UNIVERSITY OF PIRAEUS - DEPARTURE OF DIGITAL SYSTEMS

Heart disease prediction system based on SVM classifier

UNIVERSITY OF PIRAEUS

SUPERVISOR: Dimosthenis Kyriazis

Andreas Priftis

E15129

Contents

List of abbreviations

SVM: Support Vector Machine

WHO: World Health Organization

GA: Genetic Algorithm

KNN: K-Nearest Neighbor

AI: Artificial Intelligence

UCI: University of California, Irvine

cp: chest pain type

trestbps: the resting blood pressure

chol: cholesterol

fbs: fasting blood sugar

restecg: resting electrocardiographic

thalach: maximum heart rate

exang: exercise induced angina

ca: number of major vessels colored by fluoroscopy

thal: heart rate

ANN: Artificial Neural Network

AIS: Artificial Immune System

PCA: Principal Component Analysis

NA: Not Available

BSD: Berkeley Software Distribution

RBF: Radial Basis Function

ACS: Acute Coronary Syndrome

List of figures

List of tables

Abstract

In the health care field, mortality as a result of heart disease remains extremely high. Based on the specific fact, it is essential to distinguish the most important factors for predicting the risk of death in patients who suffer from this prevalent disease. Therefore, the effort to visualize the medical databases and construct models through soft computing tools that enable us to represent how the principal driving factors relate to patients' outcomes is to a great extent. The purpose of obtaining patterns that are conformed to predictor's variables of the health field and its databases is based on data mining. Existing data mining techniques are utilized to resolve complicated and dynamic procedures. In the specific research, we suggest SVM classifier using several attributes in order to achieve a large percentage of accuracy and effectiveness during the diagnosis of the recurrence of heart disease. The data records, which are used to this study, are obtained from the UCI repository and they are called Cleveland dataset. According to the results provoked by the SVM classifier, we could conclude that this model provides a high predictive accuracy and an increased performance about the reliability of the system. Furthermore, we should mention that due to this work, I was able to develop a heart disease prediction system which is a truly remarkable tool with a user-friendly interface and improved scalability.

1. Introduction

One of the important aspects of medical research is the prediction of unlimited diseases and the analysis of factors that create them. In the specific paper, we deal with heart disease using the Cleveland dataset. A lot of researches have been conducted on this dataset in order to have better predictions. In medical industry, there is a huge availability of data, but people are not able to find out the significant information about the factors that provoke heart disease. For this reason, Data mining intends to help doctors and patients to conduct their diagnosis. It is going to assist doctors to obtain more useful information and predict the disease faster. Simultaneously, it is possible to play an important role in health care industry in order to set up health systems which would be able to use data and analytics in order to detect inefficiencies and improve the current status. The majority of hospitals utilize various information systems to handle tremendous amounts of patient's data. As a result, it is important for the section of health industry to brainstorm how data could be stored, prepared, organized and allocated to organizations in order to improve the benefits of health care data mining systems and applications. Heart disease prediction models, based on data mining methods, have the chance to assist practitioners to detect the current status of the patient and cope with several problems which are prevalent in decision-making whether patients suffer from heart disease or not. One of the most usual and effective data mining classification techniques in order to predict heart disease is called Support Vector Machine (SVM) which is the main algorithm in the specific paper. Last but not least, it is essential to emphasize that according to survey of WHO (World Health Organization), because of workaholism, mental stress and various problems, the total number of deaths due to heart disease is approximately calculated to 17 million in a lot of countries [4]. Diagnosis is a very complex procedure to be conducted and it is important to take everything into consideration accurately and effectively.

In 2016, Berikol et al. attempted to calculate the accuracy of 4 different classifiers (SVM, ANN, Naïve Bayes and logistic regression) in order to predict the ACS (Acute Coronary Syndrome) using 228 records and 8 attributes. The highest accuracy provoked by the SVM classifier with a percentage of 99%, afterward ANN advanced to 90%, Naïve Bayes to 89% and logistic regression to 91%.

In 2011, Kumari & Godara, assessed the performance of 4 different classification algorithms for the purpose of predicting the cardiovascular disease which are the RIPPER, Decision Tree, SVM and ANN. Although there was no preprocessing during the variable selection, the results demonstrated that SVM classifier has better predictive accuracy in a dataset of 296 records and 14 attributes [47].

