Contents

[Figure List 2](#_Toc42379600)

[Table List 2](#_Toc42379601)

[Introduction 3](#_Toc42379602)

[System Configuration 3](#_Toc42379603)

[Data Preprocess 4](#_Toc42379604)

[Step I – Data Cleaning 4](#_Toc42379605)

[Step II – Selecting Top Genes by Class 4](#_Toc42379606)

[Model Selection 6](#_Toc42379607)

[Step III - Finding the best classifier/best gene set combination 6](#_Toc42379608)

[*Naïve Bayes Classifier* 6](#_Toc42379609)

[*Decision Tree Classifier* 7](#_Toc42379610)

[*Random Forest Classifier* 7](#_Toc42379611)

[*Neural Network (Multi-Layer Perceptron) Classifier* 8](#_Toc42379612)

[*K-Nearest Neighbor (KNN) Classifier* 8](#_Toc42379613)

[Evaluate 9](#_Toc42379614)

[Step IV 9](#_Toc42379615)

# Figure List

[FIgure 1 - Gene distribution over classes 4](#_Toc42379504)

[FIgure 2 - Classifier Accuracies 9](#_Toc42379505)

# Table List

[Table 1 - Some T-Test Results 5](#_Toc42379574)

[Table 2 - Test Dataset Evaluation Result 9](#_Toc42379575)

Section I

# Introduction

Studies on gene sequences and distributions and the use of these studies in disease treatment have gradually increased in the recent past. Machine learning and, more specifically, deep learning models with increased computational power to produce faster results have added speed to the studies.

Although unfortunately, traditional machine learning models have lost their former popularity with the start of digital data becoming easily accessible and processable. It is currently used to solve most problems. Since deep learning models need more data than other models, machine learning results more efficiently and faster than deep learning, in areas that have trouble with the amount and variety of data.

In this study, we will use machine learning and deep learning models to test this hypothesis, while finding the relationship between gene sequences and diseases of some of the patients we have.

Classifiers used in the project:

* Naïve Bayes classifier,
* Decision Tree classifier,
* K-NN classifier,
* Random Forest classifier and,
* Neural networks

These classifiers have been preferred as they are often used in most models.

# System Configuration

Python's scikit-learn library was used mainly for all machine learning models, while Ubuntu 20.04 was used as the operating system. All required libraries are specified in the "requirements.txt" file.

Section II

# Data Preprocess

## Step I – Data Cleaning

In the first step, we will use the threshold method to examine the data more efficiently, so we need to select the data we have to be within a specific range. The reason we do this is that we do not want small values that will be ineffective or values that will form peaks in the value distribution in the model entry. Here, we do the threshold by not considering values less than 20 or greater than 12000. Besides, we need to clear the data according to the minimum-maximum values and the variance of the data. For this, we also eliminate genes whose max/min ratio is less than 3 and whose mean-variance of the data is less than 20%.

