Phenotype Prediction Using Genotype Information

Salteris Georgios

Moulopoulos Antonios

Diploma Thesis

Department of Computer Science & Engineering

University of Ioannina

Advisor: P. Tsaparas

Ioannina, July 2018

ΤΜΗΜΑ ΜΗΧ. Η/Υ & ΠΛΗΡΟΦΟΡΙΚΗΣ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
UNIVERSITY OF IOANNINA

Acknowledgment

We would first like to thank our thesis advisor Associate Professor Panayiotis Tsaparas of the Department of Computer Science & Engineering at the University of Ioannina. The door to Prof. Tsaparas office was always open, whenever we ran into a trouble spot or had a question about our research or writing. He consistently allowed this paper to be our work but steered us in the right direction, whenever he thought we needed it.

We also like to thank Professor Evangelou, who helped us by providing the Data, as well as the validation for this research. Without his passionate participation and input, the validation of our research could not have been successfully conducted.

Finally, we must express our profound gratitude to our parents for providing us with unfailing support and continuous encouragement throughout our years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.

Abstract

Cancer Melanoma is a type of cancer that develops in the skin and rarely occurs in the mouth. Different genetic variations and environmental reasons are responsible for its appearance. The goal of this thesis is to study whether it is possible to predict if someone will develop cancer melanoma (phenotype), by machine learning algorithms using the genotype information of the individual, in the form of SNPs.

We constructed a pipeline that collects and processes the data, to produce the features and performs feature selection by identifying unique SNPs. We considered a variety of classification algorithms for the prediction task, and we studied the effect of the different parameters. To scale our approach to big data, we also implemented our pipeline in Apache Spark. We studied manually the extracted SNP's, and we compared the validity of our results with existing studies.

We hope this to be a first step towards the use of machine learning for detecting cancer melanoma at its earliest stages.

Table of contents

Chapter 1. Introduction

The grand goal of our thesis is to predict if someone is predisposed to develop a disease in the future based on their genotype (gene) information. Our goal is to make an assisting tool in order to help the doctors extract useful results for diagnosing whether a patient has or not a disease, in our case, cancer melanoma. Today, Machine Learning tools are used to assist doctors in evaluating their results, and there are several prototypes for spotting breast cancer in x-rays, not visible to doctors.

However, doctors and other data scientists are already attempting, using several techniques, to discover genes that are responsible for various diseases [1]. Furthermore, related projects to ours, seek to predict future diseases. The purpose of achieving something like that, has a vast benefit for our society. If we can predict that an individual has a significant probability of developing cancer, for example, then the whole cure procedure would change dramatically, and hundreds of lives could turn to the best (It is easier for someone to avoid or heal from an illness, if he knows that he is going to be infected by it in the future).

So, besides the prediction part of our thesis, we try also to find those genes that are causing the disease.

In the future, we imagine an all in one tool that, with the given data, will predict several diseases, but also an algorithm with the ability to quarantine the causing DNA snippets.

1.1 Thesis Contribution

We received from the Medical Department of the University of Ioannina our data as well as a phenotype for patients with Melanoma. Cancer Melanoma is a type of cancer that develops in the skin and rarely occurs in the mouth. Different genetic variations and environmental reasons are responsible for its appearance.

We constructed a pipeline that collects and processes the data to produce the features and performs feature selection by identifying unique SNPs. We considered a variety of classification algorithms for the prediction task, and we study here the effect of the different parameters. To scale our approach to big data, we also implemented our pipeline in Apache Spark.

Moreover, we balance our data with different proportion of patients and healthy people, as well as different Θ thresholds in our correlation techniques to keep only highly correlated features.

Then, using classification algorithms, we try to predict if someone is predisposed to develop cancer melanoma in the feature and with various metrics, we evaluate our results

We study manually the extracted SNP's from our top features, and we compare the validity of our results with existing studies [1]. We hope this will be a first step towards the use of machine learning for detecting cancer melanoma at its earliest stages. Also, an assisting tool for doctors in their studies to find more SNP's.

1.2 Outline

The contents of the Thesis are organized as follows.

Chapter 2: In this chapter, we survey the necessary biological background for this Thesis, such as DNA, Allele, SNP.

Chapter 3: In this chapter, we describe the input data and the data processing for extracting features. We also introduce the PLINK tool and describe our data processing pipeline.

Chapter 4: In this chapter we describe our methodology for feature selection. Our approach relies on outlier detection algorithms, where we view outlier SNPs as more informative.

Chapter 5: In this Chapter we consider different classification algorithms and we perform an experimental evaluation of how well they can predict melanoma. We consider the effect of different parameters on the classification performance, and manually study the SNPs that are important in the classification.

Chapter 6: In this Chapter we consider a larger dataset, and the use of Apache Spark for classification. We also study the performance of the classifiers in a real-world setting, where the number of cancer patients is a small fraction of the total population.

Chapter 7: Finally, in this chapter, we summarize our results and conclude the paper.

Chapter 2. Biological Background

In this Chapter we provide some basic biology background that is necessary for the Thesis.

2.1 Definition of DNA

DNA, or deoxyribonucleic acid, is the hereditary material in humans and almost all other organisms.

DNA is a long molecule that contains each person's unique genetic code. It holds the instructions for building the proteins that are essential for our bodies to function. DNA instructions are passed from parent to child, with roughly half of a child's DNA originating from the father and half from the mother. DNA is a two-stranded molecule that appears twisted, giving it a unique shape referred to as the **double helix**.

Each of the two strands is a long sequence of **nucleotides** or individual units made of:

- a phosphate molecule
- a sugar molecule called deoxyribose, containing five carbons
- a nitrogen-containing region

There are four types of nitrogen-containing regions called **bases**:

- adenine (A)
- cytosine (C)
- guanine (G)
- thymine (T)

The order of these four bases forms the genetic code, which is our instructions for life. The bases of the two strands of DNA are stuck together to create a ladder-like shape. Within the ladder, A always sticks to T, and G always sticks to C to create the "rungs". The length of the ladder is formed by the sugar and phosphate groups.

An important property of DNA is that it can replicate or make copies of itself. Each strand of DNA in the double helix can serve as a pattern for duplicating the sequence of bases. This is critical when cells divide because each new cell needs to have an exact copy of the DNA present in the old cell." [2] [3]

2.2 Definition of Chromosome

Chromosomes are thread-like structures in which DNA is tightly packaged within the nucleus. DNA is coiled around proteins called histones, which provide the structural support. Chromosomes help ensure that DNA is replicated and distributed appropriately during cell division. Each chromosome has a centromere, which divides the chromosome into two sections – the p (short) arm and the q (long) arm. The centromere is located at the cell's constriction point, which may or may not be the center of the chromosome.

In humans, 46 chromosomes are arranged in 23 pairs, including 22 pairs of chromosomes called autosomes. Autosomes are labeled 1-22 for reference. Each chromosome pair consists of one chromosome inherited from the mother and one from the father.

In addition to the 22 numbered autosomes, humans also have one pair of sex chromosomes called an allosome. Instead of labeling these chromosome pairs with numbers, allosomes are labeled with letters such as XX and XY. Females have two copies of the X chromosome (one inherited from the mother and one from the father). Males have one copy of the X chromosome (inherited from the mother) and one copy of the Y chromosome (inherited from the father).

Arranged on the chromosomes are genes. Genes are made of DNA and contain the instructions for building proteins and are integral in making and maintaining the human body. [4]

2.3 Definition of Gene

A gene is the basic physical and functional unit of heredity. Genes, which are made up of DNA, act as instructions to make molecules called proteins. In humans, genes vary in size from a few hundred DNA bases to more than 2 million bases. The Human Genome Project has estimated that humans have between 20,000 and 25,000 genes.

Every person has two copies of each gene, one inherited from each parent. Most genes are the same in all people, but a small number of genes (less than 1 percent of the total) are slightly different between people. Alleles are forms of the same gene with small differences in their sequence of DNA bases. These small differences contribute to each person's unique physical features. [5]

2.4 Definition of Alleles

Alleles are different forms of a gene occupying a specific spot on a chromosome. That spot is called locus. They can be dominant or recessive.

Some characteristics, such as eye color and the shape of the earlobe, are controlled by a single gene. These genes may have different forms.

Different forms of the same gene are called alleles (pronounced al-eels). The gene for eye color has an allele for blue eye color and an allele for brown eye color.

Alleles are dominant or recessive:

- the characteristic controlled by a **dominant** allele develops if the allele is present on **one or both** chromosomes in a pair
- the characteristic controlled by a **recessive** allele develops only if the allele is present on **both** chromosomes in a pair

For example, the allele for brown eyes is dominant, while the allele for blue eyes is recessive. An individual who inherits one or two alleles for brown eyes will have brown eyes. An individual will only have blue eyes if they inherit two copies of the allele for blue eyes. [6] [7]

2.5 Definition of Genotype

“Your genotype is your complete heritable genetic identity; it is your unique genome that would be revealed by personal genome sequencing. However, the word genotype can also refer just to a particular gene or set of genes carried by an individual. For example, if you carry a mutation that is linked to diabetes, you may refer to your genotype just with respect to this mutation without consideration of all the other gene variants that you may carry“ [8]

Modern DNA analyzing techniques such as genome-wide association, or several machine learning appliances have made it easier to identify which segments of DNA are responsible for various phenotypes. “A genotype has different alleles, or forms. The different alleles are produced by mutations to the DNA and may give rise to beneficial or detrimental changes. In bacteria, the DNA exists in a ring and only one allele for each genotype is present. Sometime, an allele will mutate in a beneficial way, the organism will reproduce more, and the genotype will increase in the population. In sexually reproducing organisms, there are two alleles present in each organism, which can have complex interactions with each other, and other genes. Mutations can occur in these alleles, new combinations can arise during meiosis, and infinite amount of variety can be created. These combinations of genotype give rise to the enormous variety of life on Earth.”

In order to know if a genetic variation is important or not we have major and minor allele.

Major and Minor alleles simply refer to the frequency with which an allele is found in a given population: A Minor allele is one that is expresses less often than a Major one.

On the other hand, “risk alleles” are alleles that we know is responsible for common diseases and usually defined by the minor allele.

Common risk alleles are often detected by genome-wide association studies (GWAS).

GWAS are a type of case-control study in which people with the condition being studied are compared to similar people without the condition. Each person’s complete set of DNA, or genome, is surveyed by examining a strategically selected area of genetic markers, called single nucleotide polymorphisms (SNPs). The goal is to discover new “risk alleles” responsible for certain diseases.

