Phenotype Prediction Using Genotype

Thesis Presentation

Τμήμα Μηχ. Η/Υ & Πληροφορικής

Πανεπιστήμιο Ιωαννίνων

Department of Computer Science & Engineering

University of Ioannina

Authors: Antonis Moulopoulos & Georgios Salteris

Supervisor: Panayiotis Tsaparas,Associate Professor

Introduction

Thesis Goal

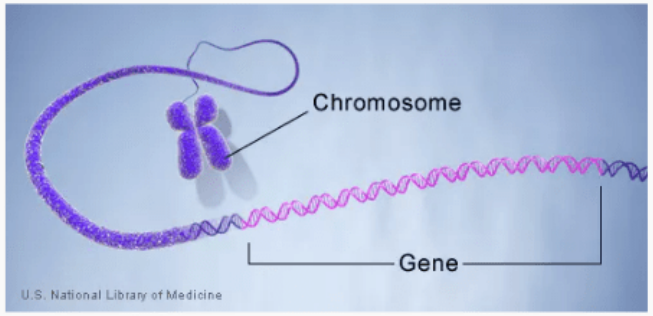
- Try to predict if someone is predisposed to develop a disease in the future

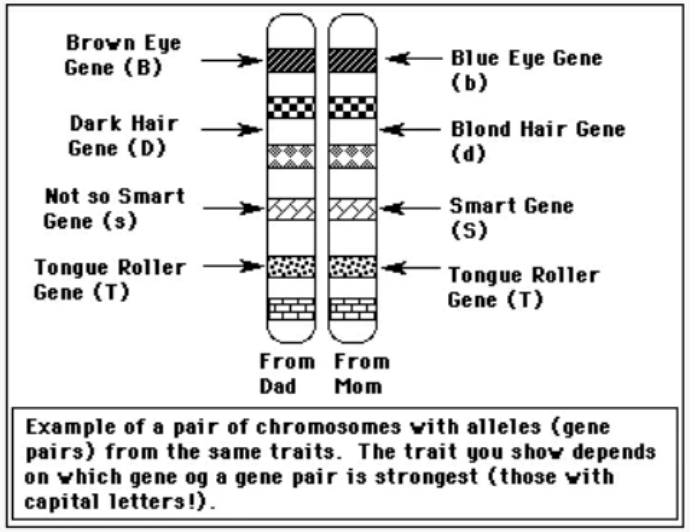
- we use several feature selection Algorithms as well as several ML algorithms.

- Cross Validate our results with existing studies

- Create an assisiting tool for Doctors

Biological Background
Part 1





Example of a pair of chromosomes with alleles (gene pairs) from the same traits. The trait you show depends on which gene og a gene pair is strongest (those with capital letters!).

Chromosomes

- 46 Chromosomes
- 23 from father
- 23 from mother

Gene

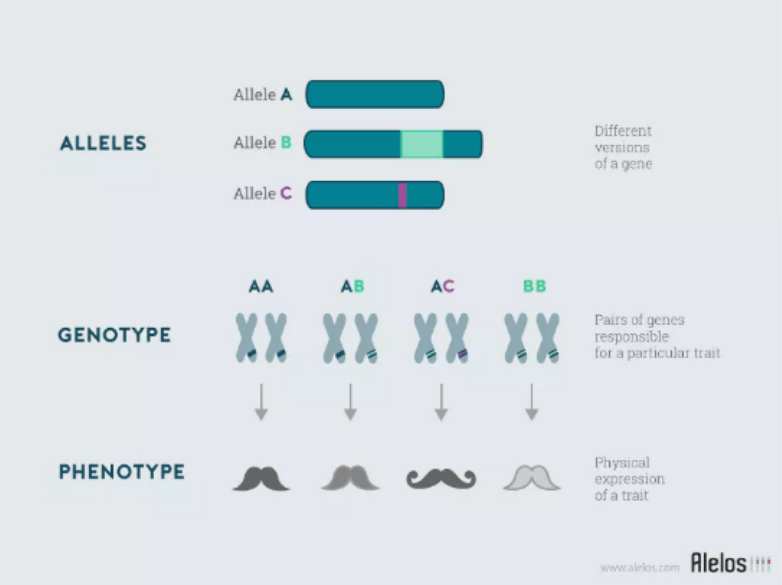
- Basic unit of heredity
- A segment in DNA describes how to make a certain protein
- Genes located at specific locus on the chromosome

Alleles

- Different forms of the same gene
- Differs from other alleles by only a few bases
- simple traits such as eye color may be caused by the interaction of only one pair of alleles

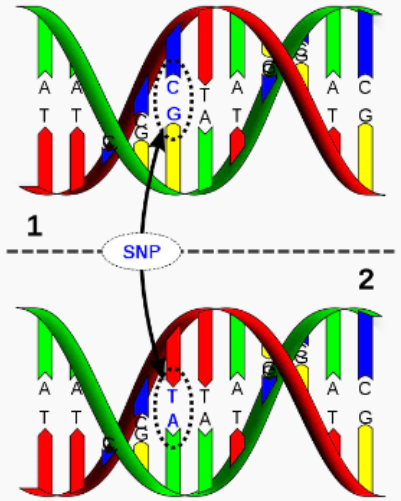
Biological Background

Part 2



Phenotype

- A description of your actual physical characteristics
- Visual Characteristics
- Diseases or allergies



SNP

- Variations in genetic sequence
- Occurs inside a gene
- it must occur to 1% of population to characterise as SNP
- A particular SNP may not cause a disorder, some SNPs are associated with certain diseases.

Feature Extraction

Data Description

First Dataset

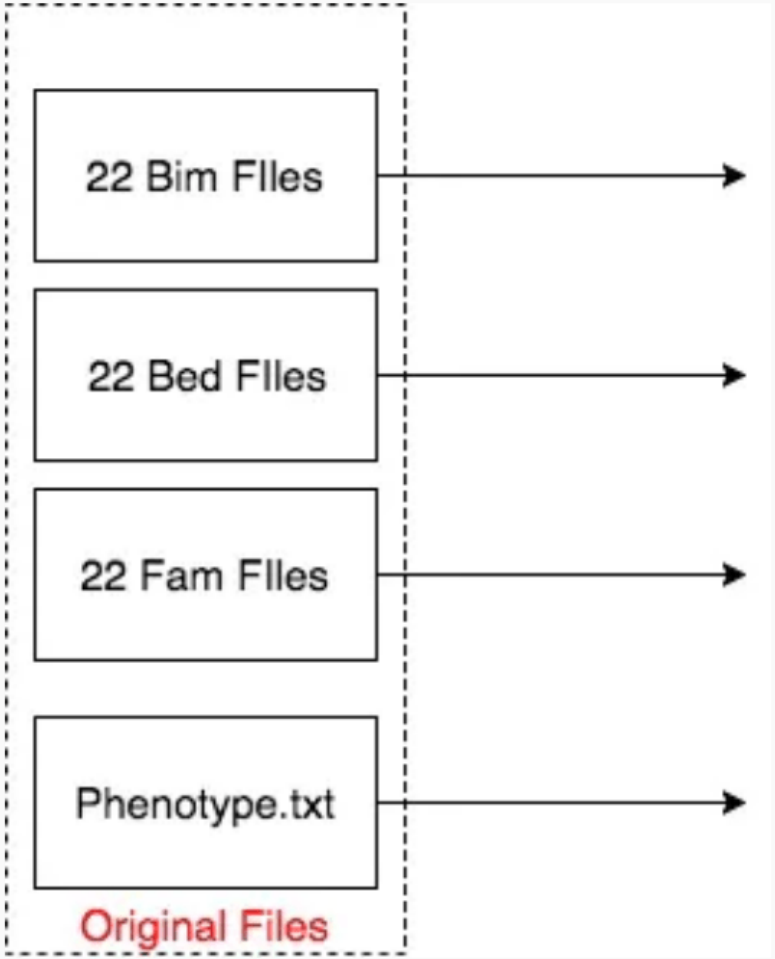
- 22 Chromosomes into 3 files type(bim,bed,fam)
- Size : 2.53TB
- 4980 columns and 7799 rows.
- Balanced already

