources we were not able to run such analysis (which would require at least $ncore$ times more the amount of RAM required to load our $ktest\_data$ set), and so a single-thread approach was used instead (only one run took place). Therefore, we are inducing lack of accuracy onto our model, as more than not having a clear k clusters value, we also did not follow the optimal approach which should consider multiple results and pick the best one of them. We could have ran multiples times our single-thread approach (one at the time), but due to lack of time to do it we have carried on this process one time only - it took approximately 53 hours to finish.

In the next chapters we complete this dissertation by highlighting main results, achieved throughout this project development, as well as their relevance taking into account a SNP-by-Sex interaction in bipolar disorder cases (Chapter 6). We also reveal some of the work left undone that will certainly reveal many other aspects relevant for this mental disorder research (Chapter 7).

Chapter 6

Conclusion

Bipolar Disorder studies regarding clinical-genetic interactions have been released quiet frequently nowadays as researchers can make progress towards their goals, such as reducing this mental disorder misdiagnosis rate. We have given evidence of a plausible SNP-by-sex interaction whilst performing a clustering analysis on WTCCC bipolar disorder cases data: we were able to retrieve subgroups out of our data set as well as revealing strong evidence of a gender-disorder relationship, which might be advantageous and helpful for further studies that rely on this same data set.

Even though our results are not clear enough we can still identify relevant patterns within this data that have not yet been revealed in the past literature: (1) there are two subgroups of patients that have revealed a high probability of sharing a common ancestor (cluster 1 and cluster 5); (2) SNPs rs1006737 and rs4765914 of CACNA1C gene have revealed a pattern in where women are part of the majority of our cohort's samples groups that have in their genotype the risk allele A and T respectively for each gene SNP; (3) we have proved that different cultural population means different ancestors, hence regarding same gene SNP genotype we might have different encoding, as it has been shown for SNP rs7680321 of GABRB1 gene, in which our European population groups are less likely to have in their genotype the risk allele C, that has been shown to be present in Han Chinese population even though both have been associated with the presence of bipolar disorder.

As per our demographic analysis it was clear that women might have a higher trending of developing this disease in comparison with men. However, we were trying to find patterns of interest that could also reveal a SNP-by-Sex interaction, for a certain set of SNPs, which can lead to a breakthrough development withing the scientific area of mental disorder studies, as it can help to reduce current misdiagnosis rate and to provide people bespoke treatments, hence a better quality of life can be achieved.

We have approached this project with an unique strategy as it was never applied such analysis in that many SNPs at the same time. Hence, we have opened a window for other researchers to continue this work and investigation in order to retrieve as many genetic patterns as possible, in behalf of the 1% of the world population that is currently diagnosed with bipolar disorder.

In the next chapter we shall reveal main strategies, that should be taken in consideration, in order to allow other researchers to pursuit further goals, within the bipolar disorder research field whilst using WTCCC data as source of their research.

Chapter 7

Future Work

We have used a bespoke and unique approach throughout this project development. However, there are still many other tasks left to be done, as well as improving current research project, such as (1) improving current algorithm used during this research, with regards to the k-means method, where we should have ran it simultaneously for other k values by using other metrics, as well as other algorithms rather than analyzing one single run for $k = 7$. For instance, it will be important to assess the k clusters value with multiple distance metrics, such as Manhattan distance, as well as using other methods beside the Elbow and the Silhouette method, such as the GAP statistic assessment method.

Therefore, it will be possible to, in one hand to match new results with current achieved results, and in the other hand to reveal new genetic patterns that have not yet been discovered whilst mining WTCCC data. In order to complete such task it will be required a more powerful machine capable of handling such big data whilst running clustering algorithms, such as k-means, in multiple threads at once.

Also, it is mandatory to retrieve as many clinical features as possible from WTCCC, regarding current data set, as we are very limited with only three of them to be considered during our analysis.

Furthermore, our study have considered a strict number of chromosomes to be analyzed, regarding references found in the literature to bipolar disorder, yet there are still others to be explored in which recent literature might have found references already for them, as well as other SNPs within current set of studied chromosomes.

We firmly believe that there is a lot more to be looked at it and future work, mentioned in this section, will certainly contribute to the current SNP-by-Sex interaction research studies that still need to approach other chromosomes data as well as other clinical features relevant to this cause.

Appendix A

Appendix

```
#######################################
# Chromossome: 3                      #
# Total SNPs: 33801                   #
#######################################

# Get Genotype per sample
if (!exists("data03_genotype_per_sample_parallel")) {
  # Set parallel cluster
  parallel_cluster <- makeCluster(detectCores()-1)
  registerDoParallel(parallel_cluster)
  data03_genotype_per_sample_parallel <-
    foreach (item = 1:length(samples), .verbose = T) %dopar% {
      data_03[data_03$SAMPLE == samples[item], "GENOTYPE"]
    }
  # Stop parallel cluster
  stopCluster(parallel_cluster)
  remove(parallel_cluster)
}

# Genotype per sample merged Dataframe
if(!exists("data03_sample_genotype_df")) {
  aux_ouput_df_pwd <- "/home/luis3m/DATA/data03_sample_genotype_df.txt"
  if(file.exists(aux_ouput_df_pwd)) {
    data03_sample_genotype_df <- read_tsv(file = aux_ouput_df_pwd, col_names = F, progress = T, col_types = cols())
    colnames(data03_sample_genotype_df) <- c("SAMPLES",unlist(unique(data_03["SNP"]), use.names = F))
    remove(aux_ouput_df_pwd)
  } else {
    cat("", file = aux_ouput_df_pwd)
    for (index in 1:length(data03_genotype_per_sample_parallel)) {
      aux_ouput_df <- data.frame(t(matrix(unlist(data03_genotype_per_sample_parallel[[index]]))))
      aux_ouput_df <- cbind(samples[index], aux_ouput_df)
      write_tsv(aux_ouput_df, path = aux_ouput_df_pwd, append = T, col_names = F)
    }
    # RUN categoricalTodiscrete.sh SCRIPT!
    beep(sound = 6)
    cat("DATA 03 - RUN categoricalTodiscrete.sh SCRIPT!\n")
    cat("Sleeping...\n")
    Sys.sleep(time = 297)
    beep(sound = 3)
    Sys.sleep(time = 3)
    cat("I'm back!\n")
    data03_sample_genotype_df <- read_tsv(file = aux_ouput_df_pwd, col_names = F, progress = T, col_types = cols())
    colnames(data03_sample_genotype_df) <- c("SAMPLES",unlist(unique(data_03["SNP"]), use.names = F))
    remove(aux_ouput_df)
    remove(aux_ouput_df_pwd)
  }
}
```

Figure A.1: Retrieve SNP's genotypes per sample with parallel processing I.

```
#######################################
# Chromossome: 4                      #
# Total SNPs: 32334                   #
#######################################

# Get Genotype per sample
if (!exists("data04_genotype_per_sample_parallel")) {
  # Set parallel cluster
  parallel_cluster <- makeCluster(detectCores()-1)
  registerDoParallel(parallel_cluster)
  data04_genotype_per_sample_parallel <-
    foreach (item = 1:length(samples), .verbose = T) %dopar% {
      data_04[data_04$SAMPLE == samples[item], "GENOTYPE"]
    }
  # Stop parallel cluster
  stopCluster(parallel_cluster)
  remove(parallel_cluster)
}

# Genotype per sample merged Dataframe
if(!exists("data04_sample_genotype_df")) {
  aux_ouput_df_pwd <- "/home/luis3m/DATA/data04_sample_genotype_df.txt"
  if(file.exists(aux_ouput_df_pwd)) {
    data04_sample_genotype_df <- read_tsv(file = aux_ouput_df_pwd, col_names = F, progress = T, col_types = cols())
    colnames(data04_sample_genotype_df) <- c("SAMPLES",unlist(unique(data_04["SNP"]), use.names = F))
    remove(aux_ouput_df_pwd)
  } else {
    cat("", file = aux_ouput_df_pwd)
    for (index in 1:length(data04_genotype_per_sample_parallel)) {
      aux_ouput_df <- data.frame(t(matrix(unlist(data04_genotype_per_sample_parallel[[index]]))))
      aux_ouput_df <- cbind(samples[index], aux_ouput_df)
      write_tsv(aux_ouput_df, path = aux_ouput_df_pwd, append = T, col_names = F)
    }
    # RUN categoricalTodiscrete.sh SCRIPT!
    beep(sound = 6)
    cat("DATA 04 - RUN categoricalTodiscrete.sh SCRIPT!\n")
    cat("Sleeping...\n")
    Sys.sleep(time = 297)
    beep(sound = 3)
    Sys.sleep(time = 3)
    cat("I'm back!\n")
    data04_sample_genotype_df <- read_tsv(file = aux_ouput_df_pwd, col_names = F, progress = T, col_types = cols())
    colnames(data04_sample_genotype_df) <- c("SAMPLES",unlist(unique(data_04["SNP"]), use.names = F))
    remove(aux_ouput_df)
    remove(aux_ouput_df_pwd)
  }
}
```