2.6 Definition of phenotype

An individual's **phenotype** consists of the traits we can observe. These can include features of appearance, behavior, metabolism, or anything else we can detect.

On the other hand, an individual's **genotype** is what we call the genes that help to create that phenotype. There may be one gene at work, or more than one. And genes don't always tell the whole story; sometimes your environment also affects what phenotype you end up with.

Traits related to appearance are sometimes the easiest to observe. When Gregor Mendel was doing his famous experiments with pea plants, he observed the plants' appearance: the peas might be green or yellow, smooth or wrinkly. The plants could also be regular height or dwarfed.

Humans have appearance phenotypes, too; for example, your height and your eye color are both phenotypes controlled, at least partly, by your genes.

Behavior can be a phenotype, too. Border collies were bred to herd sheep, so even if they have never seen a sheep in their life, they will display herding behaviors - like running around your house collecting all your pillows.

Most genes don't do anything as flashy as changing our eye color; instead, they make enzymes deep inside our cells. These enzymes do different little jobs that are important to keeping us alive, like running chemical reactions to help us digest our food or burn energy. These chemical reactions are called our **metabolism**.

One phenotype related to metabolism is lactose intolerance. If you have a gene that makes the enzyme **lactase**, you can easily digest the sugar (**lactose**) in milk. But if you are lactose intolerant, you don't have that enzyme, so you can't digest lactose and will feel sick when you drink milk. [9]

2.7 Definition of SNP

A single nucleotide polymorphism, or SNP (pronounced "snip"), is a variation at a single position in a DNA sequence among individuals. Recall that the DNA sequence is formed from a chain of four nucleotide bases: A, C, G, and T. If more than 1% of a population does not carry the same nucleotide at a specific position in the DNA sequence, then this variation can be classified as a SNP. If a SNP occurs within a gene, then the gene is described as having more than one allele.

SNPs occur normally throughout a person’s DNA. They occur once in every 300 nucleotides on average, which means there are roughly 10 million SNPs in the human genome. Most commonly, these variations are found in the DNA between genes. They can act as biological markers, helping scientists locate genes that are associated with disease.

Although a particular SNP may not cause a disorder, some SNPs are associated with certain diseases. These associations allow scientists to look for SNPs in order to evaluate an individual's genetic predisposition to develop a disease. In addition, if certain SNPs are known to be associated with a trait, then scientists may examine stretches of DNA near these SNPs in an attempt to identify the gene or genes responsible for the trait. [10, 11]

Chapter 3. Data Description

In this chapter, we describe the data we used for our study, and the tools for obtaining, and processing the data. The data was provided to us by the Faculty of Medicine of Ioannina. We used the Plink tool an open-source tool for processing biological data.

3.1 Introduction to PLINK

PLINK is a free, open-source whole genome association analysis toolset, designed to perform a range of basic, large-scale analyses in a computationally efficient manner [12]. We use PLINK in order to manipulate our initial data files and export data files we can read in order to continue our analysis. [12] PLINK to produce the needed files for the first feature extraction needs us input BIM, BED and FAM files in our case the phenotypic information exists in a separate text file which we also provide in PLINK as input. We describe thoroughly the input files on the next chapter

3.2 Input Data

In our disposal we have 22 Chromosomes for a given population. For each chromosome, we have a BED, FAM, and BIM file. We also got a text file containing the phenotype information

In the following we describe the structure of the files that we have in our disposal.

3.2.1 Phenotype.txt

The structure of the phenotype file shown in the following table.

Eid	Sex	Birth-year	cases
1000079	0	1957	1
1000257	1	1940	0
1000569	1	1945	0

The column Eid is the unique identifier for each patient associated with the id on FAM file. The sex and birth year columns were not used in our project. Finally, the last column takes value 0 if the patient is healthy if 1 if she has the disease. This is the attribute that we want to predict.

3.2.2 BED File

BED file is the most important one of the input files as it contains in binary encoding all the needed information. More thoroughly in BED file, we have for each SNP, the information of what chromosome belongs to, and then for each patient that has the specific SNP we have his unique EID and then the whole genetic sequence.

BED file has the following structure: [12]

```
0000000: 01101100 00011011 00000001 11011100 00001111 11100111l.....
0000006: 00001111 01101011 00000001…….
```

“The first 2 bytes are fixed in order for plink to confirm if the file is really a BED file. The third byte indicates if BED file is in SNP-major mode, which means that it lists all individuals for first SNP for second, etc.

So, in general the format of a bed file will be |-magic number--| |-mode-| |--genotype data---------| “

3.2.3 BIM File

The BIM file gives us the information for each SNP what minor and major allele nucleobase we have. So, when we encode both 3 files from BIM the useful columns are the id of SNP the chromosome number and the information about minor and major allele.

The BIM file has 6 columns from which we need only 5 of them. (Table1).

1. Chromosome number or name (either an integer, or 'X'/'Y'/'XY'/'MT'; '0' indicates that we don’t have the information at which chromosome belongs to).
2. SNP unique identifier (The unique code of each SNP, one SNP may exist only in one chromosome)

3. Position

Each chromosome divided into region’s the position column indicates the region where the SNP located at as well his position.

- A 1 (usually minor Allele)
- A 2 (major Allele)

Major and Minor alleles simply refer to the frequency with which an allele is found in a given population: A Minor allele is one that is expresses less often than a Major one.

CHR	Variant identifier	Position	A1	A2
1	rs533090414	18849	C	G
1	rs529030968	798510	T	C
1	rs573301795	888960	A	G
1	rs534967597	1218124	T	G

Table 1

3.2.4 FAM File

We may give plink FAM file as input, but, in our case, we use it only as pivot table to connect BED and BIM files with phenotype.txt. That was because the medical department gave us specific phenotype to examine so all the information of FAM file replicated in Phenotype file.

Also, despite the situation PLINK requires the initial 3 files and then any additional input.

The fields in a FAM file are the following (Table 2):

1. Family ID
2. Within-family ID ('IID'; cannot be '0')

In our case Family ID is the same as EID (unique patient id) that we describe on phenotype file. In our Data we have not any relationships between the patients, so Family ID and IID are the same

1. Within-family ID of father ('0' if father isn't in dataset)
2. Within-family ID f mother ('0' if mother isn't in dataset)

3. Sex code ('1' = male, '2' = female, '0' = unknown)

4. Phenotype value ('1' = control, '2' = case, '-9'/'0'/non-numeric = missing data if case/control)

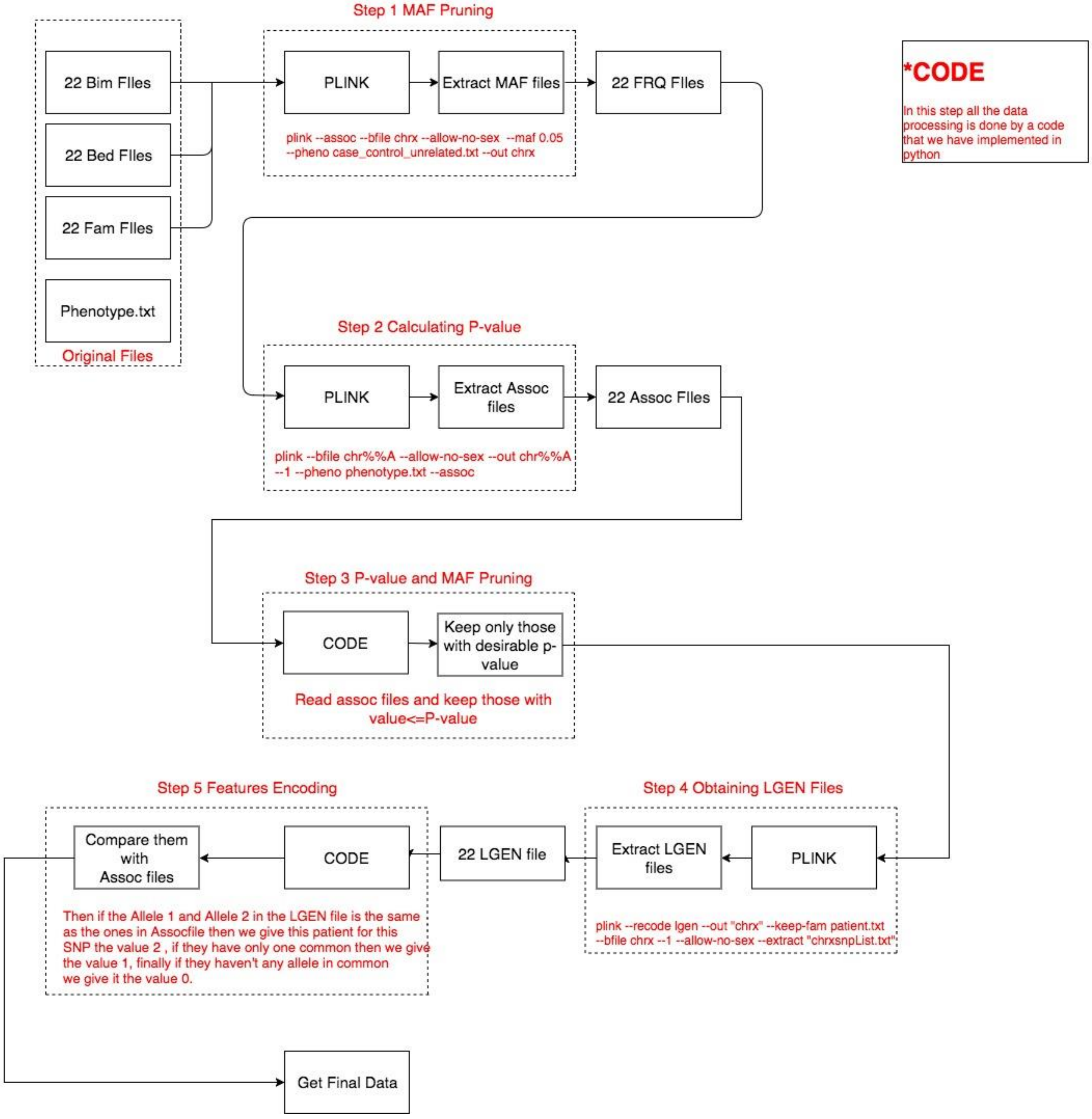
In our case we take the phenotype value from the input text file, so this column has a default value of -9.