Second Dataset

- 22 Chromosomes into 3 files type(bim,bed,fam)
- Size : 11TB
- 272176 rows and 12002 columns
- 5% of the population had the disease

Initial data

Plink
Is a tool which process genetic data in different formats



- 1 for each Chromosome
- for each SNP his Minor and Major allele

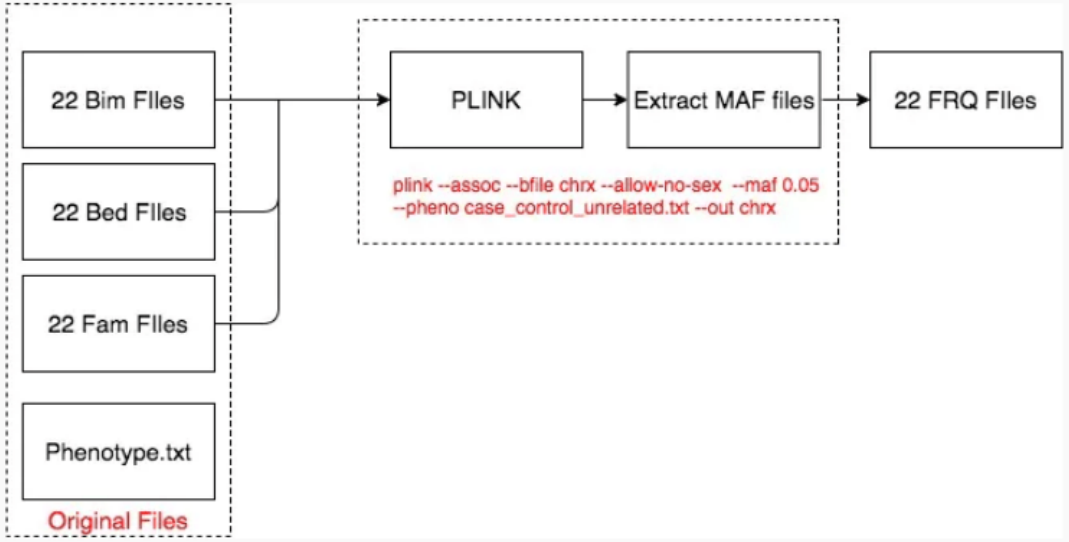
- 1 for each chromosome
- Binary encoded
- For each SNP each patient with his or her genetic sequence

- 1 for each chromosome
- for each patient has his family and phenotypic information
- In our case pivot table for Phenotype.txt

- Our population
- has(1) or not(0) the disease

Step 1 MAF Pruning

MAF or else Minor Allele Frequency is a metric that with the help of PLINK calculates the frequency of the minor allele on the given population

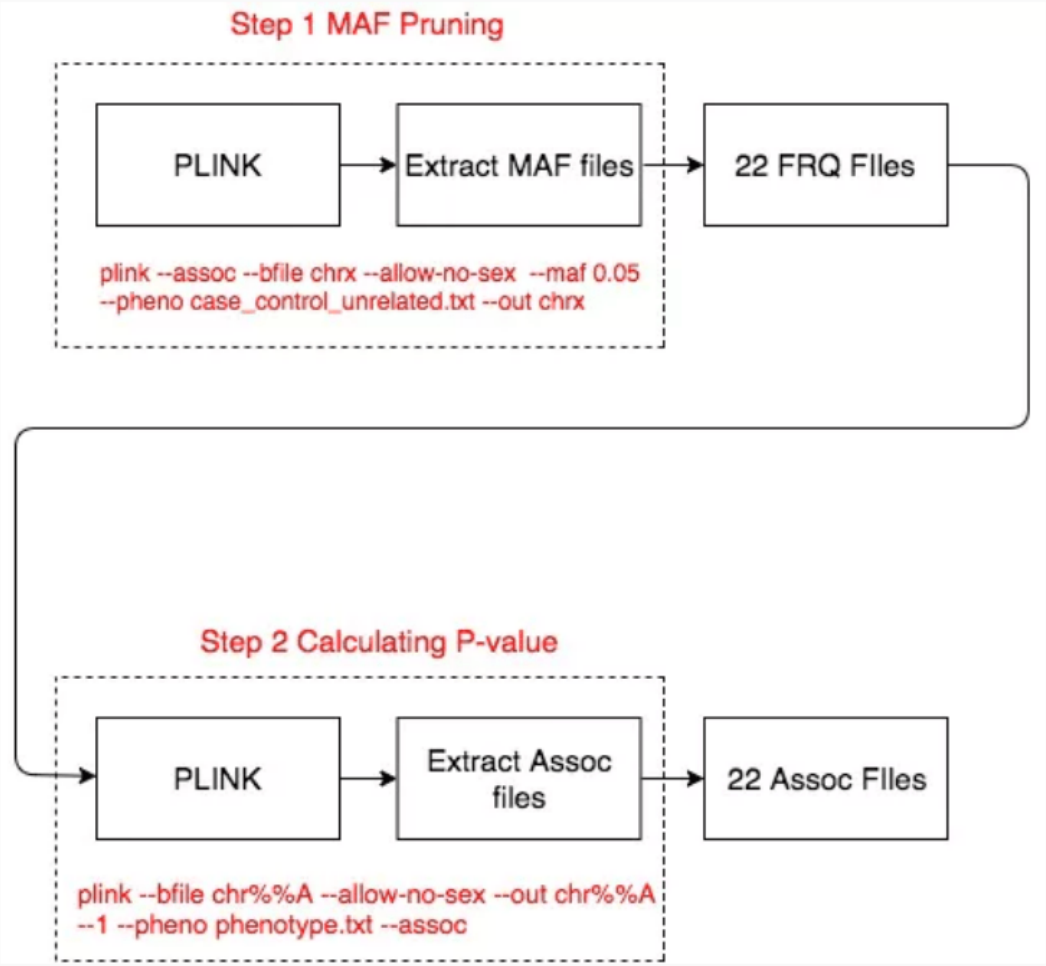


We used a threshold for MAF of 0.05, which means that we exclude from our dataset the SNP's that exist in under the 5% of our population.