Figure A.2: Retrieve SNP's genotypes per sample with parallel processing II.

```
#######################################
# Chromossome: 8                      #
# Total SNPs: 27457                   #
#######################################

# Get Genotype per sample
if (!exists("data08_genotype_per_sample_parallel")) {
  # Set parallel cluster
  parallel_cluster <- makeCluster(detectCores()-1)
  registerDoParallel(parallel_cluster)
  data08_genotype_per_sample_parallel <-
    foreach (item = 1:length(samples), .verbose = T) %dopar% {
      data_08[data_08$SAMPLE == samples[item], "GENOTYPE"]
    }
  # Stop parallel cluster
  stopCluster(parallel_cluster)
  remove(parallel_cluster)
}

# Genotype per sample merged Dataframe
if(!exists("data08_sample_genotype_df")) {
  aux_ouput_df_pwd <- "/home/luis3m/DATA/data08_sample_genotype_df.txt"
  if(file.exists(aux_ouput_df_pwd)) {
    data08_sample_genotype_df <- read_tsv(file = aux_ouput_df_pwd, col_names = F, progress = T, col_types = cols())
    colnames(data08_sample_genotype_df) <- c("SAMPLES",unlist(unique(data_08["SNP"]), use.names = F))
    remove(aux_ouput_df_pwd)
  } else {
    cat("", file = aux_ouput_df_pwd)
    for (index in 1:length(data08_genotype_per_sample_parallel)) {
      aux_ouput_df <- data.frame(t(matrix(unlist(data08_genotype_per_sample_parallel[[index]]))))
      aux_ouput_df <- cbind(samples[index], aux_ouput_df)
      write_tsv(aux_ouput_df, path = aux_ouput_df_pwd, append = T, col_names = F)
    }
    # RUN categoricalTodiscrete.sh SCRIPT!
    beep(sound = 6)
    cat("DATA 08 - RUN categoricalTodiscrete.sh SCRIPT!\n")
    cat("Sleeping...\n")
    Sys.sleep(time = 297)
    beep(sound = 3)
    Sys.sleep(time = 3)
    cat("I'm back!\n")
    data08_sample_genotype_df <- read_tsv(file = aux_ouput_df_pwd, col_names = F, progress = T, col_types = cols())
    colnames(data08_sample_genotype_df) <- c("SAMPLES",unlist(unique(data_08["SNP"]), use.names = F))
    remove(aux_ouput_df)
    remove(aux_ouput_df_pwd)
  }
}
```

Figure A.3: Retrieve SNP's genotypes per sample with parallel processing III.

```
#######################################
# Chromossome: 10                     #
# Total SNPs: 28501                   #
#######################################

# Get Genotype per sample
if (!exists("data10_genotype_per_sample_parallel")) {
  # Set parallel cluster
  parallel_cluster <- makeCluster(detectCores()-1)
  registerDoParallel(parallel_cluster)
  data10_genotype_per_sample_parallel <-
    foreach (item = 1:length(samples), .verbose = T) %dopar% {
      data_10[data_10$SAMPLE == samples[item], "GENOTYPE"]
    }
  # Stop parallel cluster
  stopCluster(parallel_cluster)
  remove(parallel_cluster)
}

# Genotype per sample merged Dataframe
if(!exists("data10_sample_genotype_df")) {
  aux_ouput_df_pwd <- "/home/luis3m/DATA/data10_sample_genotype_df.txt"
  if(file.exists(aux_ouput_df_pwd)) {
    data10_sample_genotype_df <- read_tsv(file = aux_ouput_df_pwd, col_names = F, progress = T, col_types = cols())
    colnames(data10_sample_genotype_df) <- c("SAMPLES",unlist(unique(data_10["SNP"]), use.names = F))
    remove(aux_ouput_df_pwd)
  } else {
    cat("", file = aux_ouput_df_pwd)
    for (index in 1:length(data10_genotype_per_sample_parallel)) {
      aux_ouput_df <- data.frame(t(matrix(unlist(data10_genotype_per_sample_parallel[[index]]))))
      aux_ouput_df <- cbind(samples[index], aux_ouput_df)
      write_tsv(aux_ouput_df, path = aux_ouput_df_pwd, append = T, col_names = F)
    }
    # RUN categoricalTodiscrete.sh SCRIPT!
    beep(sound = 6)
    cat("DATA 10 - RUN categoricalTodiscrete.sh SCRIPT!\n")
    cat("Sleeping...\n")
    Sys.sleep(time = 297)
    beep(sound = 3)
    Sys.sleep(time = 3)
    cat("I'm back!\n")
    data10_sample_genotype_df <- read_tsv(file = aux_ouput_df_pwd, col_names = F, progress = T, col_types = cols())
    colnames(data10_sample_genotype_df) <- c("SAMPLES",unlist(unique(data_10["SNP"]), use.names = F))
    remove(aux_ouput_df)
    remove(aux_ouput_df_pwd)
  }
}
```

Figure A.4: Retrieve SNP's genotypes per sample with parallel processing IV.

```
########################################
# Chromossome: 12                      #
# Total SNPs: 24594                    #
########################################

# Get Genotype per sample
if (!exists("data12_genotype_per_sample_parallel")) {
  # Set parallel cluster
  parallel_cluster <- makeCluster(detectCores()-1)
  registerDoParallel(parallel_cluster)
  data12_genotype_per_sample_parallel <-
    foreach (item = 1:length(samples), .verbose = T) %dopar% {
      data_12[data_12$SAMPLE == samples[item], "GENOTYPE"]
    }
  # Stop parallel cluster
  stopCluster(parallel_cluster)
  remove(parallel_cluster)
}

# Genotype per sample merged Dataframe
if(!exists("data12_sample_genotype_df")) {
  aux_ouput_df_pwd <- "/home/luis3m/DATA/data12_sample_genotype_df.txt"
  if(file.exists(aux_ouput_df_pwd)) {
    data12_sample_genotype_df <- read_tsv(file = aux_ouput_df_pwd, col_names = F, progress = T, col_types = cols())
    colnames(data12_sample_genotype_df) <- c("SAMPLES",unlist(unique(data_12["SNP"]), use.names = F))
    remove(aux_ouput_df_pwd)
  } else {
    cat("", file = aux_ouput_df_pwd)
    for (index in 1:length(data12_genotype_per_sample_parallel)) {
      aux_ouput_df <- data.frame(t(matrix(unlist(data12_genotype_per_sample_parallel[[index]]))))
      aux_ouput_df <- cbind(samples[index], aux_ouput_df)
      write_tsv(aux_ouput_df, path = aux_ouput_df_pwd, append = T, col_names = F)
    }
    # RUN categoricalTodiscrete.sh SCRIPT!
    beep(sound = 6)
    cat("DATA 12 - RUN categoricalTodiscrete.sh SCRIPT!\n")
    cat("Sleeping...\n")
    Sys.sleep(time = 297)
    beep(sound = 3)
    Sys.sleep(time = 3)
    cat("I'm back!\n")
    data12_sample_genotype_df <- read_tsv(file = aux_ouput_df_pwd, col_names = F, progress = T, col_types = cols())
    colnames(data12_sample_genotype_df) <- c("SAMPLES",unlist(unique(data_12["SNP"]), use.names = F))
    remove(aux_ouput_df)
    remove(aux_ouput_df_pwd)
  }
}
```