Family ID	IID	ID of father	ID of mother	Sex code	Phenotype value
4612776	4612776	0	0	0	-9
4822852	4822852	0	0	0	-9
1659770	1659770	0	0	0	-9
5890646	5890646	0	0	0	-9
2672059	2672059	0	0	0	-9

Table 2

3.3 Feature Extraction

Based on the initial files BIM BED FAM and phenotype.txt we do five steps to do our first feature extraction. Those steps appear on the next schema(pipeline) and then we explain each individual step in more details.



```
plink --assoc --bfile chrx --allow-no-sex  --maf 0.05
--pheno case_control_unrelated.txt --out chrx
```

In this step all the data processing is done by a code that we have implemented in python

```
plink --bfile chr%%A --allow-no-sex --out chr%%A
--1 --pheno phenotype.txt --assoc
```

Read assoc files and keep those with value<=P-value

```
plink --recode lgen --out "chrx" --keep-fam patient.txt
--bfile chrx --1 --allow-no-sex --extract "chrxsnpList.txt"
```

Then if the Allele 1 and Allele 2 in the LGEN file is the same as the ones in Assocfile then we give this patient for this SNP the value 2 , if they have only one common then we give the value 1, finally if they haven't any allele in common we give it the value 0.

3.3.1 Step 1 MAF Pruning

MAF or else Minor Allele Frequency is a metric which calculate the frequency of minor allele in the given population. MAF is widely used in population genetics studies because it provides information to differentiate between standard and rare variants in the population.

We need to prune from our data the rare SNP's for the given phenotype because otherwise, we are going get a statistical error in the calculation of p-value on the next step. Also, when we examine a disease, we do not need to take under examination the SNP's which doesn't affect the manifestation of the disease

We used a threshold for MAF of 0.05, which means that we exclude from our dataset the SNP's that exist in under the 5% of our population.

To calculate MAF we use PLINK which take as input files the 22 BIM, BED, FAM and Phenotype file and produce 22 FRQ files. The command used is **plink --assoc --bfile chrx --allow-no-sex --maf 0.05 --pheno phenotype.txt --out chrx**. The result is 22 FRQ files

The FRQ file has the following format: [12]

- CHR: Chromosome Code
- SNP: SNP unique identifier
- A1: Allele 1 (usually minor)
- A2: Allele 2

Major and Minor alleles simply refer to the frequency with which an allele is found in a given population: A Minor allele is one that is expresses less often than a Major one.

- MAF: Allele 1 frequency

The calculated minor allele frequency based on the population given from the Phenotype file. By default, plink keeps on the extracted files only those SNP's with **MAF above a specific threshold** (0.05 on our experiments)

CHR	SNP	A1(minor allele)	A2(major allele)	MAF
1	rs367896724	AC	A	0.3706
1	rs555500075	TA	T	0.3428
1	rs62635286	G	T	0.0819
1	rs200579949	G	A	0.0819

Table 5

3.3.2 Step 2 Calculating P-value

After we got the 22 FRQ files we calculate p-value based on the given population . P-value is a metric which exclude the random possibility of an effect to occur.

Moreover, the null hypothesis, H0 is the commonly accepted fact, for example in our case the null hypothesis is that all the SNP's has not a visible linkage between them in causing of cancer, so we give our solution to the problem or else the alternative hypothesis.

The alternative hypothesis is the one you would believe if the null hypothesis is concluded to be untrue. The evidence in the trial is your data and the statistics that go along with it. All hypothesis tests ultimately use a p-value to weigh the strength of the evidence

The p-value is a number between 0 and 1 and interpreted in the following way:

- A small p-value (typically ≤ 0.05) indicates strong evidence against the null hypothesis, so you reject the null hypothesis.
- A significant p-value (> 0.05) indicates weak evidence against the null hypothesis, so you fail to reject the null hypothesis.
- p-values very close to the cutoff (0.05) are considered to be marginal (could go either way).

In genetics to validate our results we need a p-value <= 0.01. In our experiments we used a p-value threshold<=0.001.

To calculate P-value we use PLINK which take as input files the 22 BIM, BED, FAM and Phenotype file and produce 22 Assoc files. The command used is **plink –bfile chr%%A --allow-no-sex --out chr%%A --1 --pheno phenotype.txt --assoc fisher --pfilter 0.001**

The result is 22 ASSOC files

The ASSOC file has the following format in our experiment we the following columns. (Table 3): [12]

- CHR: Chromosome Code
- SNP: SNP unique identifier
- A1: Allele 1 (minor allele)
- A2: Allele 2(Major allele)

Major and Minor alleles simply refer to the frequency with which an allele is found in a given population: a Minor allele is one that is expresses less often than a Major one.

- CHISQ: Allelic test chi-square statistic
- P: Allelic test p-value

Plink produce a chi-square statistic score for each SNP then based on the produced score calculates p-value.

CHR	SNP	A1	A2	*CHISQ*	P
1	rs533090414	C	G	21.22	4.086e-06
1	rs529030968	T	C	19.46	1.025e-05
1	rs573301795	A	G	18.38	1.811e-05
1	rs534967597	T	G	18.4	1.788e-05

Table 3

3.3.3 Step 3 P-value and MAF Pruning

After we produce the 22 ASSOC files, we give FRQ files ASSOC files as input in our code (described at Appendix) and we check the given thresholds. If the SNP tested has a p-value more than 0.001 we remove it from ASSOC file, we repeat the procedures for all 22 ASSOC files.

In genetics smaller p-value means for the given population with given phenotype that with there is a high probability, this SNP to be relevant with the occurrence of the disease.

After we remove from ASSOC files the SNP's with p-value bigger than our threshold we read again the files from our code and this time we remove the SNP's from ASSOC that do not exist on the corresponding MAF file.

The outcome of this procedures is the genetic information we need for each SNP without the rare SNP's and with a p-value less than 0.001.

Why don't use only P-value for Pruning

When we have for an individual too much missing genotype data at BED file, then PLINK try to replicate the missing information.

This attempt of PLINK has an outcome in some cases of extremely low p-value which means for our experiments that an in other cases irrelevant SNP becomes very important.

So we calculate MAF at first to know which SNP's are rare in our case.

Once individuals with too much missing genotype data have been excluded, subsequent analyses can be set to automatically exclude SNPs on the basis of MAF (minor allele frequency).

So PLINK continues in our experiments to replicate wrong p-values but then with MAF we exclude them from the final files.

3.3.4 Step 4 Obtaining LGEN Files

We also calculate the LGEN file which contains the IID of the patient, his SNP, and allele 1 and allele 2. ASSOC and LGEN in format may look alike, but they are not, in ASSOC for each chromosome and each SNP we have which allele appear to be dominant and which appear to be minor. On the other hand, on LGEN file, we have for a patient in a specific SNP his genotype sequence. These means that if for a random patient we have allele 1 G and allele 2 C that the sequence may look like GCGCGCGCGGCG…..

The LGEN file has the following format: [12]

- Family ID
- Within-family ID

In our case Family ID is the same as EID (unique patient id) that we describe on phenotype file. In our Data we have not any relationships between the patients, so Family ID and IID are the same

- SNP unique identifier
- Allele 1
- Allele 2

Family ID	Within-family ID	SNP	Allele 1	Allele 2
4632109	4632109	rs533090414	G	G
4632109	4632109	rs529030968	C	C
4632109	4632109	rs573301795	G	G
4632109	4632109	rs534967597	G	G

Table4

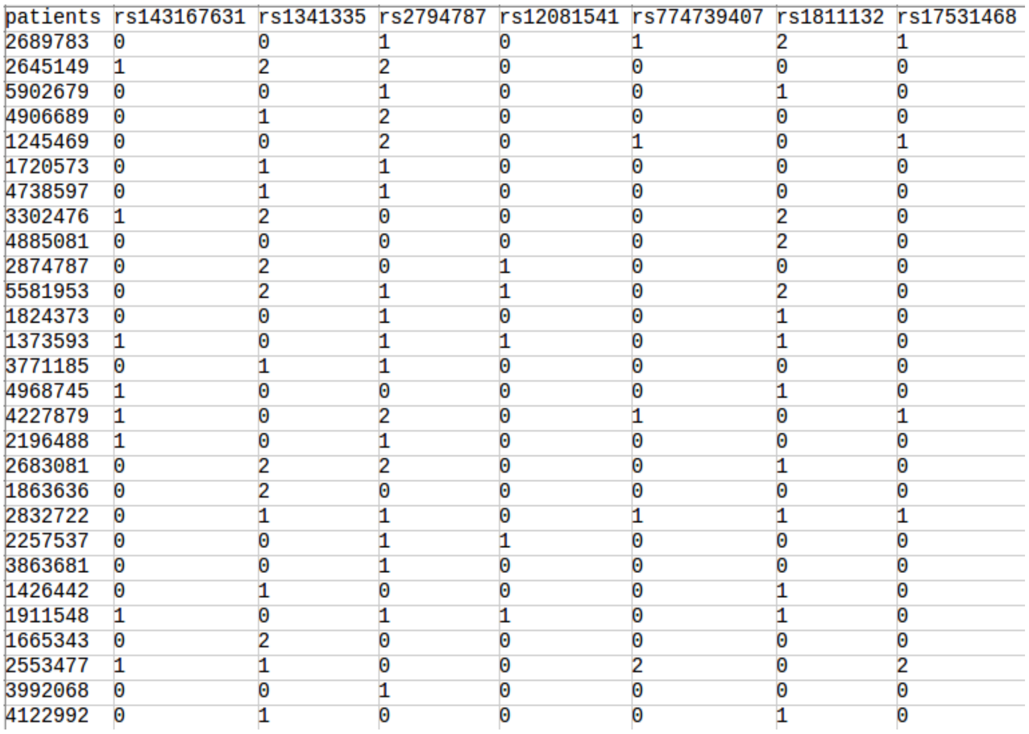
3.3.5 Step 5 Features Encoding

After we get the LGEN and ASSOC files, we input them in our code and compare LGEN file with the pruned ASSOC file for each SNP for each patient. Using this data, we create an M×N matrix, where M is the number of people and N is the number of SNPs. For each patient-SNP entry in the matrix, we give a value 0, 1 or 2. [9, 10].

If a patient has a value 2 for a particular SNP, we weighted it as relevant, a value of 0 means that this SNP is irrelevant.