In the specific study, I make an effort to eliminate the gap between health field and technology especially in prediction process. My objective is to assist medical practitioners and doctors to predict the probability of heart disease to their patients using SVM classifier and data mining techniques. Dataset includes medical and socio-environmental variables in order to make more reliable predictions [2]. Nevertheless, variables that have no association to the predictive process could prevent the training phase and the general capabilities of the algorithm. Moreover, we should mention that models, which are obligated to use all variables, should remove a high number of individuals due to missing data. For this reason, we would use a dataset of 303 processed records and resolve the specific issue. On the other hand, one of the most frequent methods to restrict this problem is to implement a heuristic variable selection that could recognize some variable values effectively.

This paper is structured as follows. In the first section, we would include some basic terms related to data mining in order readers are capable of understanding important definitions. The second section contains the experimental environment and some dataset details. Afterward, we would analyze some common data mining techniques which are used to a great extent in various fields. The next section makes reference to the components of the algorithm and the performance evaluation and comparison between different models. The next to last section includes the user manual of the prediction system. In the Conclusion, final remarks are presented and future works summarized.

2. Related Work

There are several factors that play important role in appearance of heart disease. For this reason, we should discern them in two categories:

1. Controllable risk factors
2. Uncontrollable risk factors

The first category includes smoking, cholesterol, diabetes, weight and blood pressure because people are able to control them effectively. On the other hand, uncontrollable risk factors contain the age, sex and family history of the patient. Nowadays, in health care field, the highest percentage of data is the milestone in order to extract and analyze knowledge that enables to make better decisions and contribute to cost-savings. Based on the above, data mining techniques are coming to resolve many problems with medical data in order to find patterns, make more efficient predictions and classifications. Although data mining methods are integrated late into the health care industry because they are implemented to different scientific fields such as sales forecasting, they have provided a lot of intelligent systems and decision-making applications in order to predict the severity of various diseases, for example heart disease, and remote health monitoring.

A high number of data mining techniques have been developed in recent years in order to predict heart disease. Scientists have utilized unlimited methods and data mining techniques such as classification and clustering to enhance diagnosis of diseases. The aim of this effort is to improve the accuracy of the results and eliminate the error rate in probabilities. The specific literature, in essence, regards that data mining techniques via classification are more effective in heart disease prediction as compared to other equivalent methods and express researcher's need for better accuracy and reliable model in this section of the health care [7].

Yan and Zheng had the opportunity to suggest a real-coded GA-system in order to diagnose various heart diseases while applying important clinical features sub-setting. The aforementioned prediction system was built in aim of diagnosis of five major heart diseases, using 352 heart disease instances and their respective diagnosis weights in case of successful diagnosis of heart disease or not. Furthermore, we should relate that Austin et al, based on machine learning principles and data mining methods, provided alternative classification schemes which was composed of RFs, boosting, aggregation bagging and support vector machines (SVMs), proving that modern data mining techniques provide a lot of benefits in high accuracy [3].

In 2013, Abhishek Taneja et al. proposed a heart disease prediction system utilizing data mining tool WEKA 3.6.4 and implementing various data mining techniques. For instance, J48 technique achieved 95.56% accuracy, Naive Bayes technique achieved 92.42% and Neural

Networks achieved 94.85%. The data set of 7339 instances was collected from PGI, Chandigarh with a total of 15 attributes. Of the 15, only 8 attributes were used [4].

In 2012, Akhil Jabbar et al. developed a real breakthrough in the research for heart disease prediction. They utilized Association Rule Mining implementing clustering in order to predict the heart diseases in Andhra Pradesh population [4].

In 2011, Jyoti Soni et al. conducted a research using renowned data mining techniques like Decision Tree, Neural Network, Naive Bayes, classification via clustering and KNN. The most effective method was proved the Decision Tree. After the preprocessing of the data set utilizing a genetic algorithm, it was concluded that the accuracy of the decision tree and Naive Bayes has been improved in a great extent [4].

In 2010, M. Anbarasi et al. developed an enhanced prediction of heart disease with Feature Subset Selection via genetic algorithm. Initially, the number of features utilized to predict heart diseases was 13. Afterward, due to the integration of feature subset selection with high model construction time using genetic algorithm, they used 10 attributes for the predictions. Finally, we should mention that Naive Bayes, classification by clustering and Decision Tree were used to their research [4].

3. Basic terms related to data mining

3.1. Classification

Classification is a data mining function which is responsible to assign items in a collection in order to be classified in the proper category. In essence, it's a procedure in which a new observation belongs to the base on the training set of data containing observations and whose categories and classes membership are known. For instance, a model which is created by classification principles could be used in order to identify people who have predisposition to be found with heart disease in low, medium or high level.

The whole procedure starts with a data set in which the class assignments are predetermined. For example, a classification model which is assigned to predict the heart disease risk can be built according to observed data for many patients for a particular period. Furthermore, the data is likely to track the family history and some likenesses among the family members to the risk factors.