Figure A.5: Retrieve SNP's genotypes per sample with parallel processing V.

```
#######################################
# Chromossome: 22                     #
# Total SNPs: 6207                    #
#######################################

# Get Genotype per sample
if (!exists("data22_genotype_per_sample_parallel")) {
  # Set parallel cluster
  parallel_cluster <- makeCluster(detectCores()-1)
  registerDoParallel(parallel_cluster)
  data22_genotype_per_sample_parallel <-
    foreach (item = 1:length(samples), .verbose = T) %dopar% {
      data_22[data_22$SAMPLE == samples[item], "GENOTYPE"]
    }
  # Stop parallel cluster
  stopCluster(parallel_cluster)
  remove(parallel_cluster)
}

# Genotype per sample merged Dataframe
if(!exists("data22_sample_genotype_df")) {
  aux_ouput_df_pwd <- "/home/luis3m/DATA/data22_sample_genotype_df.txt"
  if(file.exists(aux_ouput_df_pwd)) {
    data22_sample_genotype_df <- read_tsv(file = aux_ouput_df_pwd, col_names = F, progress = T, col_types = cols())
    colnames(data22_sample_genotype_df) <- c("SAMPLES",unlist(unique(data_22["SNP"]), use.names = F))
    remove(aux_ouput_df_pwd)
  } else {
    cat("", file = aux_ouput_df_pwd)
    for (index in 1:length(data22_genotype_per_sample_parallel)) {
      aux_ouput_df <- data.frame(t(matrix(unlist(data22_genotype_per_sample_parallel[[index]]))))
      aux_ouput_df <- cbind(samples[index], aux_ouput_df)
      write_tsv(aux_ouput_df, path = aux_ouput_df_pwd, append = T, col_names = F)
    }
    # RUN categoricalTodiscrete.sh SCRIPT!
    beep(sound = 6)
    cat("DATA 22 - RUN categoricalTodiscrete.sh SCRIPT!\n")
    cat("Sleeping...\n")
    Sys.sleep(time = 297)
    beep(sound = 3)
    Sys.sleep(time = 3)
    cat("I'm back!\n")
    data22_sample_genotype_df <- read_tsv(file = aux_ouput_df_pwd, col_names = F, progress = T, col_types = cols())
    colnames(data22_sample_genotype_df) <- c("SAMPLES",unlist(unique(data_22["SNP"]), use.names = F))
    remove(aux_ouput_df)
    remove(aux_ouput_df_pwd)
  }
}
```

Figure A.6: Retrieve SNP's genotypes per sample with parallel processing VI.

```
#!/bin/bash
#
#
# Luis Moreira 2018-08-17

# Data Genotypes
# "AA" "AC" "AG" "AT" "CC" "CG" "CT" "GG" "GT" "TT"

if [ $# -eq 0 ]; then
        echo -e "[ - ] \t No args supplied! Exiting..."
        exit 1
fi

file=$(echo $1 | awk -F'_' '{print $1}')
if [ ${file} = "data01" ]; then
        echo -e "[ ? ] \t Data 1 Going from categorical to discrete data" &&
        (gsed -E 's/\bAA\b/1/g' $1 > temp.txt && mv temp.txt $1) &&
        (gsed -E 's/\bAC\b/2/g' $1 > temp.txt && mv temp.txt $1) &&
        (gsed -E 's/\bAG\b/3/g' $1 > temp.txt && mv temp.txt $1) &&
        (gsed -E 's/\bAT\b/4/g' $1 > temp.txt && mv temp.txt $1) &&
        (gsed -E 's/\bCC\b/5/g' $1 > temp.txt && mv temp.txt $1) &&
        (gsed -E 's/\bCG\b/6/g' $1 > temp.txt && mv temp.txt $1) &&
        (gsed -E 's/\bCT\b/7/g' $1 > temp.txt && mv temp.txt $1) &&
        (gsed -E 's/\bGG\b/8/g' $1 > temp.txt && mv temp.txt $1) &&
        (gsed -E 's/\bGT\b/9/g' $1 > temp.txt && mv temp.txt $1) &&
        (gsed -E 's/\bTT\b/10/g' $1 > temp.txt && mv temp.txt $1)
        if [ $? -eq 0 ]; then
                echo -e "[ + ] \t Data 1 Going from categorical to discrete data"
        else
                echo -e "[ - ] \t Something went wrong! Exiting..." &&
                exit 1
        fi
```

Figure A.7: Convert categorical data into discrete data.

```
#!/bin/bash
#
#
# Luis Moreira 2018-08-17

data1='/home/luis3m/DATA/data01_sample_genotype_df.txt'
data3='/home/luis3m/DATA/data03_sample_genotype_df.txt'
data4='/home/luis3m/DATA/data04_sample_genotype_df.txt'
data8='/home/luis3m/DATA/data08_sample_genotype_df.txt'
data10='/home/luis3m/DATA/data10_sample_genotype_df.txt'
data12='/home/luis3m/DATA/data12_sample_genotype_df.txt'
data22='/home/luis3m/DATA/data22_sample_genotype_df.txt'
aux1='/home/luis3m/DATA/aux1.txt'
aux2='/home/luis3m/DATA/aux2.txt'
aux3='/home/luis3m/DATA/aux3.txt'
aux4='/home/luis3m/DATA/aux4.txt'
aux5='/home/luis3m/DATA/aux5.txt'
aux6='/home/luis3m/DATA/aux6.txt'
mergedData='/home/luis3m/DATA/merged_sample_genotype_df.txt'

echo -e "[ ? ] \t Merging data frames..." &&
paste -d"\t"  ${data1} ${data3} > ${aux1} &&
paste -d"\t"  ${aux1} ${data4} > ${aux2} &&
paste -d"\t"  ${aux2} ${data8} > ${aux3} &&
paste -d"\t"  ${aux3} ${data10} > ${aux4} &&
paste -d"\t"  ${aux4} ${data12} > ${aux5} &&
paste -d"\t"  ${aux5} ${data22} > ${aux6}

if [ $? -eq 0 ]; then
    echo -e "[ + ] \t Merging data frames"
else
    echo -e "[ - ] \t Something went wrong! Exiting..." &&
    exit 1
fi

echo -e "[ ? ] \t Moving aux file to main data file"
if [ ! -f ${mergedData} ]; then
    mv -v ${aux6} ${mergedData}
    echo -e "[ + ] \t Moving aux file to main data file"
else
    echo -e "[ - ] \t File already exists. Please check current data!"
    exit 1
fi

echo -e "[ ? ] \t Removing aux file" &&
rm -vf ${aux} &&
rm -vf ${aux1} &&
rm -vf ${aux2} &&
rm -vf ${aux3} &&
rm -vf ${aux4} &&
rm -vf ${aux5} &&
rm -vf ${aux6} &&
echo -e "[ + ] \t Removing aux file"

exit 0
```

Figure A.8: Merge chromosome data files.

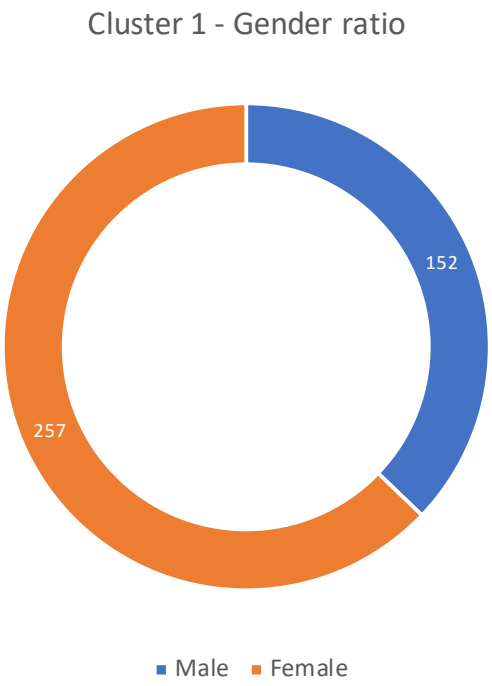


Figure A.9: Kmeans cluster 1 gender ratio I.