The values generated as follows. We check the two alleles of a patient from LGEN file and compare them with allele1 of ASSOC file. We give the value 2 when both of LGEN alleles are the same as the minor allele from ASSOC file, value 1 when one of them is the same, and value 0 when neither of them is the same. The created matrix will be used as input to the classification models.

The output file has the following format:(image 1)



patients	rs143167631	rs1341335	rs2794787	rs12081541	rs774739407	rs1811132	rs17531468
2689783	0	0	1	0	1	2	1
2645149	1	2	2	0	0	0	0
5902679	0	0	1	0	0	1	0
4906689	0	1	2	0	0	0	0
1245469	0	0	2	0	1	0	1
1720573	0	1	1	0	0	0	0
4738597	0	1	1	0	0	0	0
3302476	1	2	0	0	0	2	0
4885081	0	0	0	0	0	2	0
2874787	0	2	0	1	0	0	0
5581953	0	2	1	1	0	2	0
1824373	0	0	1	0	0	1	0
1373593	1	0	1	1	0	1	0
3771185	0	1	1	0	0	0	0
4968745	1	0	0	0	0	1	0
4227879	1	0	2	0	1	0	1
2196488	1	0	1	0	0	0	0
2683081	0	2	2	0	0	1	0
1863636	0	2	0	0	0	0	0
2832722	0	1	1	0	1	1	1
2257537	0	0	1	1	0	0	0
3863681	0	0	1	0	0	0	0
1426442	0	1	0	0	0	1	0
1911548	1	0	1	1	0	1	0
1665343	0	2	0	0	0	0	0
2553477	1	1	0	0	2	0	2
3992068	0	0	1	0	0	0	0
4122992	0	1	0	0	0	1	0

Image 1

In the first column we have the patient id, and then for each patient, we have for each SNP an indicator of how many common alleles he has as we described above more thoroughly.

Chapter 4. Features Selection

In this section we consider techniques for selecting features.

In order to reduce the size of the initial data and improve the quality of classification we consider techniques for feature selection. We note that we have already performed some feature selection already, by thresholding on the MAF and the p-value.

Our feature selection techniques rely on the idea that SNPs that are very different from the rest of the SNPs will be good predictors. To find such SNPs we need a way to measure similarity, and an algorithm for finding outliers.

4.1 Similarity

We use some similarities techniques to compare the features between them.

4.1.1 Pearson's Correlation

In our experiments, we use the Pearson's correlation of two SNPs in order to select features and reduce the $M \times N$matrix to a $M \times D$matrix ($D < N$). The correlation coefficient is a metric that measures the dependence of two variables X and Y and it is defined as $\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y}) / \sqrt{\Sigma_{i=1}^{n}(x_i - \bar{x})^2(y_i - \bar{y})^2}$

The Pearson's correlation takes values between -1 and 1, where 1 means perfect correlation, while -1 means negative correlation. Value of 0 implies independence.

4.1.2 Cramér's V

In statistics, Cramér's V (sometimes referred to as Cramér's phi) is a measure of association between two nominal variables, giving a value between 0 and +1. It is based on Pearson's chi-squared statistic.Cramér's V is computed by taking the square root of the chi-squared statistic divided by the sample size and the minimum dimension minus 1:

$$V = \sqrt{\frac{\frac{\chi^2}{n}}{\min(k-1, r-1)}}$$

where:

- χ^2 is derived from Pearson's chi-squared test
- **n** is the grand total of observations and
- **k** being the number of columns.
- **r** being the number of rows.

4.2 Feature Selection

We use two algorithms for feature extraction. The two algorithms are described below.

4.2.1 K-NN Outlier Detection

The distance to the *k*th nearest neighbor can also be seen as a local density estimate and thus is also a popular outlier score in anomaly detection. The larger the distance to the *k*-NN, the lower the local density, the more likely the query point is an outlier To take into account the whole neighborhood of the query point, the average distance to the *k*-NN can be used. Although quite simple, this outlier model, along with another classic data mining method, local outlier factor, works quite well also in comparison to more recent and more complex approaches, according to a large scale experimental analysis.

We use the k-nn outlier detection algorithm in order to find the outliers of our data. We apply k-nn outlier detection to Pearson Correlation Matrix and to Cramer V Matrix. For i-th SNP we find the k-neighbors with the higher similarities. The highest similarity of k-neighbors is the score for i-th SNP. After we calculate the scores of SNPs, we keep the SNP_i if a $score_i$ is high or equal to a threshold θ. In order to set the threshold θ we make some plots for different thetas and make our decision about theta based in plots.

4.2.2 Algorithm 2

We compute the Pearson Correlation for all $\binom{N}{2}$ pairs, where N is the number of SNPs. For a chosen threshold θ, we keep the SNPs which do not have correlation coefficient higher or equal to θ with any other SNP or with a percentage (we defined) of SNPs. For example, the i-th SNP S_i belongs to the low correlation category if $CORR(S_i, S_j) < \theta$ for all SNPs S_j $(1 \leq j \leq N)$. In this way, we want extract SNPs that behave in a very unique way. Another reason is to avoid the classification overfitting.

Chapter 5. Experiments

In this Chapter, we study experimentally different classification algorithms for phenotype prediction using genotype information. The goals of the experiments are the following: (1) To test different classification algorithms for the prediction task; (2) To test the effect of different parameters on the classification accuracy; (3) To test the effect of feature selection on the classification accuracy. The results below calculated after a 10-cross validation

5.1 Dataset

In this set of experiments, we used a dataset for xxx individuals with a total size of 2.53 TB. After the data manipulation that we described in Chapter 3 we obtain a table with 4980 rows and 7799 columns. The 4890 rows are our patients that take them from phenotype file. The 7799 columns are the SNP from feature extraction.

5.2 Software

In this set of experiments, we use Python 3 and jupyter notebook as a user interface. We used the scikit[11] learn library and the implemented classification algorithms easily. Scikit-learn (formerly scikits.learn) is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy. We also used NumPy for the data frames and arrays we import the data.

5.3 Evaluation Metrics

5.3.1 CONFUSION MATRIX

	True Condition	True Condition
	Condition Positive	Condition Negative
Predicted condition Positive	True Positives	False Positives
Predicted condition Negative	False negatives	True Negatives

5.3.2 Accuracy:

Accuracy is also used as a statistical measure of how well a binary classification test correctly identifies or excludes a condition. That is, the accuracy is the proportion of true results (both true positives and true negatives) among the total number of cases examined.

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{True\ Positives + True\ Negatives}}{(\boldsymbol{True\ Prositives + True\ Negatives + False\ Negatives + False\ Negatives})}$$

5.3.3 RECALL:

$$\boldsymbol{Recall} = \frac{\boldsymbol{True\ Positives}}{(\boldsymbol{TruePrositives + False\ Negatives})}$$

Recall refers to the test's ability to correctly detect ill patients who do have the condition.

5.3.4 PRECISION:

$$\boldsymbol{Precision} = \frac{\boldsymbol{True\ Positives}}{(\boldsymbol{True\ Prositives + False\ Positives})}$$

Precision is the probability that patients who predicted to have the disease, truly have the disease.

5.3.5 F score:

$$F_{Score} = 2 * \frac{\boldsymbol{Precision * Recall}}{(\boldsymbol{Precision + Recall})}$$

The F_1 score is the harmonic average of the precision and recall, where an F_1 score reaches its best value at 1 (perfect precision and recall) and worst at 0.

5.3.6 AUC (Area Under the Curve)

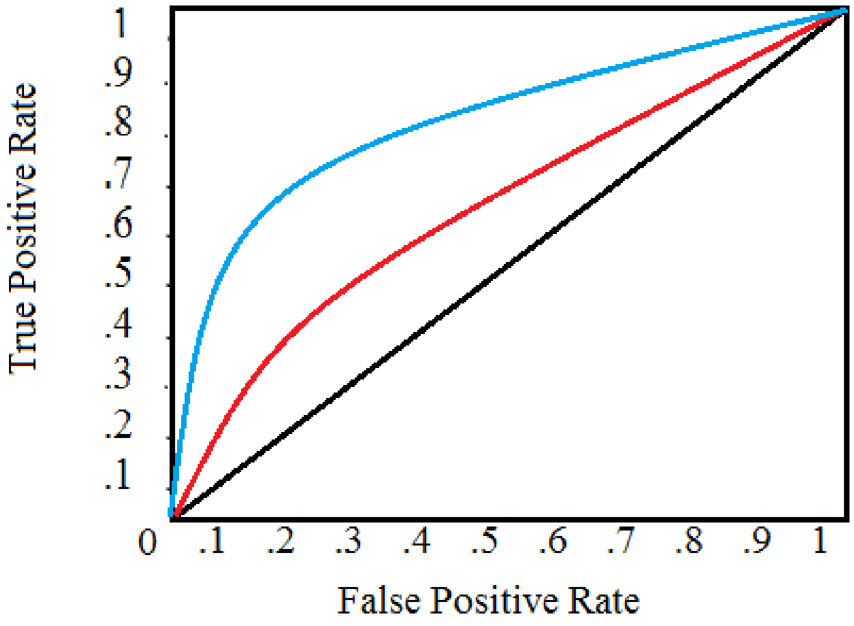
The area under the curve is the area under the roc curve. The roc curve has on x-axis the False Positive Rate and on y-axis the True Positive Rate. False Positive Rate defined as

$$\boldsymbol{False\ Positive\ Rate} = \frac{\boldsymbol{True\ Positives}}{\boldsymbol{all\ positives}}$$

and True Positive Rate defined as

$$\boldsymbol{True\ Positive\ Rate} = \frac{\boldsymbol{False\ Positives}}{\boldsymbol{all\ negatives}}$$

We plot roc curve on **points(True Positive Rate, False Positive Rate).** We set a threshold and the classifier with curve which is less close to black diagonal line is the better. Good classifier has an AUC of around 0.8, a very poor classifier has an AUC of around 0.5, and the perfect classifier has an AUC of close to 1.