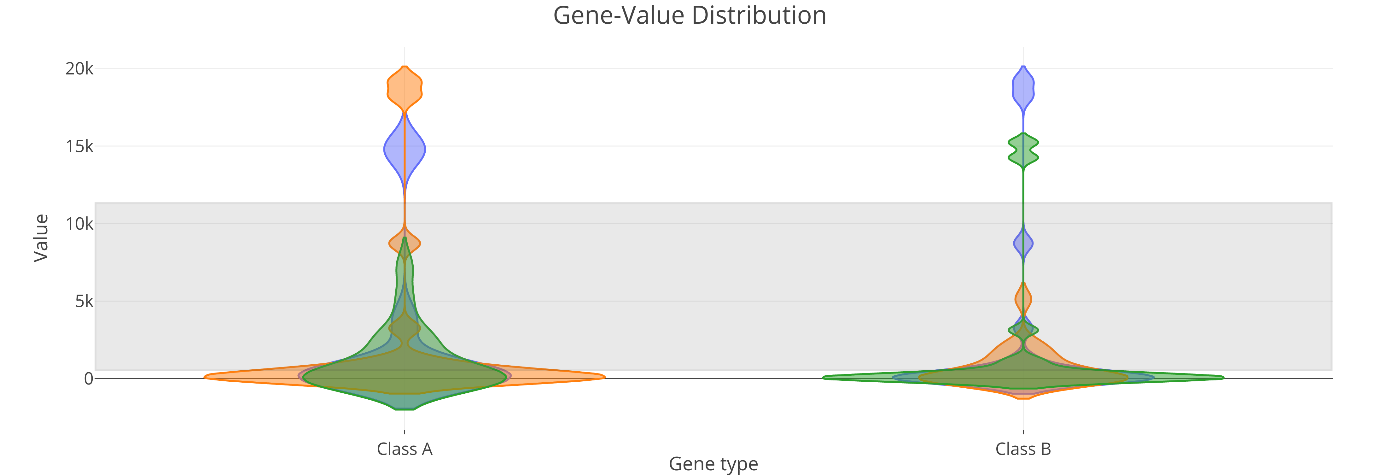


FIgure 1 - Gene distribution over classes

## Step II – Selecting Top Genes by Class

This step is the essential step in terms of selecting the data to enter the models, affecting the performance of the model and training time. By statistically testing the relationship of gene values belonging to disease classes with each other using the "Student t-Test" method, we need to select the best gene pairs with the lowest p and highest absolute t values as a result. Some results are shown in Table-1 below.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Class** | **Gene 1** | **ID Gene 1** | **Gene 2** | **ID Gene 2** | **t-Value** | **p-Value** |
| EPD | AB000409\_at | 4 | AB000468\_at | 9 | 5,6017E+15 | 5,82269E+15 |
| EPD | AB000409\_at | 4 | AB000895\_at | 10 | 1,4291E+16 | 1,70101E+15 |
| EPD | AB000409\_at | 4 | AB001325\_at | 15 | 7,28775E+15 | 900646423,6 |
| EPD | AB000409\_at | 4 | AB002559\_at | 22 | -3,65802E+15 | 7,18776E+15 |
| EPD | AB000409\_at | 4 | AB003102\_at | 23 | -2,77441E+16 | 1,25032E+16 |
| EPD | AB000409\_at | 4 | AB004884\_at | 27 | 4,72747E+15 | 1,6803E+15 |
| EPD | AB000409\_at | 4 | AC000061\_cds2\_at | 29 | -1,47741E+15 | 8,84189E+14 |
| EPD | AB000409\_at | 4 | AC002045\_xpt1\_at | 33 | 5,81984E+15 | 16333009531 |
| EPD | AB000409\_at | 4 | AC002073\_cds1\_at | 34 | 1,05046E+16 | 4164766,95 |
| EPD | AB000409\_at | 4 | AC002077\_at | 35 | 2,26991E+15 | 3,5735E+16 |
| EPD | AB000409\_at | 4 | AC002115\_cds1\_at | 36 | -8,16635E+15 | 1825831454 |
| EPD | AB000409\_at | 4 | AD000092\_cds1\_at | 40 | -2,16639E+16 | 4,39479E+15 |
| EPD | AB000409\_at | 4 | AF000177\_at | 44 | 9,05873E+15 | 3,7697E+15 |
| EPD | AB000409\_at | 4 | AF000231\_at | 45 | -1,10963E+16 | 2,81765E+15 |
| EPD | AB000409\_at | 4 | AF002700\_at | 54 | 7,61952E+15 | 487079913,8 |
| EPD | AB000409\_at | 4 | AF003743\_at | 55 | 9,49106E+15 | 198300572,5 |
| EPD | AB000409\_at | 4 | AF006609\_at | 64 | 6,30163E+16 | 6112558178 |
| EPD | AB000409\_at | 4 | AF014958\_at | 74 | 9,19167E+15 | 32146357,52 |
| EPD | AB000409\_at | 4 | AF015910\_at | 75 | -3,37472E+15 | 3,37519E+16 |
| EPD | AB000409\_at | 4 | AF015950\_at | 77 | -4,38752E+16 | 3,55134E+16 |
| EPD | AB000409\_at | 4 | AJ000480\_at | 78 | -6,5214E+15 | 3943938876 |
| EPD | AB000409\_at | 4 | D00632\_at | 84 | 1,1798E+16 | 9,0739E+14 |
| EPD | AB000409\_at | 4 | D00763\_at | 91 | -1,14943E+16 | 10060291,68 |
| EPD | AB000409\_at | 4 | D10511\_at | 94 | 3,73706E+15 | 1,50883E+15 |
| EPD | AB000409\_at | 4 | D10522\_at | 95 | -7,65614E+15 | 4555426975 |

Table 1 - Some T-Test Results

The stages followed when achieving these results are respectively:

1. We are separating gene samples according to classes. So, we can find the best gene for each class,
2. Putting a selected gene into the test with all the remaining genes, respectively.
3. It is passing on the next gene to continue until there are no genes left that have not been tested.
4. We are saving gene indices, t-values, and p-values.
5. Repeat the test until all classes are finished by moving to the next class.
6. We repeat the above steps to take the best 2,4,6,8,10,12,15,20,25,30 genes, respectively.
7. Saving all best-N train datasets.

Section III

# Model Selection

## Step III - Finding the best classifier/best gene set combination

### *Naïve Bayes Classifier*

The Bayesian theorem is a critical subject studied within probability theory. This theorem shows the relationship between conditional probabilities and marginal probabilities in a probability distribution for a random variable.