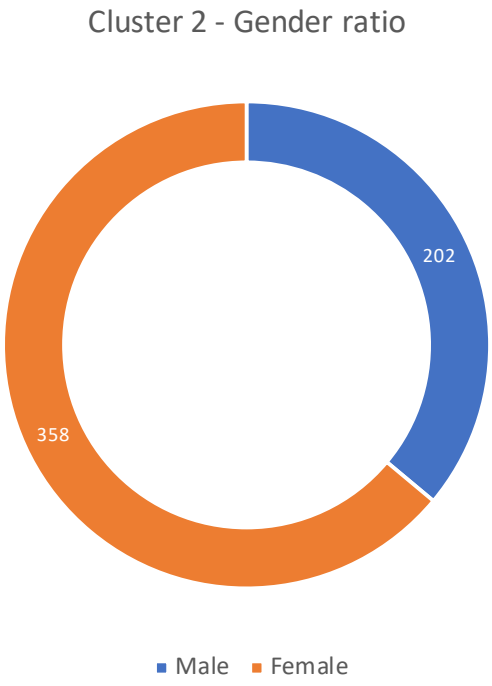


Figure A.10: Kmeans cluster 2 gender ratio II.

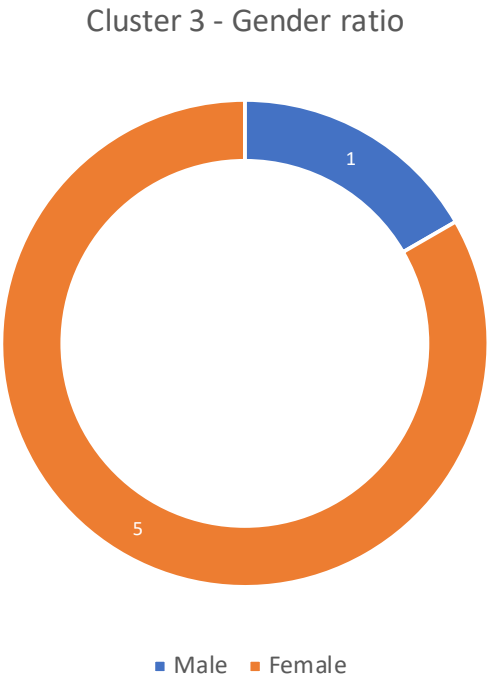


Figure A.11: Kmeans cluster 3 gender ratio III.

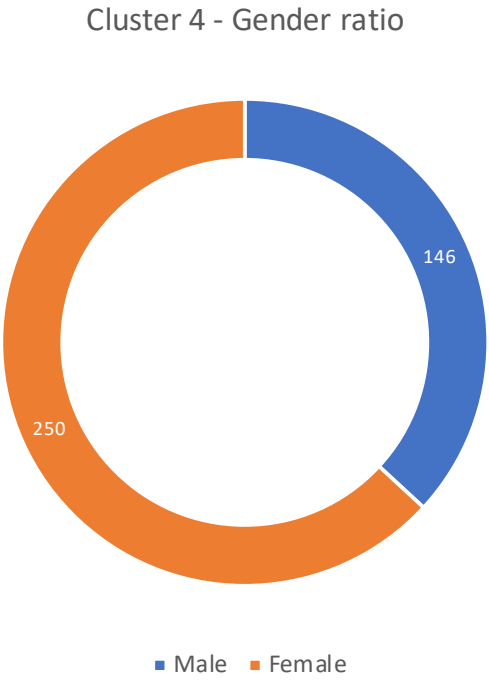


Figure A.12: Kmeans cluster 4 gender ratio IV.

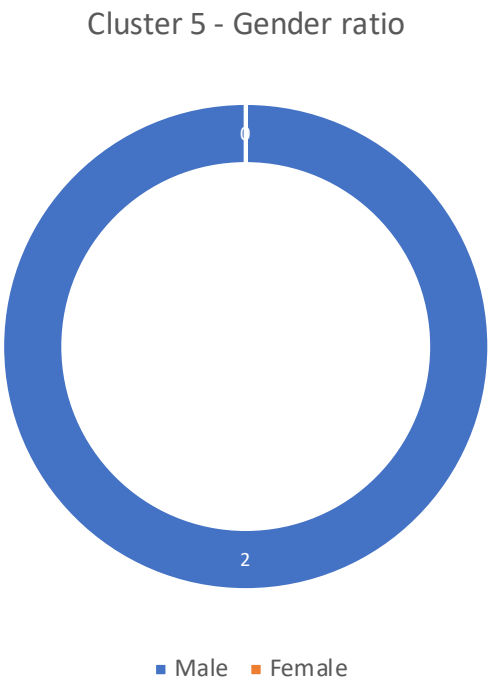


Figure A.13: Kmeans cluster 5 gender ratio V.

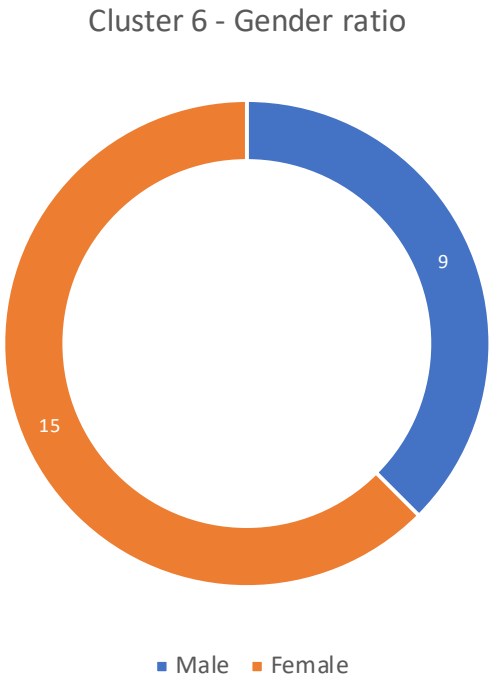


Figure A.14: Kmeans cluster 6 gender ratio VI.

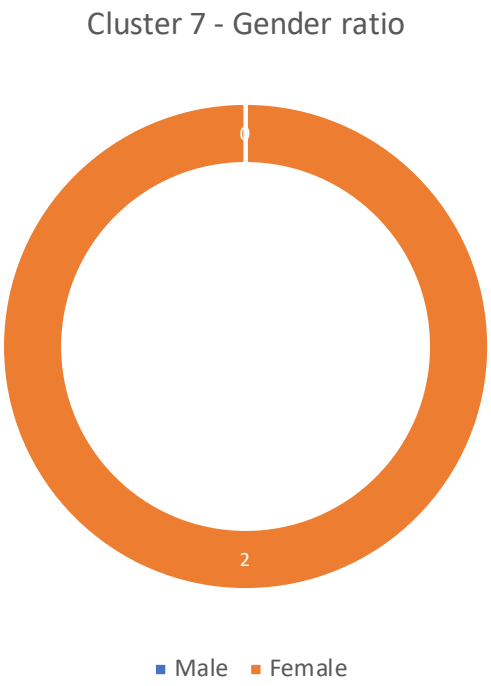


Figure A.15: Kmeans cluster 7 gender ratio VII.

```r
# CLUSTER 2
aux_samples_region <- c()
for (patient in cluster_2$SAMPLES) {
  aux_samples_region <- append(aux_samples_region,sampleBD[sampleBD$SAMPLE==patient,"REGION"])
}
cluster_2$REGION <- unlist(aux_samples_region, recursive = T, use.names = F)
remove(aux_samples_region)

aux_samples_age <- c()
for (patient in cluster_2$SAMPLES) {
  aux_samples_age <- append(aux_samples_age,sampleBD[sampleBD$SAMPLE==patient,"AGE_RECRUITMENT"])
}
cluster_2$AGE_RECRUITMENT <- unlist(aux_samples_age, recursive = T, use.names = F)
remove(aux_samples_age)

aux_samples_gender <- c()
for (patient in cluster_2$SAMPLES) {
  aux_samples_gender <- append(aux_samples_gender,sampleBD[sampleBD$SAMPLE==patient,"GENDER"])
}
cluster_2$GENDER <- unlist(aux_samples_gender, recursive = T, use.names = F)
remove(aux_samples_gender)
```