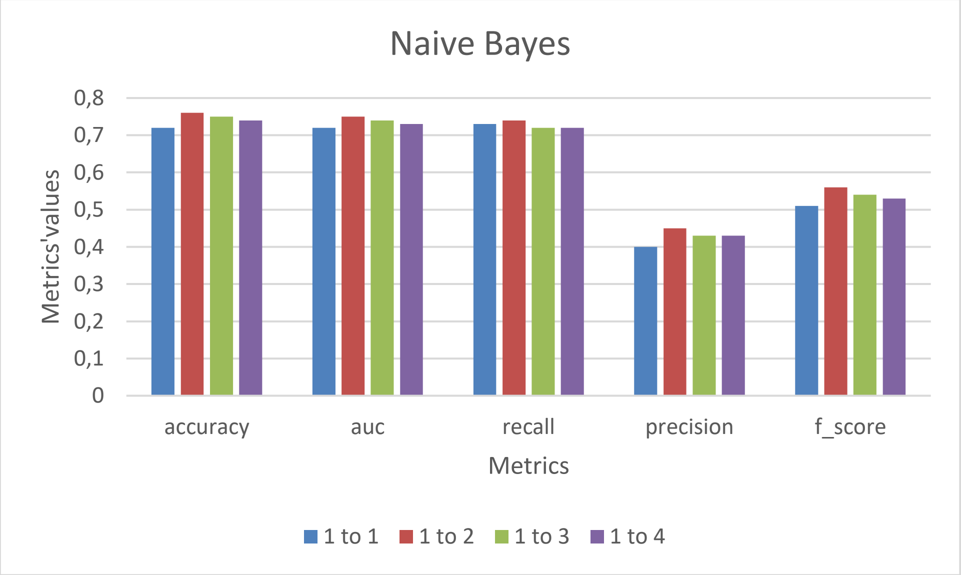


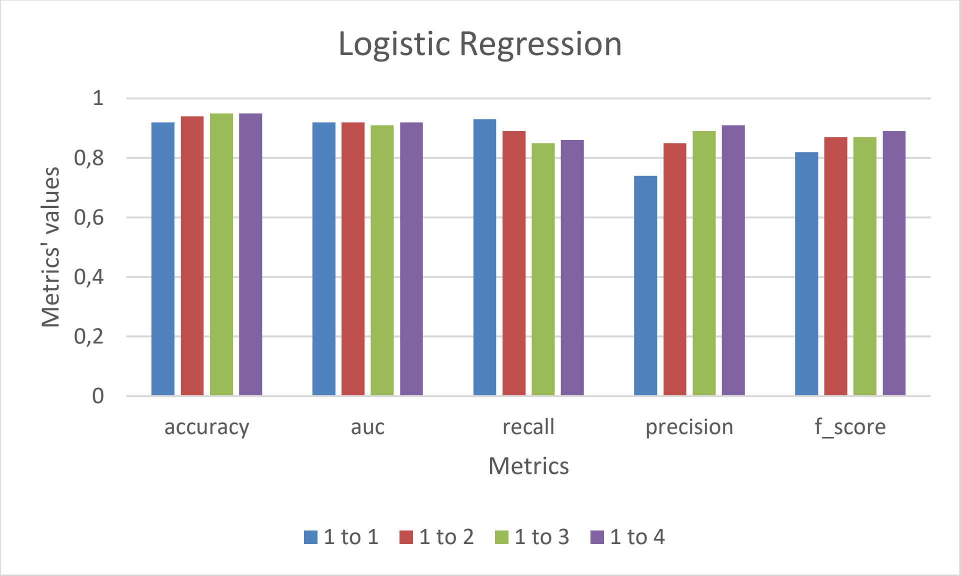
5.4 Results

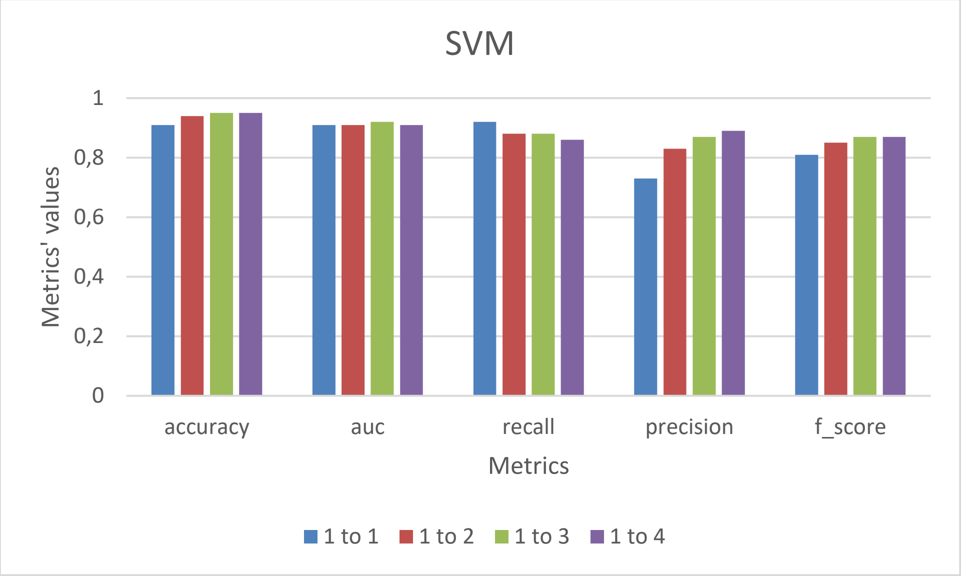
In our experiments, we use Logistic Regression, SVM, Decision Tree, Gaussian Bayes Bernoulli classifiers. We run different experiments: we run experiments with all features, we run experiments and we run experiments after features selection using Features Selection or Crammer V techniques. Before experiments start, we have to prepare our dataset. At first set, we run an experiment for p-value 0.001. All metrics' results below are the average after a 10-cross-validation.

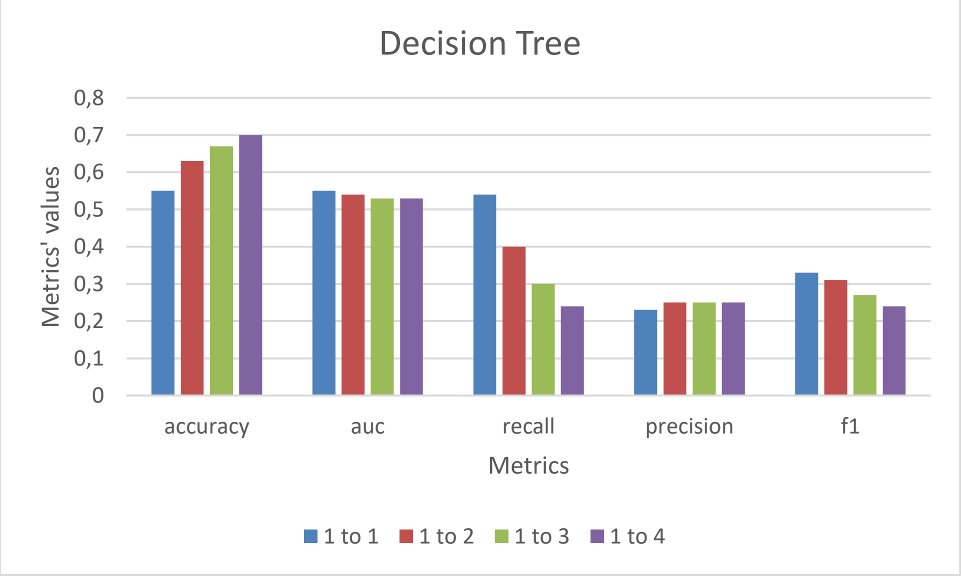
5.4.1 All features Experiment

In this experiment, we want to evaluate algorithms' prediction ability using all features. Our goal is to compare different classifiers using all features with different balances. The balances are one healthy for one patient, two healthies for one patient, three healthies for one patient and four healthies for one patient. As we see from the plots below, the best balance is one patient for two healthies, because the metrics' values are improved. As we see Logistic Regression and SVM classifiers can predict much better than Bayes Bernoulli and Decision Tree classifiers. The table below has the results of classifiers with balance 1 patient for 2 healthies. As we see from the table below Logistic Regression and SVM classifier are better than Naïve Bayes and Decision Tree.




Logistic Regression
1
0,8
0,6
0,4
0,2
0
Metrics' values
accuracy
auc
recall
precision
f_score
Metrics
1 to 1
1 to 2
1 to 3
1 to 4

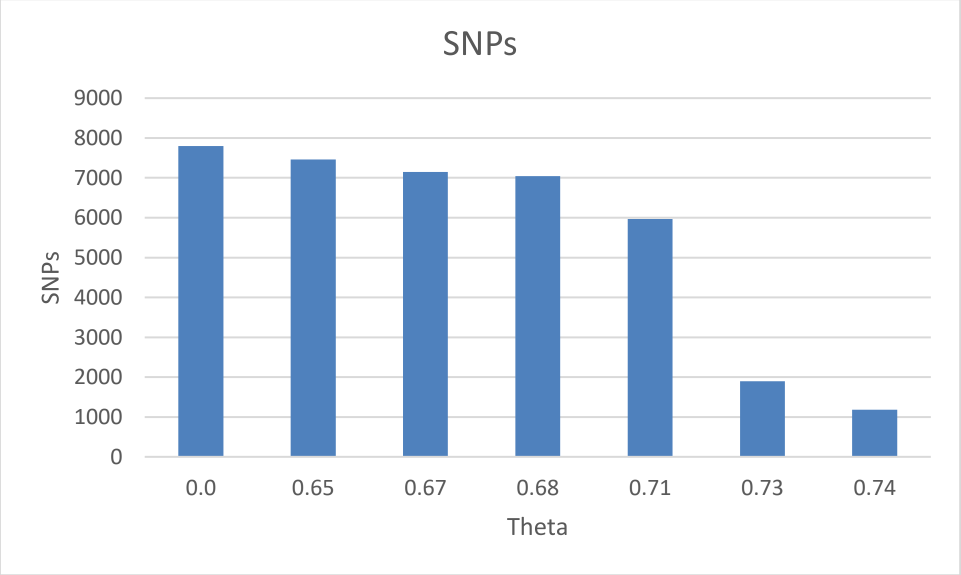

SVM
1
0,8
0,6
0,4
0,2
0
Metrics' values
accuracy
auc
recall
precision
f_score
Metrics
1 to 1
1 to 2
1 to 3
1 to 4


Decision Tree
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
Metrics' values
accuracy
auc
recall
precision
f1
Metrics
1 to 1
1 to 2
1 to 3
1 to 4