* - the probability of event A occurring when event B occurs.
* - the probability that event A will occur.
* - the probability of event B occurring when event A occurs.
* - the probability that event B will occur.

The basis of the Naïve Bayesian classifier is based on the Bayesian theorem. It is a lazy learning algorithm that can also work on unbalanced datasets. The way the algorithm works is to calculate the probability of each state for an element and classify it according to the highest probability value. He can do very well with a little education. If a value in the test set has a value that cannot be observed in the training set, it returns 0 as the probability value, meaning it cannot predict. This condition is commonly known as Zero Frequency. Correction techniques can be used to resolve this situation. One of the most straightforward correction techniques is known as Laplace estimation.

Examples of uses include real-time prediction, multi-class prediction, text classification, spam filtering, sensitivity analysis, and suggestion systems.

### *Decision Tree Classifier*

The decision tree algorithm is one of the data mining classification algorithms. They have a predefined target variable. In their structure, they offer a strategy that goes from the top to the bottom.

In decision tree learning, during the learning of the tree, the set studied on it is divided into subsets according to various properties, this process is repeated recursively. It lasts until the repetition operation does not affect the prediction. This process is called recursive partitioning. In general, the way data comes in during data mining is as follows:

According to this notation, values from to are inputs of the system, whereas the value is the desired output of the system.

In data mining, decision tree learning is used for two primary purposes. The specific names for these purposes and decision tree learning for this purpose are as follows:

* Classification problems
  + Classification Tree: Problems that aim to place the target set in one of the various classes, such as the estimation of a person's level of education from their maps, and make the class definition.
* Principle problems:
  + Regression Trees: As a result, data mining problems returning a numeric value instead of a class.

### *Random Forest Classifier*

The random forest algorithm, which is an ensemble learning method, is an algorithm that aims to increase the classification value by generating multiple decision trees during the classification process. Individually created decision trees come together to form a decision forest. The decision trees here are randomly selected subsets of the data set to which they are connected.

Random forest method is unique in known machine learning methods that provide predictive validity and model interpretability. Random sampling and to include improved features of techniques in community methods due to random forest method offers better generalizations, and current estimates can be found. Reasons for the certainty of predictions of the random forest method bias give low results and is a low correlation between trees. Low the amount of bias is derived from the creation of relatively large trees. Low correlation by creating as many different trees as possible a community is obtained in its structure.

### *Neural Network (Multi-Layer Perceptron) Classifier*

Multi-layer Perceptron is a supervised learning algorithm that learns a function:

By training on a dataset, where  is the number of dimensions for input and  is the number of dimensions for output. Given a set of features:

And a target , it can learn a non-linear function approximator for either classification or regression. It is different from logistic regression, in that between the input and the output layer, there can be one or more non-linear layers, called hidden layers.

### *K-Nearest Neighbor (KNN) Classifier*

The KNN algorithm is one of the easy-to-implement supervised learning algorithms. Although it is used in the solution of both classification and regression problems, it is mostly used in the industry in the solution of classification problems.

When calculating the distance of the new data that will be included in the sample data set according to the existing data, the nearest neighbor of the K number is looked at. 3 types of distance functions are used for distance calculations:

* Euclidean Distance
* Manhattan Distance
* Minkowski Distance.

KNN is one of the most popular machine learning algorithms because it is resistant to Old, noisy, and straightforward educational data. However, there are also disadvantages. For example, because it stores all states when calculating distance, it requires an ample memory space when used for extensive data.

Section IV

# Evaluate

## Step IV

You can see the training results in the figure below.

sabit, kalem içeren bir resim

Açıklama otomatik olarak oluşturuldu

FIgure 2 - ClassIfIer AccuracIes

The best classifier selected is the Random Forest Classifier, and the best gene groups are those in their 30s.

Classification of the test data prepared using the 30 best gene groups with the random forest classifier after training with the training data:

|  |  |  |
| --- | --- | --- |
| Sample: 0 - MED | Sample: 8 - MGL | Sample: 16 - MED |
| Sample: 1 - EPD | Sample: 9 - RHB | Sample: 17 - EPD |
| Sample: 2 - MED | Sample: 10 - RHB | Sample: 18 - EPD |
| Sample: 3 - MED | Sample: 11 - MED | Sample: 19 - MED |
| Sample: 4 - MED | Sample: 12 - EPD | Sample: 20 - MED |
| Sample: 5 - MED | Sample: 13 - MED | Sample: 21 - MED |
| Sample: 6 - MED | Sample: 14 - MED | Sample: 22 - MED |
| Sample: 7 - EPD | Sample: 15 - MED |

Table 2 - Test Dataset EvaluatIon Result