Figure A.16: R code used to retrieve samples clinical data of each cluster I.

```r
# CLUSTER 3
aux_samples_region <- c()
for (patient in cluster_3$SAMPLES) {
  aux_samples_region <- append(aux_samples_region,sampleBD[sampleBD$SAMPLE==patient,"REGION"])
}
cluster_3$REGION <- unlist(aux_samples_region, recursive = T, use.names = F)
remove(aux_samples_region)

aux_samples_age <- c()
for (patient in cluster_3$SAMPLES) {
  aux_samples_age <- append(aux_samples_age,sampleBD[sampleBD$SAMPLE==patient,"AGE_RECRUITMENT"])
}
cluster_3$AGE_RECRUITMENT <- unlist(aux_samples_age, recursive = T, use.names = F)
remove(aux_samples_age)

aux_samples_gender <- c()
for (patient in cluster_3$SAMPLES) {
  aux_samples_gender <- append(aux_samples_gender,sampleBD[sampleBD$SAMPLE==patient,"GENDER"])
}
cluster_3$GENDER <- unlist(aux_samples_gender, recursive = T, use.names = F)
remove(aux_samples_gender)
```

Figure A.17: R code used to retrieve samples clinical data of each cluster II.

```r
# CLUSTER 4
aux_samples_region <- c()
for (patient in cluster_4$SAMPLES) {
  aux_samples_region <- append(aux_samples_region,sampleBD[sampleBD$SAMPLE==patient,"REGION"])
}
cluster_4$REGION <- unlist(aux_samples_region, recursive = T, use.names = F)
remove(aux_samples_region)

aux_samples_age <- c()
for (patient in cluster_4$SAMPLES) {
  aux_samples_age <- append(aux_samples_age,sampleBD[sampleBD$SAMPLE==patient,"AGE_RECRUITMENT"])
}
cluster_4$AGE_RECRUITMENT <- unlist(aux_samples_age, recursive = T, use.names = F)
remove(aux_samples_age)

aux_samples_gender <- c()
for (patient in cluster_4$SAMPLES) {
  aux_samples_gender <- append(aux_samples_gender,sampleBD[sampleBD$SAMPLE==patient,"GENDER"])
}
cluster_4$GENDER <- unlist(aux_samples_gender, recursive = T, use.names = F)
remove(aux_samples_gender)
```

Figure A.18: R code used to retrieve samples clinical data of each cluster III.

```r
# CLUSTER 5
aux_samples_region <- c()
for (patient in cluster_5$SAMPLES) {
  aux_samples_region <- append(aux_samples_region,sampleBD[sampleBD$SAMPLE==patient,"REGION"])
}
cluster_5$REGION <- unlist(aux_samples_region, recursive = T, use.names = F)
remove(aux_samples_region)

aux_samples_age <- c()
for (patient in cluster_5$SAMPLES) {
  aux_samples_age <- append(aux_samples_age,sampleBD[sampleBD$SAMPLE==patient,"AGE_RECRUITMENT"])
}
cluster_5$AGE_RECRUITMENT <- unlist(aux_samples_age, recursive = T, use.names = F)
remove(aux_samples_age)

aux_samples_gender <- c()
for (patient in cluster_5$SAMPLES) {
  aux_samples_gender <- append(aux_samples_gender,sampleBD[sampleBD$SAMPLE==patient,"GENDER"])
}
cluster_5$GENDER <- unlist(aux_samples_gender, recursive = T, use.names = F)
remove(aux_samples_gender)
```

Figure A.19: R code used to retrieve samples clinical data of each cluster IV.

```r
# CLUSTER 6
aux_samples_region <- c()
for (patient in cluster_6$SAMPLES) {
  aux_samples_region <- append(aux_samples_region,sampleBD[sampleBD$SAMPLE==patient,"REGION"])
}
cluster_6$REGION <- unlist(aux_samples_region, recursive = T, use.names = F)
remove(aux_samples_region)

aux_samples_age <- c()
for (patient in cluster_6$SAMPLES) {
  aux_samples_age <- append(aux_samples_age,sampleBD[sampleBD$SAMPLE==patient,"AGE_RECRUITMENT"])
}
cluster_6$AGE_RECRUITMENT <- unlist(aux_samples_age, recursive = T, use.names = F)
remove(aux_samples_age)

aux_samples_gender <- c()
for (patient in cluster_6$SAMPLES) {
  aux_samples_gender <- append(aux_samples_gender,sampleBD[sampleBD$SAMPLE==patient,"GENDER"])
}
cluster_6$GENDER <- unlist(aux_samples_gender, recursive = T, use.names = F)
remove(aux_samples_gender)
```

Figure A.20: R code used to retrieve samples clinical data of each cluster V.

```r
# CLUSTER 7
aux_samples_region <- c()
for (patient in cluster_7$SAMPLES) {
  aux_samples_region <- append(aux_samples_region,sampleBD[sampleBD$SAMPLE==patient,"REGION"])
}
cluster_7$REGION <- unlist(aux_samples_region, recursive = T, use.names = F)
remove(aux_samples_region)

aux_samples_age <- c()
for (patient in cluster_7$SAMPLES) {
  aux_samples_age <- append(aux_samples_age,sampleBD[sampleBD$SAMPLE==patient,"AGE_RECRUITMENT"])
}
cluster_7$AGE_RECRUITMENT <- unlist(aux_samples_age, recursive = T, use.names = F)
remove(aux_samples_age)

aux_samples_gender <- c()
for (patient in cluster_7$SAMPLES) {
  aux_samples_gender <- append(aux_samples_gender,sampleBD[sampleBD$SAMPLE==patient,"GENDER"])
}
cluster_7$GENDER <- unlist(aux_samples_gender, recursive = T, use.names = F)
remove(aux_samples_gender)
```