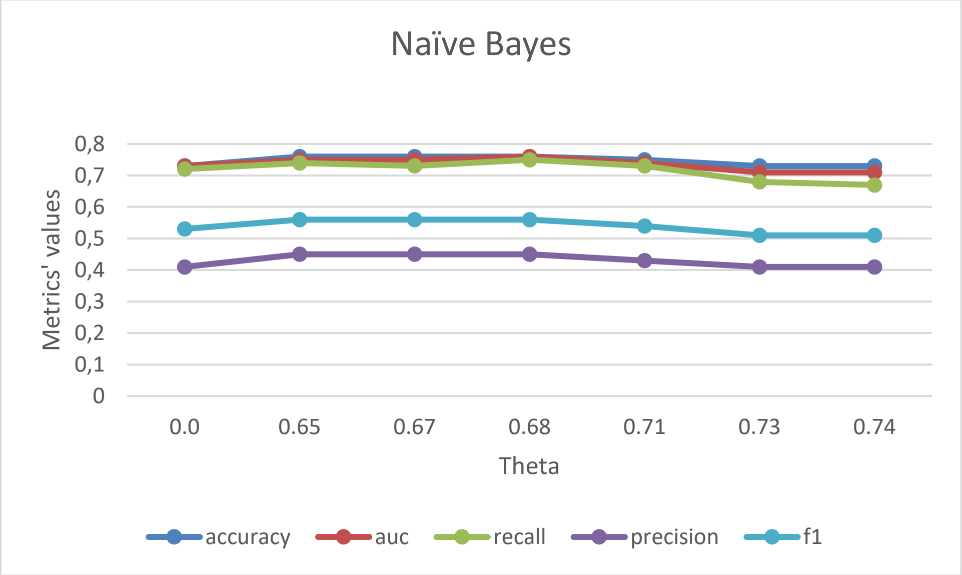
	accuracy	auc	recall	precision	F_score	Number of features
Naïve Bayes	0.73	0.73	0.73	0.41	0.53	7799
Logistic Regression	0.94	0.92	0.88	0.84	0.86	7799
SVM	0.94	0.92	0.90	0.84	0.87	7799
Decision Tree	0.63	0.55	0.42	0.26	0.32	7799

5.4.2 K-NN Outlier Detection

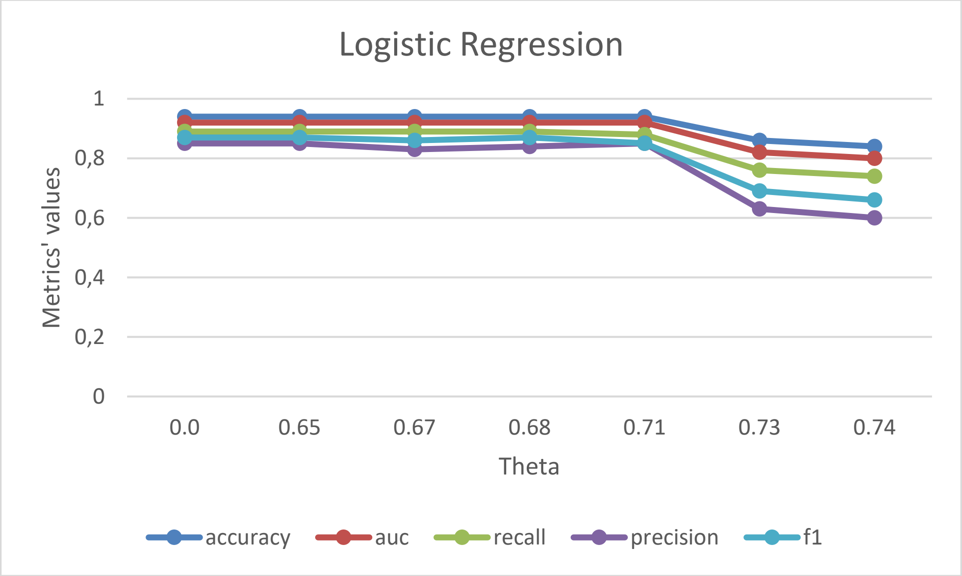
In this experiment, we use k-nn outlier detection to see predictabilities of classifiers. The plots below show the metrics' values of different classifiers for different thetas. As we see from the plots, the different evaluation metrics change when the threshold theta change. The table below has the results after features selection. We calculate the Cramer V test between all SNPs. We make some plots for different θ as we see below to choose the best θ. We make also a plot for numbers of SNPs based on different theta. The θ we choose according to plots is 0.71. The SNPs that remain are 6621 from 7799 and the results are almost the same. Feature selection does not help us to improve the results. We can say that with less features make almost the same prediction.



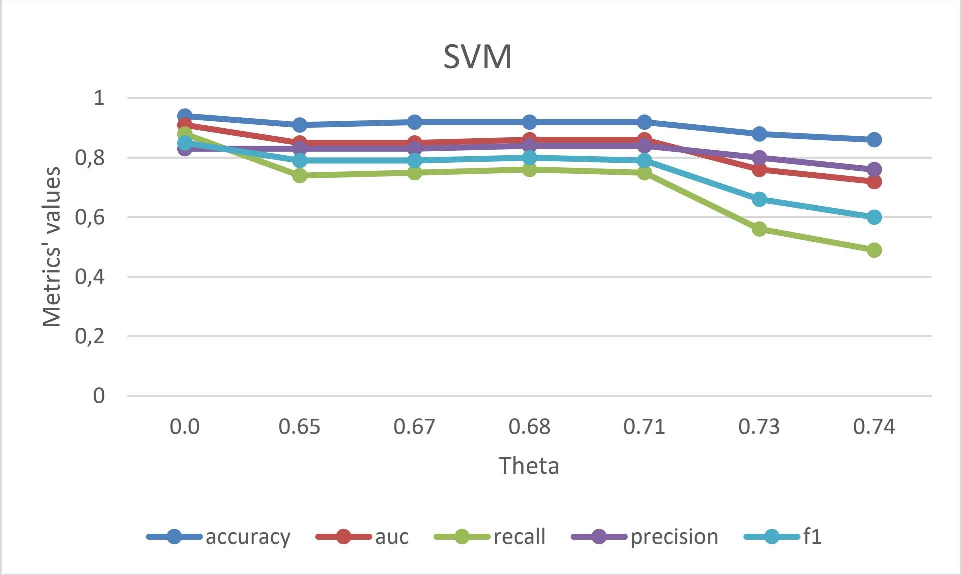
Number of SNPs based on thetas



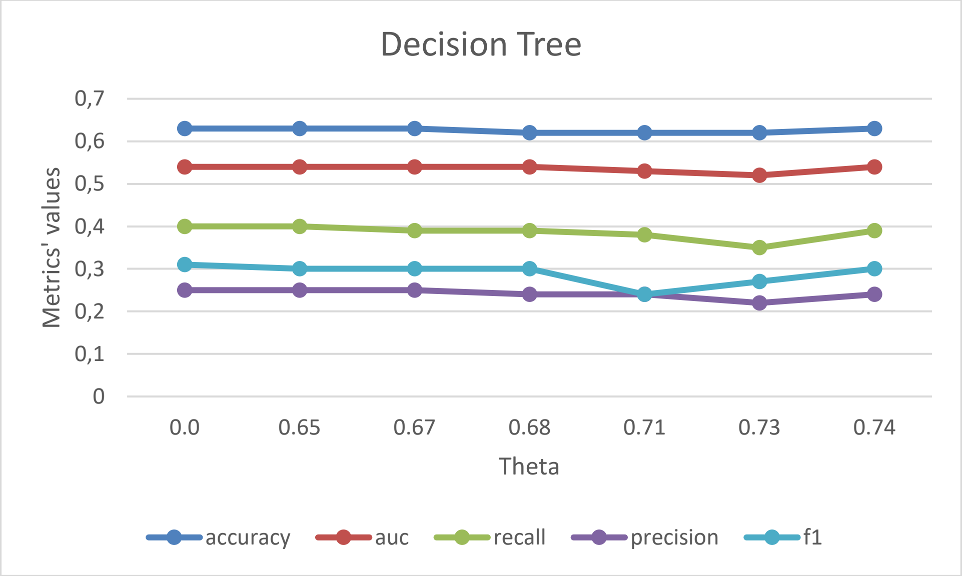
Naive Bayes based on thetas



Logistic Regression based on thetas



SVM based on thetas



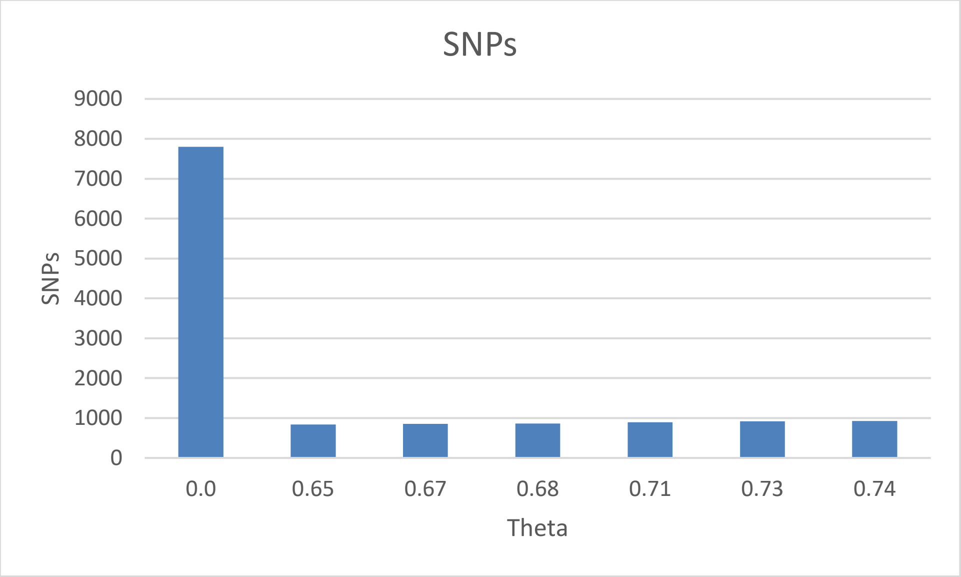
Decision Tree based on thetas

	accuracy	**auc**	**recall**	**precision**	**F_score**	**Number of features**
Naïve Bayes	0.75	0.74	0.74	0.43	0.54	6621
Logistic Regression	0.94	0.92	0.88	0.85	0.86	6621
SVM	0.92	0.86	0.75	0.84	0.79	6621
Decision Tree	0.63	0.53	0.36	0.23	0.28	6621