FRQ

- 1 for each Chromosome
- For each SNP at each CHR his minor allele frequency

Step 2 Calculating P-value

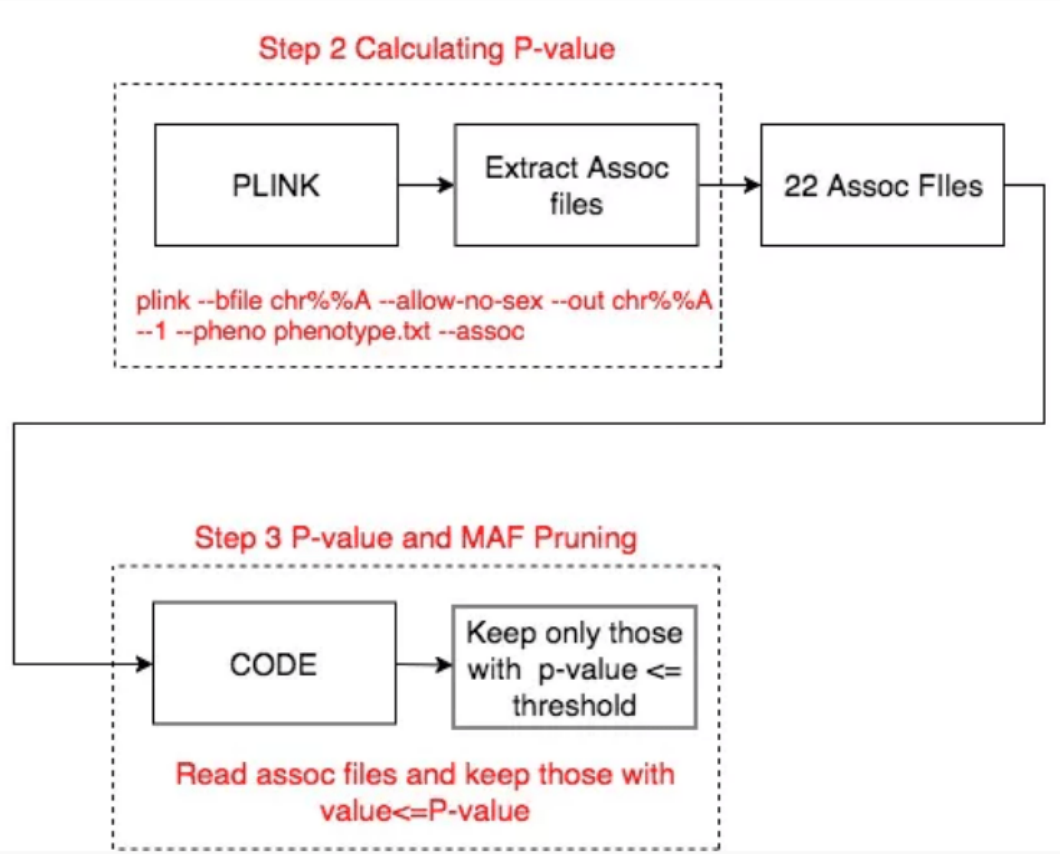


- P-value is a metric which exclude the random possibility of an effect to occur.
- The goal is to reject the null Hypothesis
- In genetics the null hypothesis is that all the SNP's has not a visible linkage between them
- We use P-value of 0.001

ASSOC

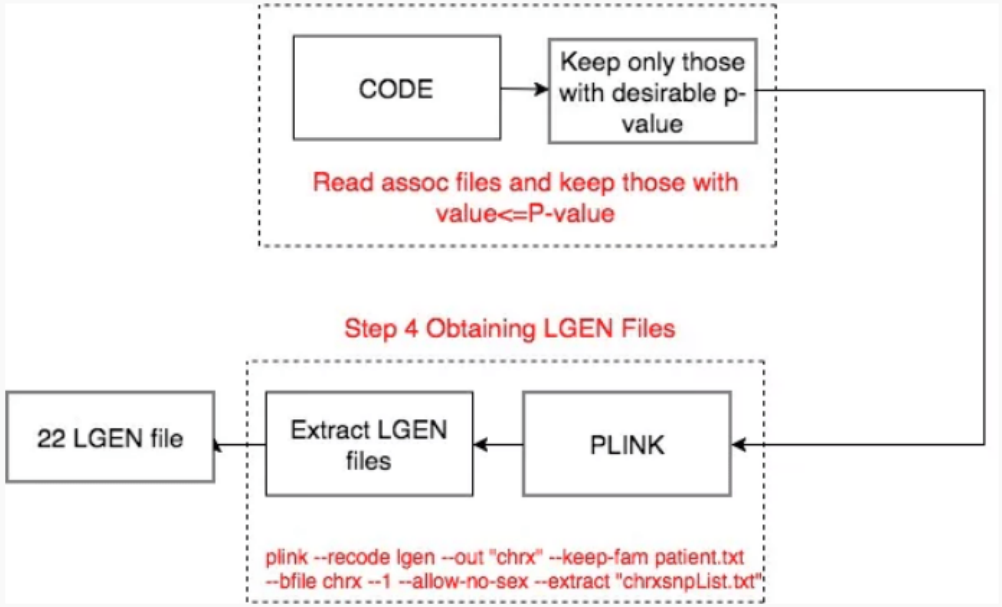
- 1 for each chromosome
- For each SNP his P-value based on given population