Figure A.21: R code used to retrieve samples clinical data of each cluster VI.

```
cluster_1_genotype <-
  cluster_2_genotype <-
  cluster_3_genotype <-
  cluster_4_genotype <-
  cluster_5_genotype <-
  cluster_6_genotype <-
  cluster_7_genotype <- c()
s_time <- Sys.time()
# CLUSTER 1
parallel_cluster <- makeCluster(detectCores()*0.5)
registerDoParallel(parallel_cluster)
cluster_1_genotype <- foreach (index = 1:length(cluster_1$SAMPLES), .verbose = T) %dopar% {
  append(cluster_1_genotype,data_03[data_03$SAMPLE==cluster_1$SAMPLES[index] & data_03$SNP=="rs1523041","GENOTYPE"])
}
stopCluster(parallel_cluster)
Sys.sleep(120)
s_time <- Sys.time()
# CLUSTER 2
parallel_cluster <- makeCluster(detectCores()*0.5)
registerDoParallel(parallel_cluster)
cluster_2_genotype <- foreach (index = 1:length(cluster_2$SAMPLES), .verbose = T) %dopar% {
  append(cluster_2_genotype,data_03[data_03$SAMPLE==cluster_2$SAMPLES[index] & data_03$SNP=="rs1523041","GENOTYPE"])
}
stopCluster(parallel_cluster)
Sys.sleep(120)
# CLUSTER 3
parallel_cluster <- makeCluster(detectCores()*0.5)
registerDoParallel(parallel_cluster)
cluster_3_genotype <- foreach (index = 1:length(cluster_3$SAMPLES), .verbose = T) %dopar% {
  append(cluster_3_genotype,data_03[data_03$SAMPLE==cluster_3$SAMPLES[index] & data_03$SNP=="rs1523041","GENOTYPE"])
}
stopCluster(parallel_cluster)
Sys.sleep(120)
# CLUSTER 4
parallel_cluster <- makeCluster(detectCores()*0.5)
registerDoParallel(parallel_cluster)
cluster_4_genotype <- foreach (index = 1:length(cluster_4$SAMPLES), .verbose = T) %dopar% {
  append(cluster_4_genotype,data_03[data_03$SAMPLE==cluster_4$SAMPLES[index] & data_03$SNP=="rs1523041","GENOTYPE"])
}
stopCluster(parallel_cluster)
Sys.sleep(120)
# CLUSTER 5
parallel_cluster <- makeCluster(detectCores()*0.5)
registerDoParallel(parallel_cluster)
cluster_5_genotype <- foreach (index = 1:length(cluster_5$SAMPLES), .verbose = T) %dopar% {
  append(cluster_5_genotype,data_03[data_03$SAMPLE==cluster_5$SAMPLES[index] & data_03$SNP=="rs1523041","GENOTYPE"])
}
stopCluster(parallel_cluster)
Sys.sleep(120)
# CLUSTER 6
parallel_cluster <- makeCluster(detectCores()*0.5)
registerDoParallel(parallel_cluster)
cluster_6_genotype <- foreach (index = 1:length(cluster_6$SAMPLES), .verbose = T) %dopar% {
  append(cluster_6_genotype,data_03[data_03$SAMPLE==cluster_6$SAMPLES[index] & data_03$SNP=="rs1523041","GENOTYPE"])
}
stopCluster(parallel_cluster)
Sys.sleep(120)
# CLUSTER 7
parallel_cluster <- makeCluster(detectCores()*0.5)
registerDoParallel(parallel_cluster)
cluster_7_genotype <- foreach (index = 1:length(cluster_7$SAMPLES), .verbose = T) %dopar% {
  append(cluster_7_genotype,data_03[data_03$SAMPLE==cluster_7$SAMPLES[index] & data_03$SNP=="rs1523041","GENOTYPE"])
}
stopCluster(parallel_cluster)
e_time <- Sys.time()
ex_time <- e_time-s_time
cluster_1$rs1523041 <- unlist(cluster_1_genotype)
cluster_2$rs1523041 <- unlist(cluster_2_genotype)
cluster_3$rs1523041 <- unlist(cluster_3_genotype)
cluster_4$rs1523041 <- unlist(cluster_4_genotype)
cluster_5$rs1523041 <- unlist(cluster_5_genotype)
cluster_6$rs1523041 <- unlist(cluster_6_genotype)
cluster_7$rs1523041 <- unlist(cluster_7_genotype)
```

Figure A.22: R code used to retrieve samples' genotype of SNP rs1523041.

References

[1] R. Agrawal, R. Srikant, et al. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499, 1994.

[2] I. P. Arribas, K. Saunders, G. Goodwin, and T. Lyons. A signature-based machine learning model for bipolar disorder and borderline personality disorder. *arXiv preprint arXiv:1707.07124*, 2017.

[3] American Psychiatric Association et al. *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub, 2013.

[4] A. Beheshti, M. Hashmi, H. Dong, and W. E. Zhang. *Service Research and Innovation*. Springer International Publishing, 2018.

[5] G. Blokland, C.-Y. Chen, J. Smoller, T. Petryshen, J. Goldstein, Schizophrenia Working Group Psychiatric Genomics Consortium, Major Depressive Disorder Working Group Psychiatric Genomics Consortium, Bipolar Disorder Working Group Psychiatric Genomics Consortium, et al. T226. genotype-by-sex interaction effects in the risk for schizophrenia, major depressive disorder, and bipolar disorder. *Biological Psychiatry*, 83(9):S216, 2018.

[6] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM, 1992. doi:10.1145/130385.130401.

[7] P. Brambilla and J.C. Soares. Journal of affective disorders. 226:1 – 360, January 2018. ISSN: 0165-0327. doi:https://doi.org/10.1016/S0165-0327(17)32118-3.

[8] P. R. Burton, D. G. Clayton, L. R. Cardon, N. Craddock, P. Deloukas, A. Duncanson, D. P. Kwiatkowski, M. I. McCarthy, W. H. Ouwehand, N. J. Samani, J. A. Todd, P. Donnelly, J. C. Barrett, P. R. Burton, D. Davison, P. Donnelly, D. Easton, D. Evans, H. T. Leung, J. L. Marchini, A. P. Morris, C. C. Spencer, M. D. Tobin, L. R. Cardon, D. G. Clayton, A. P. Attwood, J. P. Boorman, B. Cant, U. Everson, J. M. Hussey, J. D. Jolley, A. S. Knight, K. Koch, E. Meech, S. Nutland, C. V. Prowse, H. E. Stevens, N. C. Taylor, G. R. Walters, N. M. Walker, N. A. Watkins, T. Winzer, J. A. Todd, W. H. Ouwehand, R. W. Jones, W. L. McArdle, S. M. Ring, D. P. Strachan, M. Pembrey, G. Breen, D. St Clair, S. Caesar, K. Gordon-Smith, L. Jones, C. Fraser, E. K. Green, D. Grozeva, M. L. Hamshere,

P. A. Holmans, I. R. Jones, G. Kirov, V. Moskvina, I. Nikolov, M. C. O'Donovan, M. J. Owen, N. Craddock, D. A. Collier, A. Elkin, A. Farmer, R. Williamson, P. McGuffin, A. H. Young, I. N. Ferrier, S. G. Ball, A. J. Balmforth, J. H. Barrett, D. T. Bishop, M. M. Iles, A. Maqbool, N. Yuldasheva, A. S. Hall, P. S. Braund, P. R. Burton, R. J. Dixon, M. Mangino, S. Suzanne, M. D. Tobin, J. R. Thompson, N. J. Samani, F. Bredin, M. Tremelling, M. Parkes, H. Drummond, C. W. Lees, E. R. Nimmo, J. Satsangi, S. A. Fisher, A. Forbes, C. M. Lewis, C. M. Onnie, N. J. Prescott, J. Sanderson, C. G. Mathew, J. Barbour, M. K. Mohiuddin, C. E. Todhunter, J. C. Mansfield, T. Ahmad, F. R. Cummings, D. P. Jewell, J. Webster, M. J. Brown, D. G. Clayton, G. M. Lathrop, J. Connell, A. Dominczak, N. J. Samani, C. A. Marcano, B. Burke, R. Dobson, J. Gungadoo, K. L. Lee, P. B. Munroe, S. J. Newhouse, A. Onipinla, C. Wallace, M. Xue, M. Caulfield, M. Farrall, A. Barton, I. N. Bruce, H. Donovan, S. Eyre, P. D. Gilbert, S. L. Hider, A. M. Hinks, S. L. John, C. Potter, A. J. Silman, D. P. Symmmons, W. Thomson, J. Worthington, D. G. Clayton, D. B. Dunger, S. Nutland, H. E. Stevens, N. M. Walker, B. Widmer, J. A. Todd, T. A. Frayling, R. M. Freathy, H. Lango, J. R. Perry, B. M. Shields, M. N. Weedon, A. T. Hattersley, G. A. Hitman, M. Walker, K. S. Elliott, C. J. Groves, C. M. Lindgren, N. W. Rayner, N. J. Timpson, E. Zeggini, M. I. McCarthy, M. Newport, G. Sirugo, E. Lyons, F. Vannberg, A. V. Hill, L. A. Bradbury, C. Farrar, J. J. Pointon, P. Wordsworth, M. A. Brown, J. A. Franklyn, J. M. Heward, M. J. Simmonds, S. C. Gough, S. Seal, M. R. Stratton, N. Rahman, M. Ban, A. Goris, S. J. Sawcer, A. Compston, D. Conway, M. Jallow, M. Newport, G. Sirugo, K. A. Rockett, D. P. Kwiatowski, S. J. Bumpstead, A. Chaney, K. Downes, M. J. Ghori, R. Gwilliam, S. E. Hunt, M. Inouye, A. Keniry, E. King, R. McGinnis, S. Potter, R. Ravindrarajah, P. Whittaker, C. Widden, D. Withers, P. Deloukas, H. T. Leung, S. Nutland, H. E. Stevens, N. M. Walker, J. A. Todd, D. Easton, D. G. Clayton, P. R. Burton, M. D. Tobin, J. C. Barrett, D. Evans, A. P. Morris, L. R. Cardon, N. J. Cardin, D. Davison, T. Ferreira, J. Pereira-Gale, I. B. Hallgrimsdottir, B. N. Howie, J. L. Marchini, C. C. Spencer, Z. Su, Y. Y. Teo, D. Vukcevic, P. Donnelly, D. Bentley, M. A. Brown, L. R. Gordon, M. Caulfield, D. G. Clayton, A. Compston, N. Craddock, P. Deloukas, P. Donnelly, M. Farrall, S. C. Gough, A. S. Hall, A. T. Hattersley, A. V. Hill, D. P. Kwiatkowski, C. Mathew, M. I. McCarthy, W. H. Ouwehand, M. Parkes, M. Pembrey, N. Rahman, N. J. Samani, M. R. Stratton, J. A. Todd, and J. Worthington. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, Jun 2007.