Classification algorithms discover the connection between the values of the predictors and the values of the target. Every classification algorithm implements different techniques in order to find relationships. The specific relationships are summarized in a model which is able to be implemented to various data sets with unknown classes. Simultaneously, they are tested by comparing the predicted values to known target values in a set of test data.

Finally, classification could be utilized to a lot of professional fields such as customer segmentation, marketing, credit analysis and in biomedical industry [12] [13].

3.2. Supervised learning

Supervised learning, in the artificial intelligence field, is where you have input variables X and an output variable Y and by using an algorithm, you are able to learn the mapping function during the transition from the input to the output. In mathematical terms, the above is displayed as Y = f(X) [15].

Supervised machine learning systems utilize learning algorithms in which quantities and labeled data are known in order to assist in future judgments and decisions. Some of the most acknowledged examples of supervised learning are chatbots, facial recognition, text recognition, robots, self-driving cars, stock prediction time series and most of them are affiliated with retrieval-based artificial intelligence (AI). Also, training data used to supervised learning is composed of a set of examples with related input subjects and desired output.

Moreover, supervised learning models are more capable of making decisions that humans can understand because humans included the decision-making to the systems. Contrary to the specific fact, in retrieval-based practices, supervised learning is not able to handle new information. For instance, whether a system which is created in order to recognize vehicles in two categories such as cars and buses, then if the new input is a bicycle, algorithm would classify it incorrectly in one category or the other [14] [16].

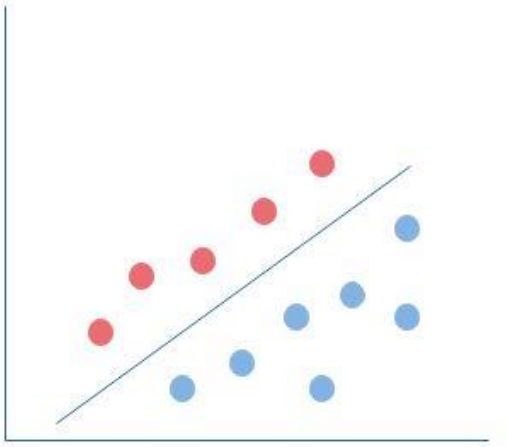


Figure 1: Supervised Learning

3.3. Unsupervised learning

Unsupervised learning, in the artificial intelligence field, is an aspect of machine learning in which test data are not labeled or classified and provide the opportunity to algorithm to act on new information without any help or guidance. At the same time, based on the fact that outcomes are unknown, there is no manner in order to calculate the accuracy of the algorithm. As a result, supervised machine learning is claimed to be more applicable and trustful to real-worlds problems and tasks [18].

However, unsupervised learning is commonly used in case you don't have the required data on desired outcomes. For example, in the marketing section, via unsupervised learning algorithms, marketeers have the chance to speculate the target customers for a new product, but whether they are willing to comprehend the prevalent situation of the status consuming, they should make good use of supervised learning methods [19] [17].

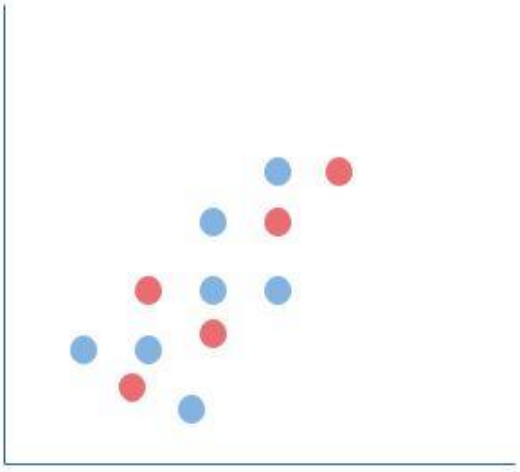


Figure 2: Unsupervised Learning

3.4. Prediction

It is one of the data mining techniques utilized to discover the relationship among to dependent and independent variables.