Step 3 P-value & MAF Pruning



1. Keep only those SNP's with P-value<=0.001
2. Remove those SNP's with MAF<=0.05

Step 4 Obtaining LGEN files

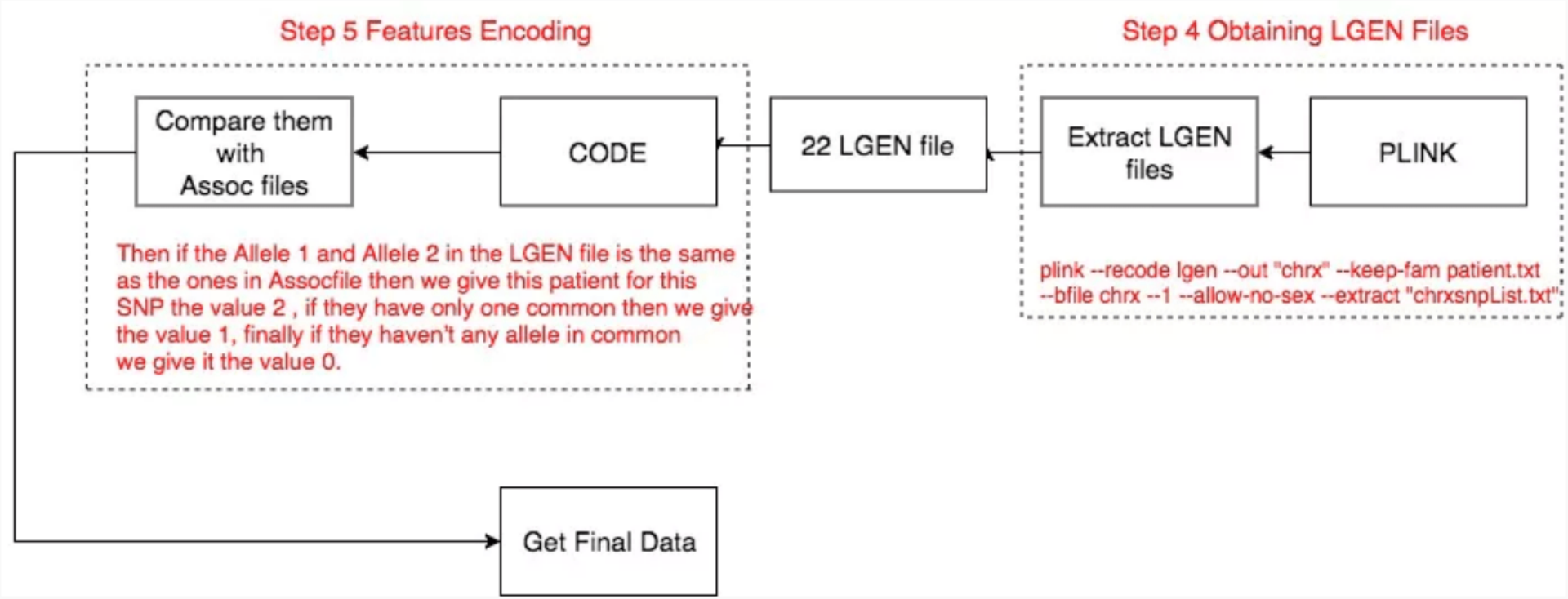


LGEN

- For each patient in each SNP allele 1 and allele 2

Step 5 Feature Encoding

Part 1



Step 5 Feature Encoding
Part 2

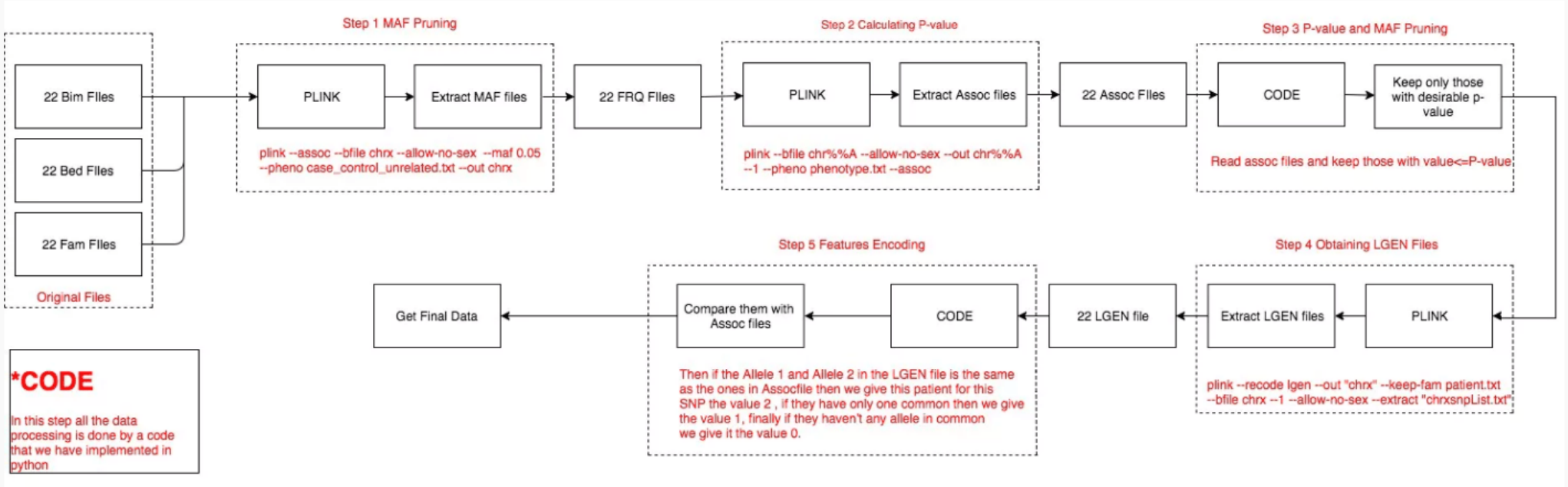
Using Code we compare LGEN file with the pruned ASSOC. We compare allele 1 and allele 2 of LGEN with minor allele of ASSOC

- If both allele 1 and allele 2 are in commom we give value 2
- If the have only one in common we give value 1
- if none we give value 0

Data after Processing

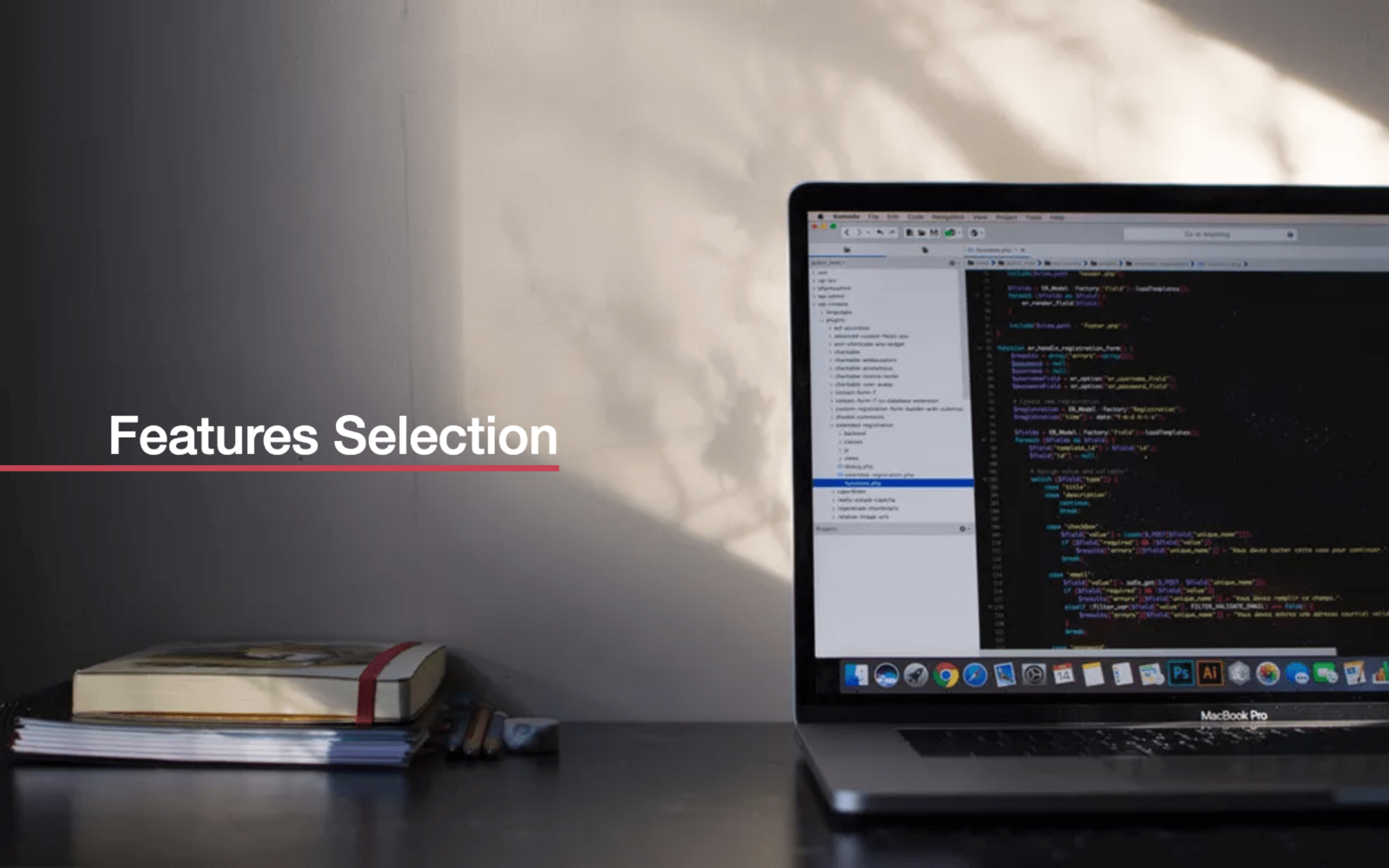
patients	rs143167631	rs1341335	rs2794787	rs12081541	rs774739407	rs1811132	rs17531468	rs12047467	rs7553896	rs12135707	rs2048425	rs561254159	rs7551128	rs539368	rs141292640	rs190960804	rs554899631	rs478406
2689783	0	0	1	0	1	2	1	0	0	0	0	0	1	2	0	1	0	1
2645149	1	2	2	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0
5902679	0	0	1	0	0	1	0	2	2	2	0	0	1	0	0	1	0	0
4906689	0	1	2	0	0	0	0	1	1	1	0	0	1	1	0	1	0	1
1245469	0	0	2	0	1	0	1	1	1	1	0	1	1	1	0	1	0	1
1720573	0	1	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	2
4738597	0	1	1	0	0	0	0	0	0	0	0	1	1	2	0	1	0	0
3302476	1	2	0	0	0	2	0	2	2	2	1	1	1	1	0	1	1	0
4885081	0	0	0	0	0	2	0	0	1	0	0	0	0	1	0	0	0	0
2874787	0	2	0	1	0	0	0	1	1	1	2	0	1	0	0	0	0	2
5581953	0	2	1	1	0	2	0	0	0	0	1	1	0	0	0	1	0	0
1824373	0	0	1	0	0	1	0	1	1	1	0	0	0	1	1	2	0	1
1373593	1	0	1	1	0	1	0	2	2	2	0	1	1	0	2	0	1	1
3771185	0	1	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
4968745	1	0	0	0	0	1	0	1	1	1	0	1	0	1	0	1	1	0
4227879	1	0	2	0	1	0	1	1	1	1	0	1	0	0	1	0	1	1
2196488	1	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	1	0
2683081	0	2	2	0	0	1	0	1	1	1	1	1	0	0	0	0	0	2
1863636	0	2	0	0	0	0	0	0	0	0	1	0	0	2	0	0	0	0
2832722	0	1	1	0	1	1	1	2	2	2	0	1	0	1	0	1	0	0
2257537	0	0	1	1	0	0	0	1	0	1	0	0	0	1	1	1	0	1
3863681	0	0	1	0	0	0	0	0	0	0	0	0	1	2	0	0	0	1
1426442	0	1	0	0	0	1	0	1	1	1	0	1	0	0	0	0	0	0
1911548	1	0	1	1	0	1	0	1	1	1	0	0	0	0	0	0	1	0
1665343	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
2553477	1	1	0	0	2	0	2	0	0	0	0	2	0	0	0	0	1	1
3992068	0	0	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
4122992	0	1	0	0	0	1	0	1	1	1	1	0	1	1	0	1	0	0

Pipeline



Steps to process orgininal data

Features Selection



Similarity

- Cramer V Test

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

- Pearson Correlation

The correlation coefficient is a metric that measures the dependence of two variables X and Y and correlation takes values between -1 and 1, where 1 means perfect correlation, while -1 means negative correlation.