[9] D. Bzdok and A. Meyer-Lindenberg. Machine learning for precision psychiatry: Opportunites and challenges. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 2017.

[10] S.-H. Chang, L. Gao, Z. Li, W.-N. Zhang, Y. Du, and J. Wang. Bdgene: a genetic database for bipolar disorder and its overlap with schizophrenia and major depressive disorder. *Biological psychiatry*, 74(10):727–733, 2013.

[11] S. H. Chang, L. Gao, Z. Li, W. N. Zhang, Y. Du, and J. Wang. BDgene: a genetic database for bipolar disorder and its overlap with schizophrenia and major depressive disorder. *Biol. Psychiatry*, 74(10):727–733, Nov 2013.

[12] M. Cloutier, M. Greene, A. Guerin, M. Touya, and E. Wu. The economic burden of bipolar i disorder in the united states in 2015. *Journal of Affective Disorders*, 226(Supplement C): 45 – 51, September 2017. ISSN: 0165-0327. doi:https://doi.org/10.1016/j.jad.2017.09.011.

[13] The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, Jun 2007. ISSN: 0028-0836. 17554300[pmid]. doi:10.1038/nature05911.

[14] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

[15] A. Diflorio and I. Jones. Is sex important? gender differences in bipolar disorder. *International Review of Psychiatry*, 22(5):437–452, 2010. doi:10.3109/09540261.2010.514601.

[16] G. L. Faedda, R. J Baldessarini, T. Suppes, L. Tondo, I. Becker, and D. S. Lipschitz. Pediatric-onset bipolar disorder: A neglected clinical and public health problem gianni. *Harvard Review of Psychiatry*, 3(4):171–195, 1995.

[17] M. A. Ferreira, M. C. O'Donovan, Y. A. Meng, I. R. Jones, D. M. Ruderfer, L. Jones, J. Fan, G. Kirov, R. H. Perlis, E. K. Green, J. W. Smoller, D. Grozeva, J. Stone, I. Nikolov, K. Chambert, M. L. Hamshere, V. L. Nimgaonkar, V. Moskvina, M. E. Thase, S. Caesar, G. S. Sachs, J. Franklin, K. Gordon-Smith, K. G. Ardlie, S. B. Gabriel, C. Fraser, B. Blumenstiel, M. Defelice, G. Breen, M. Gill, D. W. Morris, A. Elkin, W. J. Muir, K. A. McGhee, R. Williamson, D. J. MacIntyre, A. W. MacLean, C. D. St, M. Robinson, M. Van Beck, A. C. Pereira, R. Kandaswamy, A. McQuillin, D. A. Collier, N. J. Bass, A. H. Young, J. Lawrence, I. N. Ferrier, A. Anjorin, A. Farmer, D. Curtis, E. M. Scolnick, P. McGuffin, M. J. Daly, A. P. Corvin, P. A. Holmans, D. H. Blackwood, H. M. Gurling, M. J. Owen, S. M. Purcell, P. Sklar, and N. Craddock. Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. *Nat. Genet.*, 40(9):1056–1058, Sep 2008.

[18] A. Fiorentino, N. L. O'brien, D. P. Locke, A. McQuillin, A. Jarram, A. Anjorin, R. Kandaswamy, D. Curtis, R. A. Blizard, and H. M. D. Gurling. Analysis of ank3 and cacna1c variants identified in bipolar disorder whole genome sequence data. *Bipolar Disorders*, 16(6):583–591, 2014. doi:10.1111/bdi.12203.

[19] New York-Mid-Atlantic Consortium for Genetic, Newborn Screening Services, et al. *Understanding genetics: a New York, mid-Atlantic guide for patients and health professionals*. Lulu. com, 2009.

[20] G. M. Goodwin, P. M. Haddad, I. N. Ferrier, J. K. Aronson, T. R. H. Barnes, A. Cipriani, D. R. Coghill, S. Fazel, J. R. Geddes, H. Grunze, E. A. Holmes, O. Howes, S. Hudson, N. Hunt, I. Jones, I. C. Macmillan, H. McAllister-Williams, D. M. Miklowitz, R. Morriss, M. Munafò, C. Paton, B. J. Saharkian, K. E. A. Saunders, J. M. A. Sinclair, D. Taylor, E. Vieta, and A. H. Young. Evidence-based guidelines for treating bipolar disorder: revised third edition

recommendations from the british association for psychopharmacology. *J Psychopharmacol*, 30(6):495–553, Jun 2016. ISSN: 0269-8811. 26979387[pmid]. doi:10.1177/0269881116636545.

[21] H. Grunze, E. Vieta, G. M. Goodwin, C. Bowden, R. W. Licht, J.-M. Azorin, L. Yatham, S. Mosolov, H.-J. Möller, S. Kasper, et al. The world federation of societies of biological psychiatry (wfsbp) guidelines for the biological treatment of bipolar disorders: Acute and long-term treatment of mixed states in bipolar disorder. *The World Journal of Biological Psychiatry*, 19(1):2–58, 2018.

[22] R. E. Hales, S. C. Yudofsky, G. O. Gabbard, and American Psychiatric Publishing. *The American Psychiatric Publishing textbook of psychiatry*. American Psychiatric Pub, 2008.

[23] J. Han, M. Kamber, and J. Pei. *Data Mining: Concepts and Techniques*. Elsevier Inc., 2012.

[24] M. Hossain et al. *A Study on Behavioral and Biological Risk Factors Determination of Non-Communicable Diseases at Narayanganj*. PhD thesis, East West University, 2017.

[25] J. Hou, L. Acharya, D. Zhu, and J. Cheng. An overview of bioinformatics methods for modeling biological pathways in yeast. *Brief Funct Genomics*, 15(2):95–108, Mar 2016.

[26] L. V. Kessing, K. Munkholm, M. Faurholt-Jepsen, K. W. Miskowiak, L. B. Nielsen, R. Frikke-Schmidt, C. Ekstrøm, O. Winther, B. K. Pedersen, H. E. Poulsen, R. S. McIntyre, F. Kapczinski, W. F. Gattaz, J. Bardram, M. Frost, O. Mayora, G. M. Knudsen, M. Phillips, and M. Vinberg. The bipolar illness onset study: research protocol for the bio cohort study. Jun 2017. doi:10.1136/bmjopen-2016-015462.

[27] M. Khalid, T. M. Driessen, J. S. Lee, L. Tejwani, A. Rasool, M. Saqlain, P. A. Shiaq, M. Hanif, A. Nawaz, A. T. DeWan, et al. Association of cacna1c with bipolar disorder among the pakistani population. *Gene*, 664:119–126, 2018.

[28] S. H. Kim, S. Y. Shin, K. Y. Lee, E. J. Joo, J. Y. Song, Y. M. Ahn, Y. H. Lee, and Y. S. Kim. The genetic association of dusp6 with bipolar disorder and its effect on erk activity. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 37(1):41–49, 2012.

[29] R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI*, 1995.

[30] L. Kruglyak and D. A. Nickerson. Variation is the spice of life. *Nature genetics*, 27(3):234, 2001.

[31] C. N. Kuswanto, M. Y. Sum, C. R. Z. Thng, Y. B. Zhang, G. L. Yang, W. L. Nowinski, Y. Y. Sitoh, C. M. Low, and K. Sim. Grin2b gene and associated brain cortical white matter changes in bipolar disorder: a preliminary combined platform investigation. *BioMed research international*, 2013, 2013.