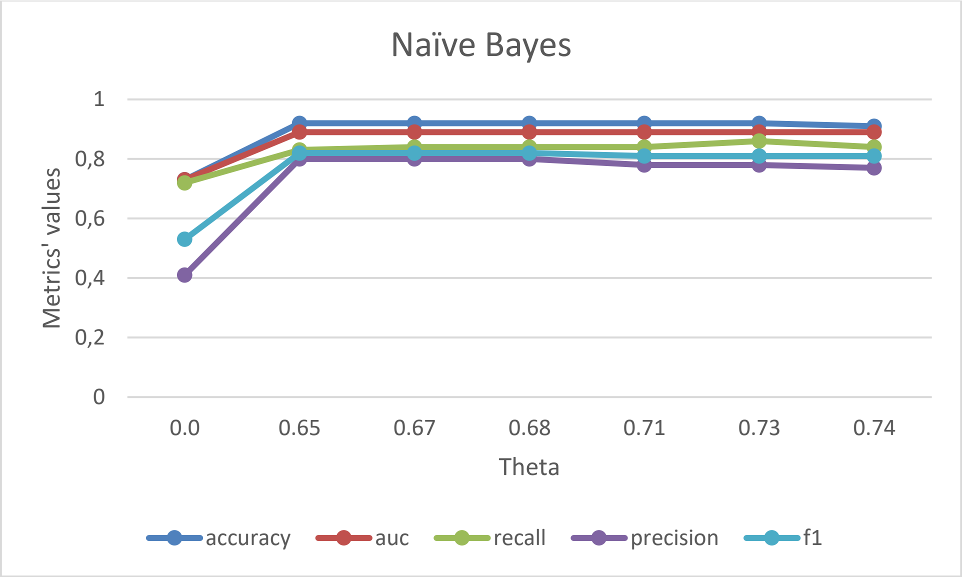
Evaluation Metrics based on ϑ = 0,71

5.4.3 Algorithm 2

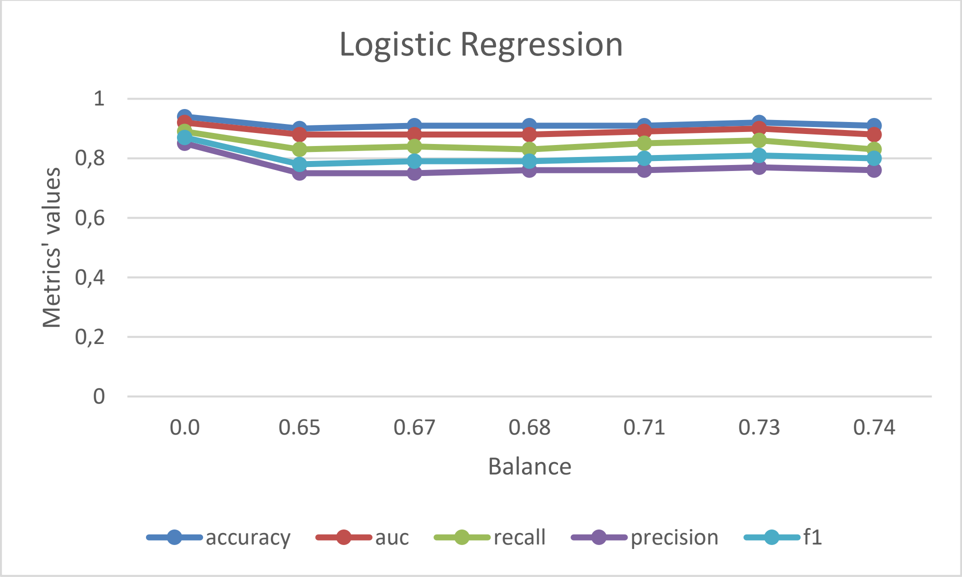
In this experiment, we use Algorithm 2 for feature selection. The resulting SNPs are significantly less than in the other experiments and the predictive ability of classifiers is almost the same. We have a plot for numbers of SNPs based on theta and 4 plots for evaluation metrics based on thetas for each classifier. The table below has the results based on Algorithm2 for reducing features. The threshold θ we set for this experiment was 0.71. According to this feature selection technique the SNPs that remain (883) are less than the original number of SNPs (7799) and the results of SVM classifier and Logistic Regression classifiers decrease. We can conclude that the Algorithm2 for feature selection does not improve the classifiers expect from Naïve Bayes. May the feature selection is good because the classifiers' performance is the almost with less features and not because improve their prediction ability.



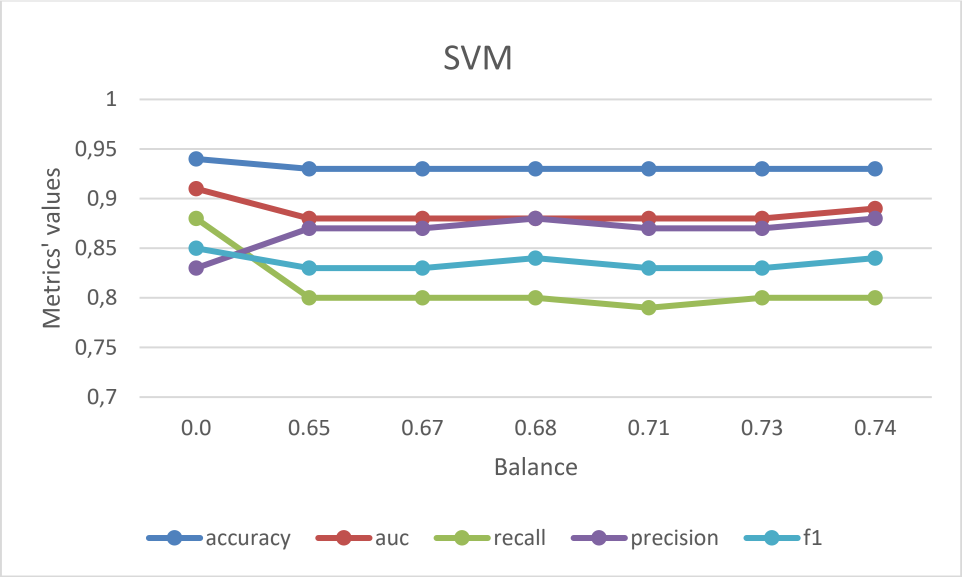
Number of SNPs based on thetas



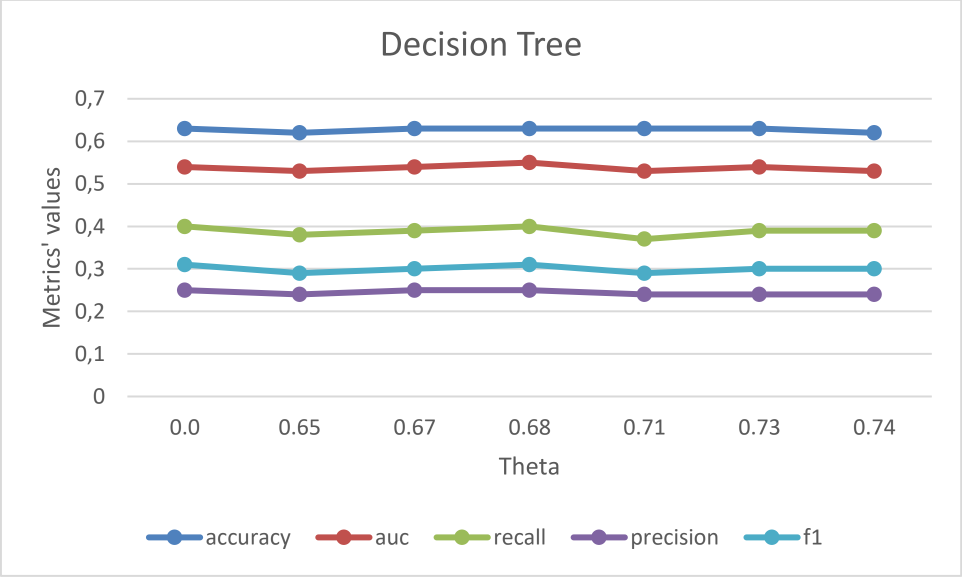
Naive Bayes based on thetas



Logistic Regression based on thetas



SVM based on thetas



Decision Tree based on thetas

	accuracy	auc	recall	Precision	F_score	Number of features
Naïve Bayes	0.92	0.89	0.84	0.79	0.81	883
Logistic Regression	0.91	0.88	0.83	0.76	0.79	883
SVM	0.93	0.89	0.81	0.87	0.84	883
Decision Tree	0.64	0.55	0.39	0.25	0.31	883

Evaluation Metrics based on ϑ = 0,71

5.5 Important SNPs

The list below has the top 30 SNPs based on coefficients values of Logistic Regression algorithm. A regression coefficient describes the size and direction of the relationship between a predictor and the response variable. Coefficients are the numbers by which the values of the term are multiplied in a regression equation We take the SNPs that we extract from K-NN Outlier Detection algorithm, and calculate the coefficients using Logistic Linear classifier. After calculating the coefficients, we choose the 30 SNPs which have the highest coefficients values than the others. These SNPs are "important" based on Logistic Regression. We compare these 30 SNPs according to Mr. Evangelou's & Alexander J Stratigos' paper **[1]** we find 3 SNPs that belong to same region: rs10965542, rs7866885and rs12289561. They are not exaclty the same according to paper. The snps we found, exist in the same region as the ones of the paper. Belongs to same region means that they behave in a simiral way.

rs201991408
rs17853865
rs545187316
rs17637117
rs5777200
rs112153839
rs571851010
rs12214534
rs70959416
rs10946654
rs5899038
rs7003912
rs78903746
rs12291321
rs150866057
rs200037943
rs265996
rs6456122
rs4398852
rs6506628
rs541534244
rs202091130
rs76144877
rs149091100
rs9341971
rs201176671
rs36090976
rs9507104
rs4361381
rs12532089

top 30 SNPs

Chapter 6. Experiments with large data

In this Chapter we experiment with a larger dataset, consisting of 272176 individuals, 2156 of whom are patients and 270020 are control cases. The second dataset have 22 chromosomes into three files each (bim, bed, fam) after the data manipulation that we explained earlier we had a table with 272176 columns and 12002 rows. So, for these experiments, we could not use anymore the previous technologies mentioned because of RAM issues, so we start using Spark to resolve those problems.