Feature Selection

- K-NN Outlier Detection

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 them if a is high or equal to a threshold θ

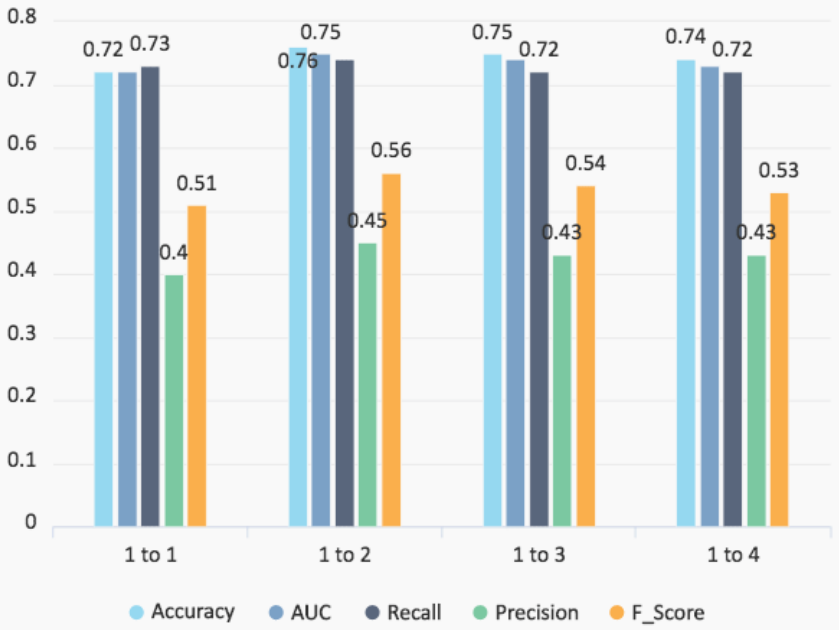
- Algorithm 2

We compute the Pearson Correlation for all pairs, where N is the number of SNPs. For a chosen threshold, we keep the SNPs which do not have correlation coefficient higher or equal to with any other SNP or with a percentage (we defined) of SNPs