[32] B. Labonté, O. Engmann, I. Purushothaman, C. Menard, J. Wang, C. Tan, J. R. Scarpa, G. Moy, Y.-H. E. Loh, M. Cahill, Z. S. Lorsch, P. J. Hamilton, E. S. Calipari, G. E. Hodes,

O. Issler, H. Kronman, M. Pfau, A. L. J. Obradovic, Y. Dong, R. L. Neve, S. Russo, A. Kasarskis, C. Tamminga, N. Mechawar, G. Turecki, B. Zhang, Li Shen, and E. J. Nestler. Sex-specific transcriptional signatures in human depression. *Nature Medicine*, 23:1102, Aug 2017.

[33] Y. Lee, R.-M. Ragguett, R. B. Mansur, J. J. Boutilier, J. D. Rosenblat, A. Trevizol, E. Brietzke, K. Lin, Z. Pan, M. Subramaniapillai, T. C. Y. Chan, D. Fus, C. Park, N. Musial, H. Zuckerman, V. C.-H. Chen, R. Ho, C. Rong, and R. S. McIntyre. Applications of machine learning algorithms to predict therapeutic outcomes in depression: A meta-analysis and systematic review. *Journal of Affective Disorders*, 241:519–532, Dec 2018. ISSN: 0165-0327. doi:10.1016/j.jad.2018.08.073.

[34] E. Lin, P.-H. Kuo, Y.-L. Liu, Y. W. Y. Yu, A. Yang, and S.-J. Tsai. A deep learning approach for predicting antidepressant response in major depression using clinical and genetic biomarkers. *Frontiers in psychiatry*, 9:290, 2018.

[35] A. Lyons-Warren, J.J. Chang, R. Balkissoon, A. Kamiya, M. Garant, J. Nurnberger, W. Scheftner, T. Reich, F. McMahon, J. Kelsoe, et al. Evidence of association between bipolar disorder and citron on chromosome 12q24. *Molecular psychiatry*, 10(9):807, 2005.

[36] O. Manor and E. Segal. Predicting disease risk using bootstrap ranking and classification algorithms. *PLoS computational biology*, 9(8):e1003200, 2013.

[37] J. A. McCain. Antidepressants and suicide in adolescents and adults: A public health experiment with unintended consequences? *P T*, 34(7):355–378, Jul 2009. ISSN: 1052-1372.

[38] E. Mellerup, M. B. Jørgensen, H. Dam, and G. L. Møller. Combinations of snp genotypes from the wellcome trust case control study of bipolar patients. *Acta Neuropsychiatrica*, 30 (2):106–110, 2018. doi:10.1017/neu.2017.36.

[39] R. Mihaescu. *Genetic risk prediction for common diseases: methodology and applications*. 2013.

[40] N. Morisada, T. Ioroi, M. Taniguchi-Ikeda, M. J. Ye, N. Okamoto, T. Yamamoto, and K. Iijima. A 12p13 grin2b deletion is associated with developmental delay and macrocephaly. *Human genome variation*, 3:16029, 2016.

[41] United Nations Department of Economic and Social Affairs/Population Division. *World Population Prospects: The 2017 Revision, Key Findings and Advance Tables*. United Nations, 2017.

[42] National Institute of Mental Health. *Depression: What You Need to Know*. page 32, November 2015. ISSN: NIH Publication No. 15-3561.

[43] G. V. D. Oliveira and M. C. Naldi. Scalable fast evolutionary k-means clustering. In *2015 Brazilian Conference on Intelligent Systems (BRACIS)*, pages 74–79, Nov 2015. doi:10.1109/BRACIS.2015.20.

[44] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999. Previous number = SIDL-WP-1999-0120.

[45] H. Ren, L. Guan, L. Zhao, Y. Lin, Y. Wang, Z. Yang, X. Li, X. Ma, X. Cheng, W. Deng, et al. Contribution of genes in the gabaergic pathway to bipolar disorder and its executive function deficit in the chinese han population. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 177(1):50–67, 2018.

[46] G. S. Sachs, A. A. Nierenberg, J. R. Calabrese, L. B. Marangell, S. R. Wisniewski, L. Gyulai, E. S. Friedman, C. L. Bowden, M. D. Fossey, M. J. Ostacher, T. A. Ketter, J. Patel, P. Hauser, D. Rapport, J. M. Martinez, M. H. Allen, D. J. Miklowitz, M. W. Otto, E. B. Dennehy, and M. E. Thase. Effectiveness of adjunctive antidepressant treatment for bipolar depression. *N. Engl. J. Med.*, 356(17):1711–1722, Apr 2007.

[47] R. Salvini, R. S. Dias, B. Lafer, and I. Dutra. A multi-relational model for depression relapse in patients with bipolar disorder. *Mednfo*, 216:741–745, 2015. doi:10.3233/978-1-61499-564-7-741.

[48] H. G. Schnack. Improving individual predictions: Machine learning approaches for detecting and attacking heterogeneity in schizophrenia (and other psychiatric diseases). *Schizophrenia research*, 2017.

[49] M. L. Seney, Z. Huo, K. Cahill, L. French, R. Puralewski, J. Zhang, R. W. Logan, G. Tseng, D. A. Lewis, and E. Sibille. Opposite molecular signatures of depression in men and women. *Biological psychiatry*, 2018.

[50] N. Shah, S. Grover, and G. P. Rao. Clinical practice guidelines for management of bipolar disorder. *Indian J Psychiatry*, 59(Suppl 1):S51–S66, Jan 2017. ISSN: 0019-5545. IJPsy-59-51[PII]. doi:10.4103/0019-5545.196974.

[51] S. Shi and H. R. Ueda. Ca2+-dependent hyperpolarization pathways in sleep homeostasis and mental disorders. *BioEssays*, 40(1):1700105, 2018.

[52] I. H. Shim, Y. S. Woo, H.-R. Wang, and W.-M. Bahk. Predictors of a shorter time to hospitalization in patients with bipolar disorder: Medication during the acute and maintenance phases and other clinical factors. *Clin Psychopharmacol Neurosci*, 15(3): 248–255, Aug 2017. ISSN: 1738-1088. cpn-15-248[PII]. doi:10.9758/cpn.2017.15.3.248.

[53] T. Singh and M. Rajput. Misdiagnosis of bipolar disorder. *Psychiatry (Edgmont)*, 3(10): 57–63, Oct 2006. ISSN: 1550-5952. 20877548[pmid].

[54] D. Sit. Women and bipolar disorder across the life span. *J Am Med Womens Assoc (1972)*, 59(2):91–100, 2004.

[55] D. J. Smith and N. Craddock. Unipolar and bipolar depression: different or the same? *British Journal of Psychiatry*, 199(4), 2011. doi:10.1192/bjp.bp.111.092726.

[56] F. Wang, A. M. McIntosh, Y. He, J. Gelernter, and H. P. Blumberg. The association of genetic variation in CACNA1C with structure and function of a frontotemporal system. *Bipolar Disord*, 13(7-8):696–700, 2011.

[57] G. Welchman. From polish bomba to british bombe: The birth of ultra. *Intelligence and National Security*, 1(1):71–110, 1986.

[58] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, S. Y. Philip, Z. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg. Top 10 algorithms in data mining. *Knowledge and information systems*, 14(1):1–37, 2008. doi:10.1007/s10115-007-0114-2.

[59] L. Zhang, X. Yu, Y.-R. Fang, G. S. Ungvari, C. H. Ng, H. F. K. Chiu, H.-C. Li, H.-C. Yang, Q.-R. Tan, X.-F. Xu, G. Wang, and Y.-T. Xiang. Duration of untreated bipolar disorder: a multicenter study. *Sci Rep*, 7:44811, Mar 2017. ISSN: 2045-2322. 28327583[pmid]. doi:10.1038/srep44811.

[60] Q. Zhao, R. Che, Z. Zhang, P. Wang, J. Li, Y. Li, K. Huang, W. Tang, G. Feng, K. Lindpaintner, et al. Positive association between grin2b gene and bipolar disorder in the chinese han population. *Psychiatry research*, 185(1-2):290–292, 2011.