4. Experimental Environment

In the specific part of the research, it is essential to provide the technologies which were utilized in order to implement the project work and looks like as follows:

Client side

- HTML5
- CSS3
- Bootstrap 4
- JavaScript – jQuery
- Python

Server Side

- PHP 7

Database

- phpMyAdmin - XAMPP

4.1. Required Packages

The required packages in order to be completed successfully the prediction procedure are the below ones:

- cycler – version: 0.10.0
- kiwisolver – version: 1.0.1
- matplotlib – version: 3.0.3
- numpy – version: 1.16.2
- pandas – version: 0.24.2
- pip – version: 19.0.3
- pyparsing – version: 2.4.0
- python-dateutil – version: 2.8.0
- pytz – version: 2018.9
- scikit-learn – version: 0.20.3
- scipy – version: 1.2.1
- setuptools – version: 40.8.0
- six – version: 1.12.0

4.2. Database Structure

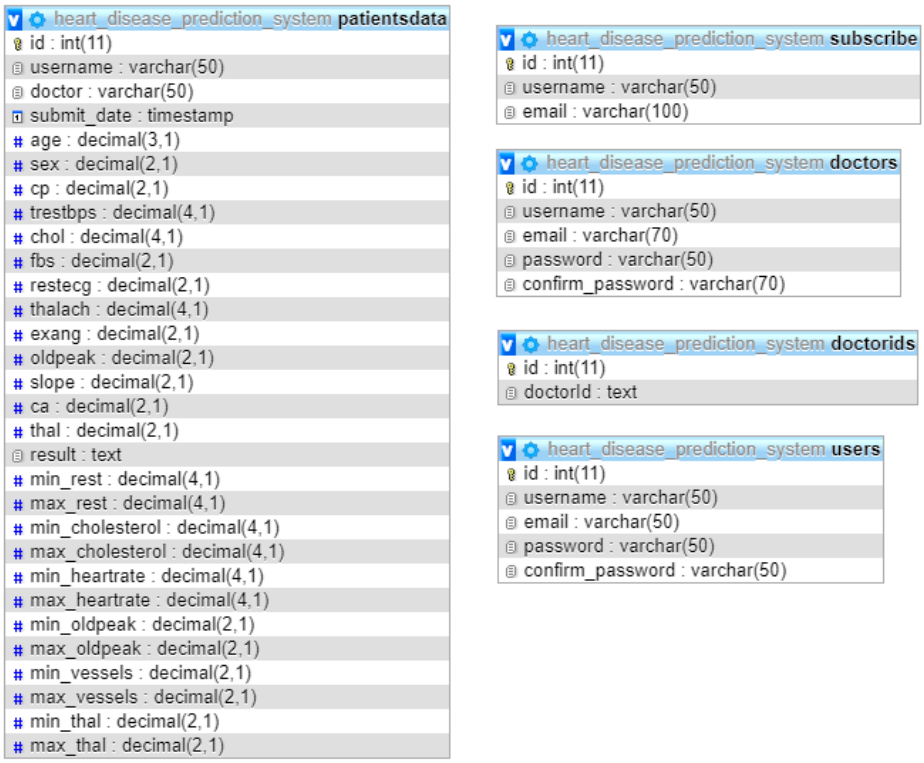


Figure 3: Database Structure

5. Proposed System

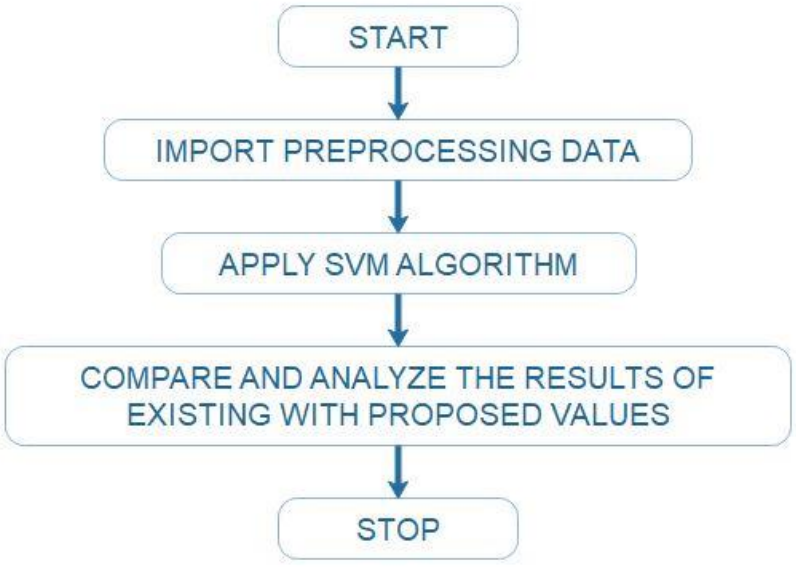


Figure 4: Flow diagram of proposed system

The above figure displays the major steps of the proposed system. Once we start the procedure, algorithm would import the preprocessing data which are represented by

the Cleveland Dataset of 303 instances. Afterward, it is deployed the SVM algorithm in order to compare and analyze the results of existing values of the dataset with the proposed input values from the medical practitioner – user. After the completion of the process and the display of the required predictions, users would have the chance to be aware of patient's situation and the probability to appear heart disease in the foreseeable future.