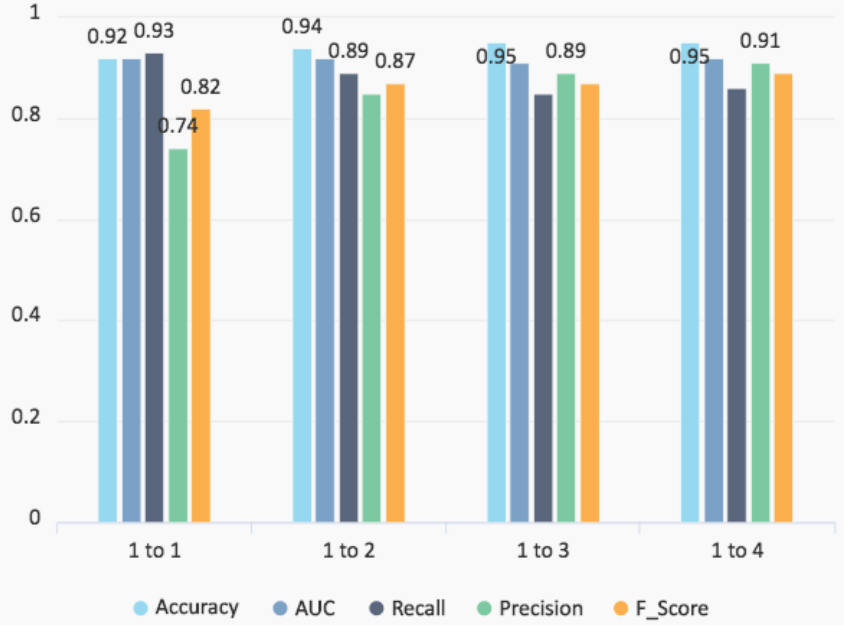
Results

Based On Balance

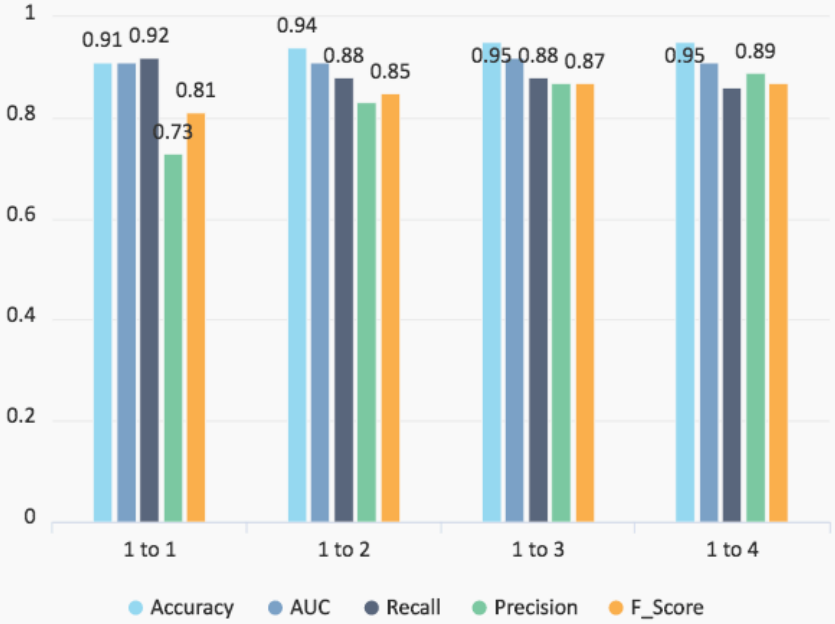
Naive Bayes



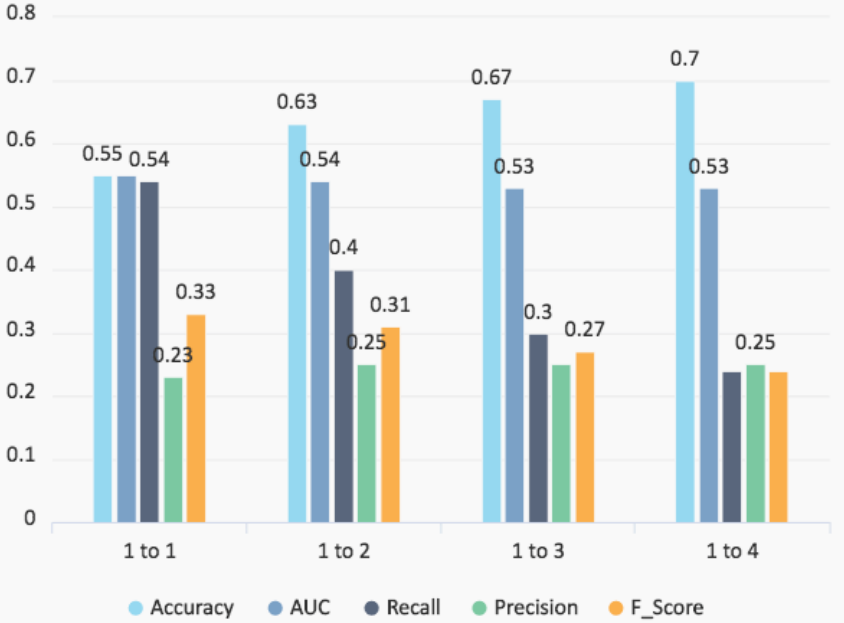
Logistic Regression



SVM



Decision Tree

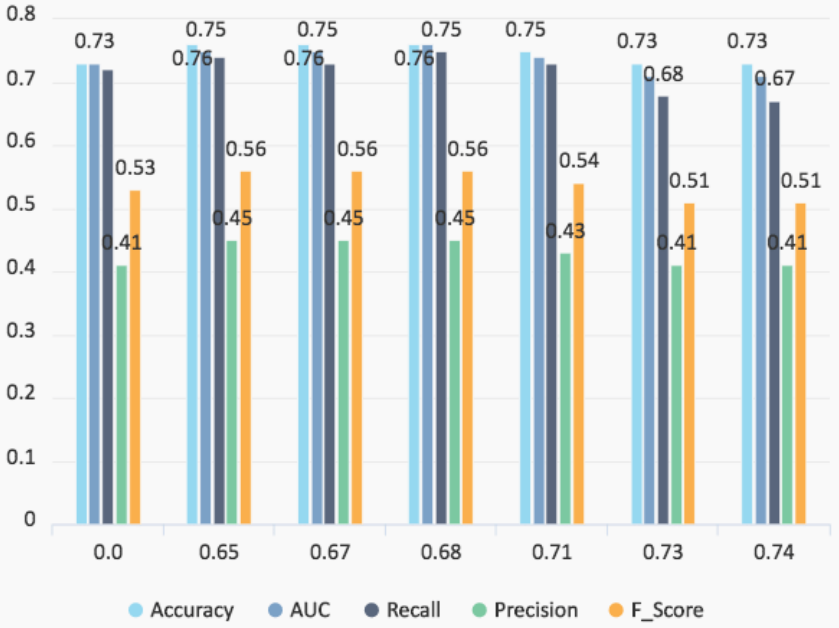


- Logistic Regression and SVM classifiers can predict much better than Bayes Bernoulli and Decision Tree classifiers
- Balance 1 to 2 is better than the others
- The recall is very good, so we can predict if someone is really healthy.
- Precision is also very good, we can predict if someone is patient.