6.1 Apache Spark

Spark is a general-purpose data processing engine that is suitable for use in a wide range of circumstances. Application developers and data scientists incorporate Spark into their applications to rapidly query, analyze, and transform data at scale. Tasks most frequently associated with Spark include interactive queries across large data sets, processing of streaming data from sensors or financial systems, and machine learning tasks.

Spark began life in 2009 as a project within the AMPLab at the University of California, Berkeley. More specifically, it was born out of the necessity to prove out the concept of Mesos, which was also created in the AMPLab. Spark was first discussed in the Mesos white paper Mesos: A Platform for Fine-Grained Resource Sharing in the Data Center, written most notably by Benjamin Hindman and Matei Zaharia.

Spark became an incubated project of the Apache Software Foundation in 2013, and it was promoted early in 2014 to become one of the Foundation's top-level projects. Spark is currently one of the most active projects managed by the Foundation, and the community that has grown up around the project includes both prolific individual contributors and well-funded corporate backers such as Databricks, IBM, and China's Huawei.

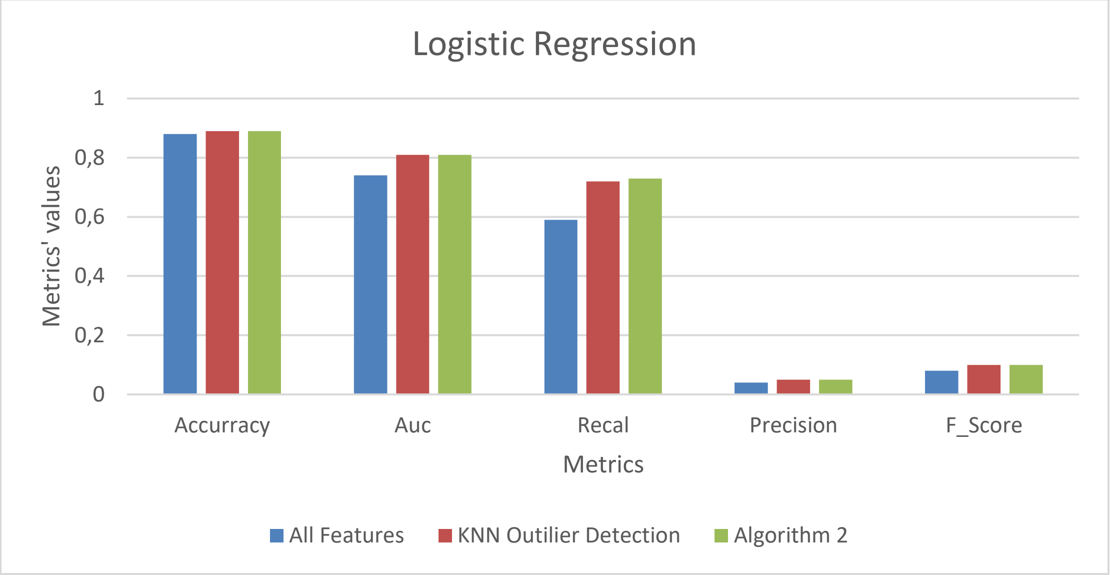
From the beginning, Spark was optimized to run in memory. It helps process data far more quickly than alternative approaches like Hadoop's MapReduce, which tends to write data to and from computer hard drives between each stage of processing. Spark's proponents claim that Spark's running in memory can be 100 times faster than Hadoop MapReduce, and also 10 times faster when processing disk-based data in a way similar to Hadoop MapReduce itself. This comparison is not entirely fair, not least because raw speed tends to be more important to Spark’s typical use cases than it is to batch processing, at which MapReduce-like solutions still excel.

6.2 Experimental Results

In this experiment we test the ability of the classifier to predict melanoma in a real-life setting, where the percentage of patients is very small.

In our data there is a huge class imbalance problem. The patients are 0,7% of the data. If we apply the classifier directly it will always predict the negative class. To deal with this problem, in every cross-validation, we balance of train data between healthy and patients people. For balancing we keep only two healthy people for every patient. In this experiment we use only Logistic Regression. When using all features, we observe that results are not good. The value of precision metric is very low because of the big difference between patients and healthy people. For example, a test data has 20000 people who the 19500 are healthy and the 500 are patients, the classifier predicts 16000 True negatives and 3500 false positives, predicts also 400 True positives and 100 False Negatives, so the precision will have low value because of that. We obtain relatively good recall (0.59), which means that we can identify patients with relatively high success rate.

When applying the feature selection algorithms, we observe that metrics improved. The low value of precision is due to big different between healthy people and patients. As we see the Logistic Regression is getting better when we use the features selection techniques. The recall that refers to the test's ability to correctly detect ill patients who do have the condition, is increased by 0,13. Precision, that is the probability that patients who predicted to have the disease truly have the disease, also increased by 0,01. We can conclude that in this experiment feature selection techniques help us to improve the prediction of Logistic Regression. The table below has the results of all features, results of k-nn outlier detection and results of algorithm2 feature selection.



Logistic Regression	accuracy	auc	recall	precision	F_score	Number of features
All features	0.88	0.74	0.59	0.04	0.08	12002
K-NN Outlier Detection	0.89	0.81	0,72	0.05	0.1	9318
Algorithm 2	0.89	0.81	0.73	0.05	0.1	9470

6.3 Important SNPs

The list below has the top 30 SNPs based on coefficients values of Logistic Regression algorithm. A regression coefficient describes the size and direction of the relationship between a predictor and the response variable. Coefficients are the numbers by which the values of the term are multiplied in a regression equation We take the SNPs that we extract from K-NN Outlier Detection algorithm, and calculate the coefficients using Logistic Linear classifier. After calculating the coefficients, we choose the 30 SNPs which have the highest coefficients values than the others. These SNPs are “important” based on Logistic Regression. We compare these 30 SNPs according to Mr. Evangelou’s & Alexander J Stratigos’ paper **[1]** we find 3 SNPs that belong to same region: rs10965542, rs7866885and rs12289561. They are not exaclty the same according to paper. The snps we found, exist in the same region as the ones of the paper. Belongs to same region means that they behave in a simiral way.

rs777418488
rs13026576
rs2167472
rs9508844
rs11816708
rs560001294
rs4615882
rs541943396
rs12500848
rs558002004
rs2329993
rs549618774
rs12504096
rs10965542
rs57115029
rs559274468
rs144525555
rs7866885
rs1032194
rs12289561
rs587676103
rs188383015
rs62510585
rs836116
rs6779726
rs3733907
rs67722884
rs6942584
rs62463054
rs116162314

top 30 SNPs

Chapter 7. Conclusion

The goal of the thesis was to predict if someone is predisposed to develop a disease in the future based on his or her genotype (gene) information. Our goal was to make an assisting tool to help the doctors extract useful results to diagnose if a patient has or not a disease in our case cancer melanoma.

To this end we built a data processing pipeline for extracting features from the input data, using the PLINK tool. We considered different classification and feature selection algorithms and different parameters for our tools. Our results show that the classifiers work well in a controlled setting of balanced data. In this setting feature selection does not appear to help in classification performance. Using the Logistic Regression coefficients, we identified important features. We observe that they agree in part with previous studies [1].

We also considered a larger dataset, and we experimented with the use of Apache Spark for handling large data. In our experiments with the larger data we tested our classification algorithms in a real-life setting where the patient population is a small fraction of the total population. We obtained good recall, but low precision. In this setting feature selection helps significantly in improving the recall and differentiate the patient genotypes.

7.1 Future Extensions

There some future work we can do in our research. First, we may try the experiments above using TensorFlow. TensorFlow is an open-source software library for dataflow programming across a range of tasks. It is a symbolic math library and is also used for machine learning applications such as neural networks. Another work is to extract MAF, ASSOC and LGEN files from PLINK using a balanced phenotype file.

Chapter 8. Bibliography

[1] Evangelos Evangelou, Alexander J. Stratigos, "Lessons from genome-wide studies of melanoma: towards precision medicine," *Expert Review of Precision Medicine and Drug Development,* pp. 443-449, 2016.

[2] DNA, "Genetic Home Reference," 3 July 2018. [Online]. Available: https://ghr.nlm.nih.gov/primer/basics/dna.

[3] T. Newman, "What is DNA and how does it work?," 111 Jan 2018. [Online]. Available: https://www.medicalnewstoday.com/articles/319818.php.

[4] "Healio learn Genomics," [Online]. Available: https://www.healio.com/hematology-oncology/learn-genomics/genomics-primer/what-are-chromosomes.

[5] Gene. [Online]. Available: https://ghr.nlm.nih.gov/primer/basics/gene.

[6] Allele, "Biology Online," 20 August 2017. [Online]. Available: https://www.biology-online.org/dictionary/Allele.

[7] BBC. [Online]. Available: http://www.bbc.co.uk/schools/gcsebitesize/science/add_aqa_pre_2011/celldivision/inheritance1.shtml.

[8] Genotype, "PGED," June 2018. [Online]. Available: https://pged.org/what-is-genotype-what-is-phenotype/.

[9] Phenotype, "study," [Online]. Available: https://study.com/academy/lesson/what-is-a-phenotype-definition-example-quiz.html.

[10] SNP2, "Nature," 9 July 2018. [Online]. Available: https://www.nature.com/scitable/definition/single-nucleotide-polymorphism-snp-295.

[11] SNP, "Genetic Home Reference," 3 July 2018. [Online]. Available: https://ghr.nlm.nih.gov/primer/genomicresearch/snp.

[12] Harvard, "plink," 25 Jan 2017. [Online]. Available: http://zzz.bwh.harvard.edu/plink/.

[13] "McMaster University," 31 Oct 2014. [Online]. Available: https://fhs.mcmaster.ca/pgp/documents/AnalysisGWASdatathroughPLINK.doc.

[14] Chromosome, "Genetic Home Reference," 3 July 2018. [Online]. Available: https://ghr.nlm.nih.gov/primer/basics/chromosome.

[15] Sun, Silke Szymczak Joanna M. Biernacka Heather J. Cordell Oscar González-Recio Inke R. König Heping Zhang Yan V., "Machine learning in genome-wide association studies," *Genetic Epidemiology,* vol. 33, no. S1, pp. 51-57, 2009.

[16] A. M. Nielsen, "Application of Machine Learning on a Genome-Wide Association Studies Dataset," KTH, School of Engineering Sciences (SCI), Mathematics (Dept.), Numerical Analysis, NA., 2015.