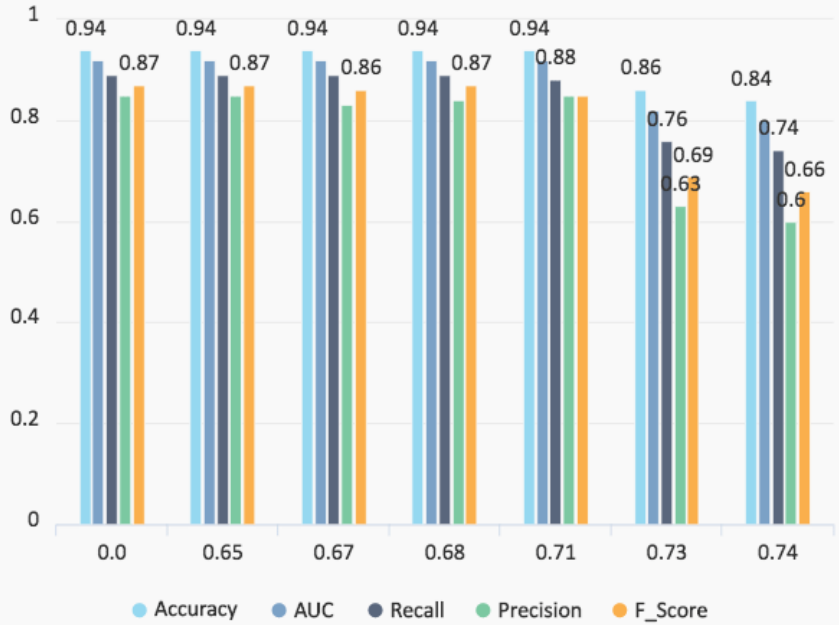
K-NN Outlier Detection

Based on thetas

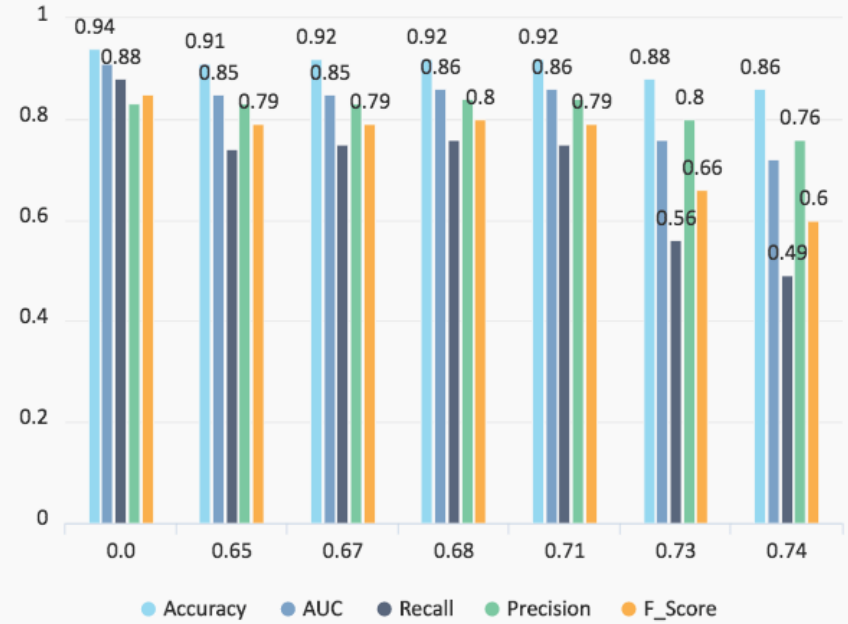
Naive Bayes



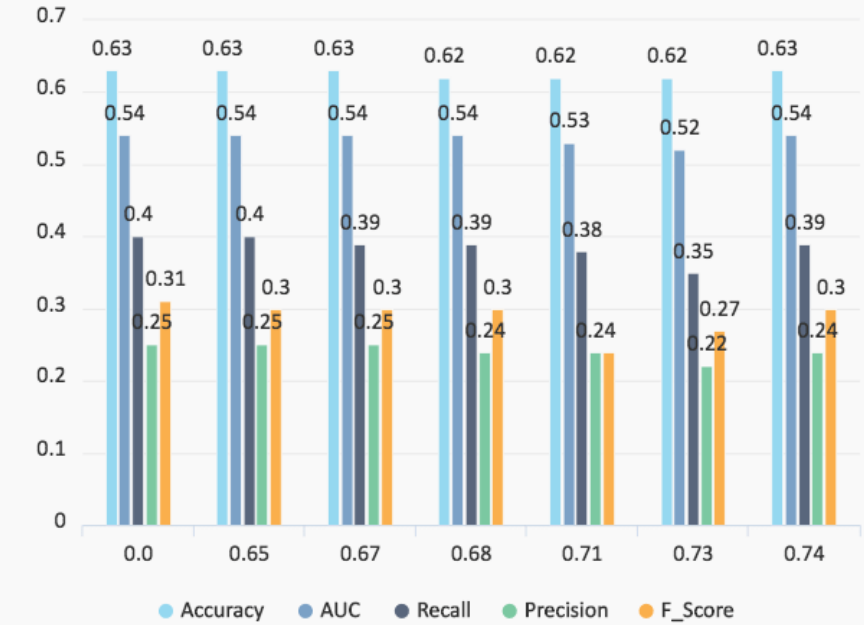
Logistic Regression



SVM



Decision Tree

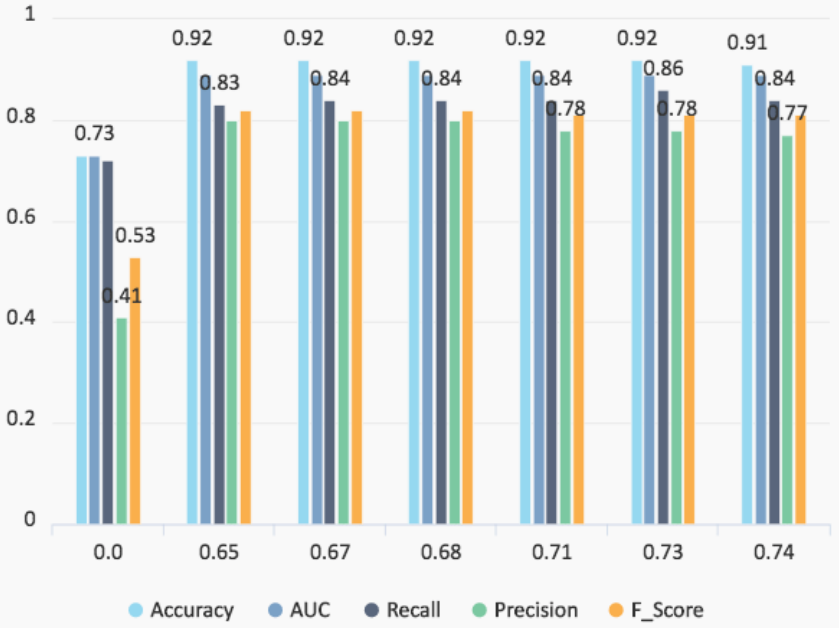


- From 7799 SNPs after feature selection remain 6621
- The results are the same.
- Feature selection does not improve classifier performance
- The recall is very good, so we can predict if someone is really healthy.
- Precision is also very good, we can predict if someone is patient.

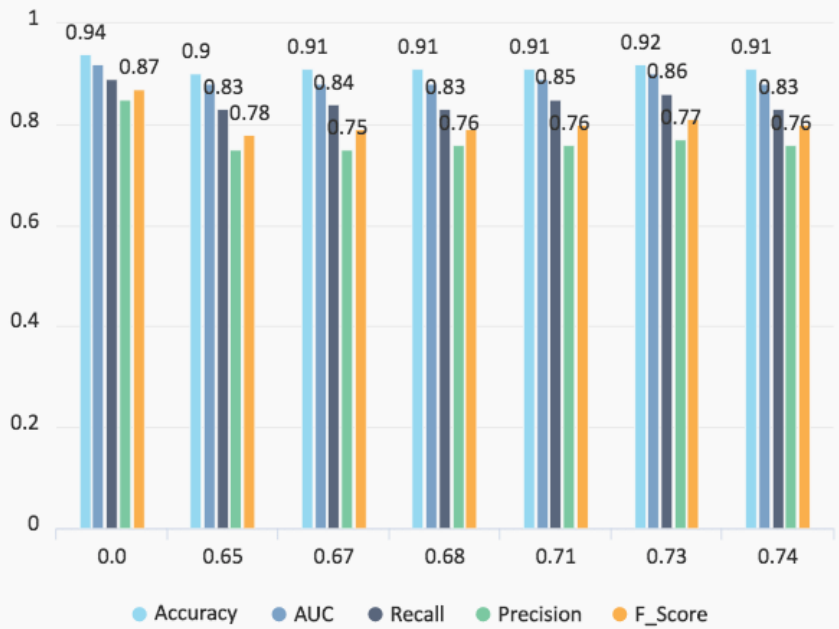
Algorithm2

Based on thetas

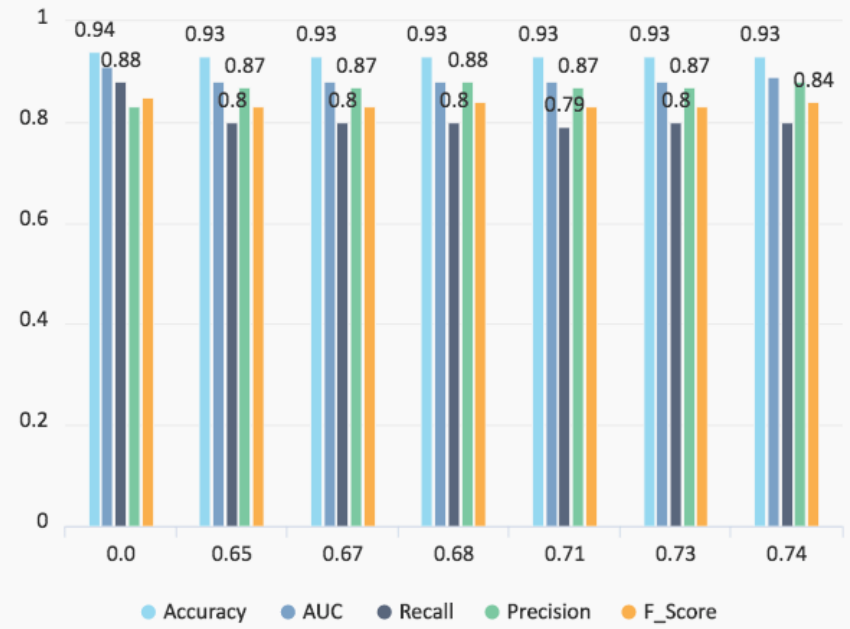
Naive Bayes



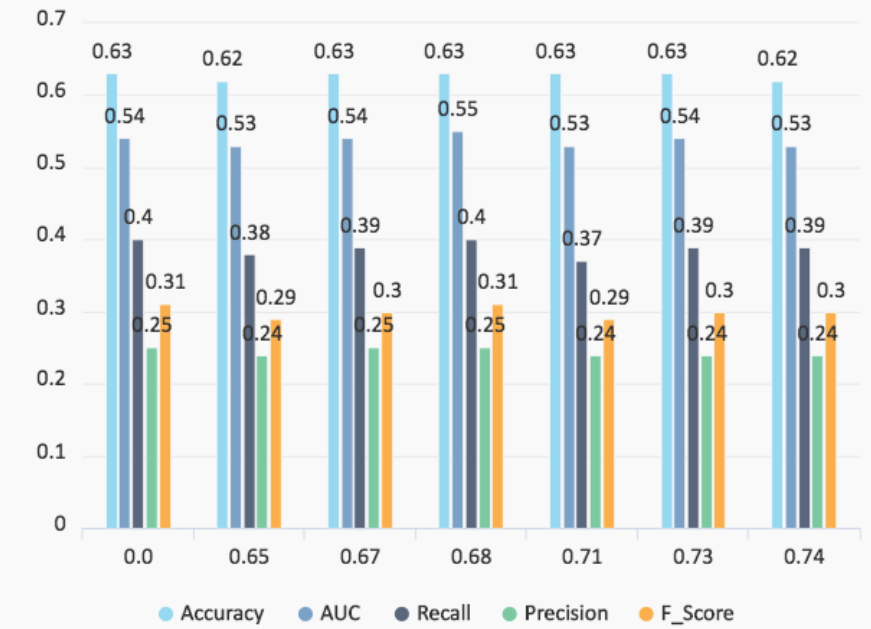
Logistic Regression



SVM



Decision Tree

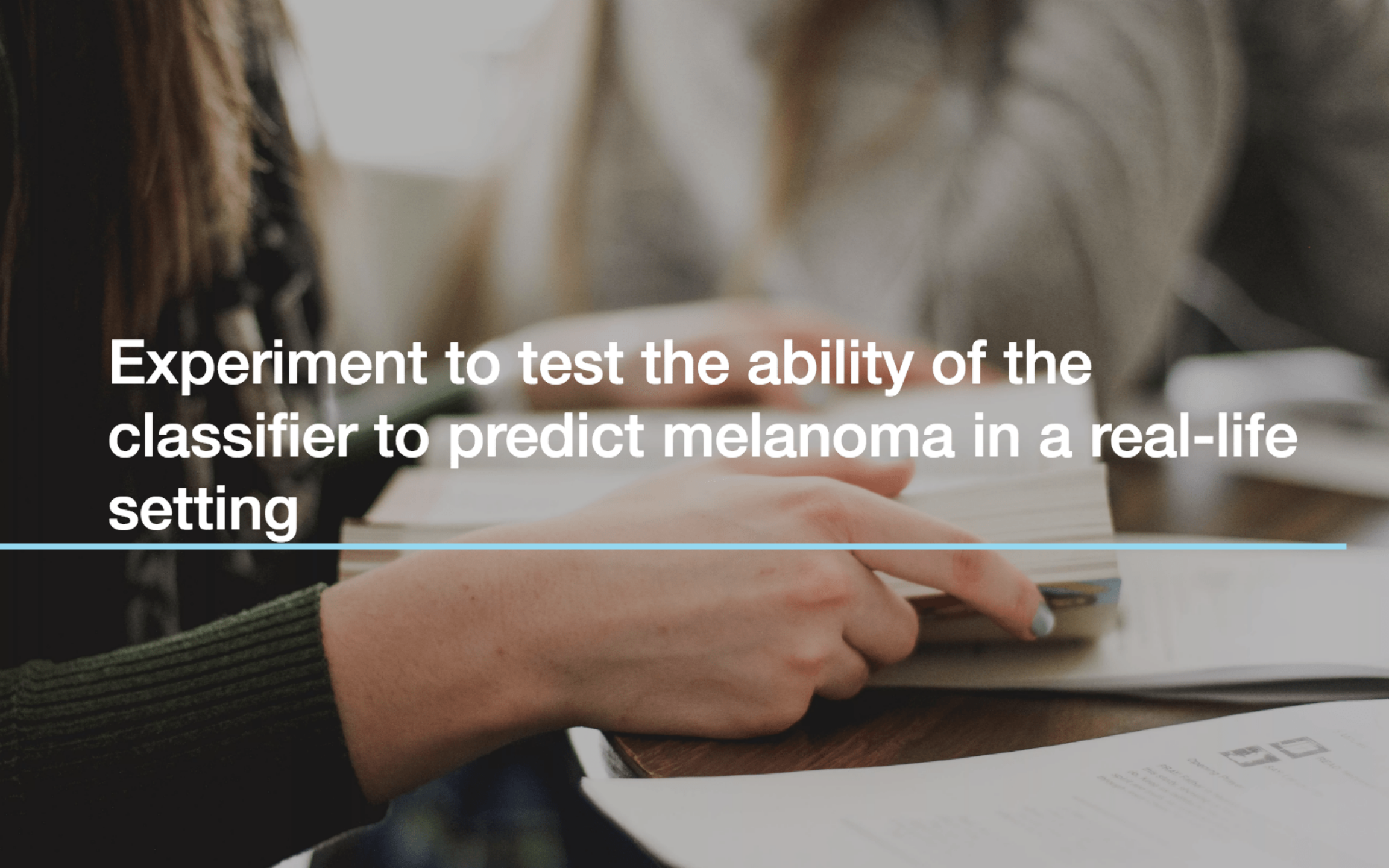


- From 7799 SNPs after feature selection remain 883
- The results are αλμοστ the same.
- Feature selection does not improve classifier performance
- The recall is very good, so we can predict if someone is really healthy.
- Precision is also very good, we can predict if someone is patient.
- Prediction ability of Naive Bayes improved.

Summarize

In this experiment, we conclude that:

- Neither K-NN Outlier detection nor Algorithm2 improves classifiers' performance.
- That we succeed is that we take the same results with fewer features.
- With Algorithm2 prediction ability of Naive Bayes improved.
- The best balance is 2 healthy for one patient.



Experiment to test the ability of the classifier to predict melanoma in a real-life setting

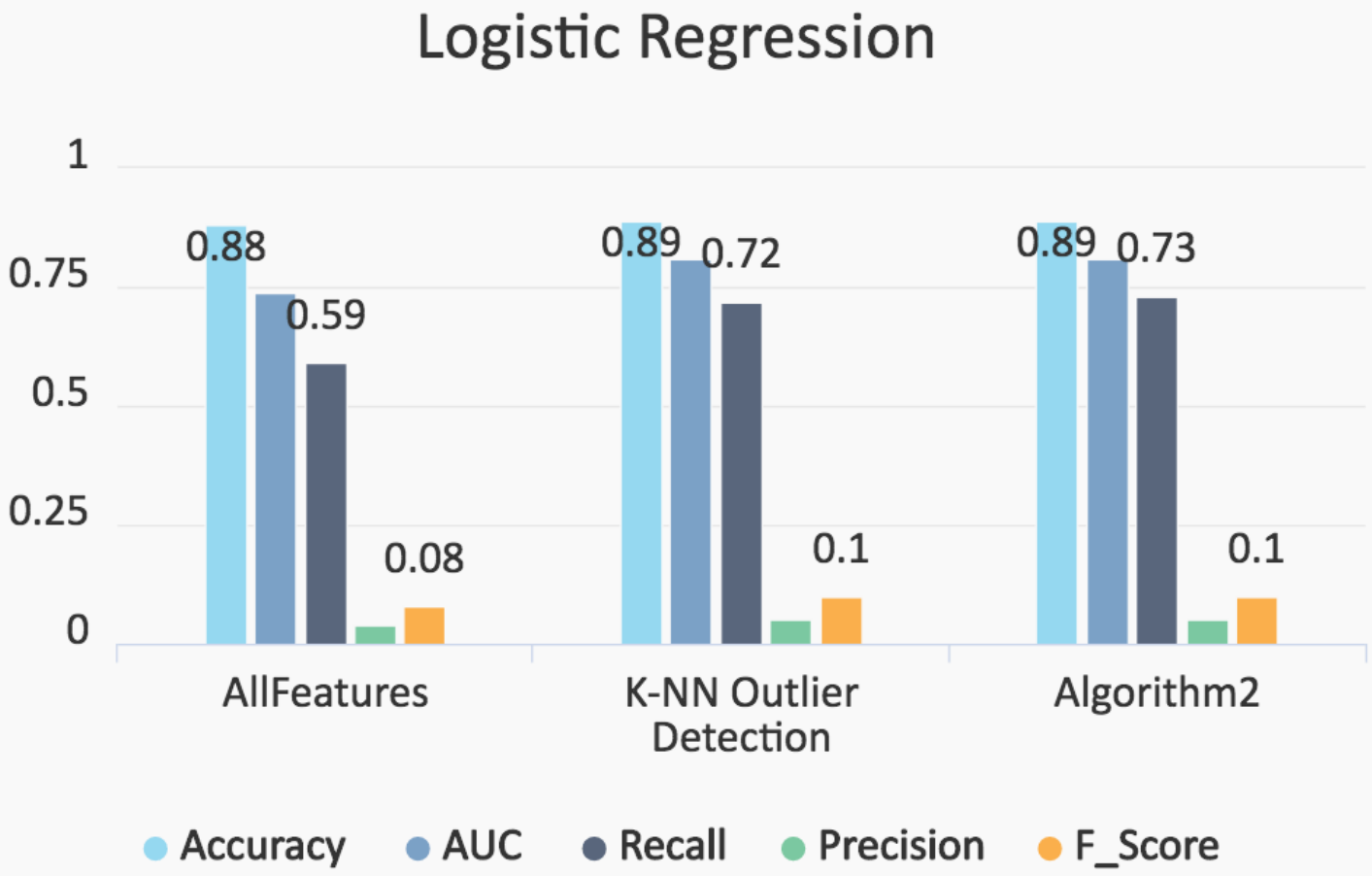
Data

- 272176 people
- 270020 healthies
- 2156 patients

- The difference between patients and healthy people is a problem because classifiers predict only the negative class.

- We solve this problem by balancing people. We keep only two healthy people for every patient.

Results



The low value of precision may be due to the big difference between patients and healthy people.


Summarize

In this experiment, we conclude that:

- feature selection improves algorithm performance
- Recall increased by 0,13
- If algorithm predicts the negative class the with high probability is a really negative class

Conclusion

- Our results show that the classifiers work well in a controlled setting of balanced 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

The End

Thank you