6. Dataset Details

6.1. Description of Cleveland Dataset

The data set for the specific research was obtained from UCI machine learning data repository which is called Cleveland heart disease database and consist of 76 attributes. Simultaneously, due to the fact that the majority of published experiments are inclined to utilize a subset of 14, I would like to select 13 of them for this paper. As detailed below, table 1 depicts the list of 13 attributes and a brief description about their meanings. According to researchers, Cleveland is claimed to be one of the most appropriate databases for creating a data mining model because it contains lesser missing values, duplicates and errors. Initially, the dataset was cleaned and preprocessed in order to assist the proposed algorithm to become more accurate and effective during the training and testing [1] [7].

Data set characteristics	**Multivariate**
Number of records	303
Missing values	YES
Number of attributes	13
Attribute characteristics	Decimal

Table 1: Information of dataset

Table 2 represents the significant risk factors acquired from databases and their corresponding values.

	Variables	**Brief Description**	**Options**
1	age	Input age in years (e.g. 67.0)	Not stable value
2	sex	Input male = 1 , female = 0	Male or Female
3	cp	Input chest pain type	Chest Pain Type (Valid values: 1 - 4): 1. Typical angina 2. Atypical angina 3. Non-anginal pain 4. Asymptomatic
4	trestbps	Input resting blood pressure in mmHg	Not stable value in mmHg
5	chol	Input serum cholesterol in mm/dL	Not stable value in mm/dL

6	fbs	Input blood sugar in mg/dL	Fasting blood sugar > 120 mg/dL -> value = 1 Otherwise, value = 0 Value 1 = TRUE Value 0 = FALSE
7	restecg	Input electrocardiographic results	Resting electrocardiographic results with values 0 -2: 0: normal 1: having ST-T wave abnormality (depression > 0.05mV) 2: showing probable or predetermined left ventricular hypertrophy
8	thalach	Input the maximum heart rate	Not stable value classified to normal and abnormal
9	exang	Input exercise induced angina	Exercise induced angina. Values are ranging from 0 to 1. Value 1 = YES Value 0 = NO
10	old peak	Input ST depression induced by exercise relative to rest	Not stable value
11	slope	Input the slope of the peak exercise ST segment	Values are ranging from: Value 1: up sloping Value 2: flat Value 3: down sloping
12	ca	Input the number of major vessels colored by fluoroscopy	Major vessels from 0 – 3
13	thal	Input the heart rate	Heart rate of the patient: Value 3: normal Value 6: fixed defect Value 7: reversible defect
14	result	Display the predicted value	Value 0: absence of heart disease Value 1: presence of heart disease

Table 2: Attributes – factors associated with patient suffered by heart disease and a brief description about their role.

7. Data preprocessing

Data preprocessing is one of the most prevalent procedures in data mining. It is responsible for the preparation, transformation and normalization of the data in order to be into appropriate form. Data preprocessing intends to discover the associations among data, decrease its size and contribute to integration.

Figure 5: Preprocessing Steps

7.1. Normalization

Normalization is claimed to be a mapping technique or a preprocessing stage, which means that we are able to regard a new range of data over the existing one. It is known that there are a lot of ways to make predictions but, every algorithm has a peculiarity of its own to make results. Therefore, normalization is capable of bridging the gap between the different algorithms. The specific method is useful to various data mining techniques such as clustering and classification. Furthermore, we should relate that there is a great availability of normalization techniques:

1. Min – Max Normalization
2. Decimal Scaling Normalization
3. Z-Score Normalization

7.2. Min - Max Normalization

In Min – Max Normalization, all features are assured to be scaled on the same level, which means that the values are normalized within the given range. It is one of the most frequent ways to normalize data because the minimum value of every feature would be modified to 0

and the maximum value to 1, respectively [20] [21]. Min – Max Normalization implements the below formula:

$$\boldsymbol{Value - Min \,/\, Max - Min}$$

7.3. Decimal Scaling Normalization

In Decimal Scaling Normalization, in essence, we move the decimal point of values of the attribute. The specific movement of decimal points depends on the maximum value between all values in the attribute [22]. Decimal Scaling Normalization, implements the below formula:

$$\textbf{Normalized value} = v^x \,/\, 10^y$$

7.4. Z-Score Normalization

Z-Score Normalization, which is called as Zero Mean Normalization, is a method that normalize data in order to resolve the outlier problem and it is based on the mean and the standard deviation [21]. The specific strategy implements the below formula:

$$\textbf{Value} - \mu \,/\, \sigma$$

μ is the mean value of the feature and σ is the standard deviation of the feature. Whether a value is equal to the mean of the feature, then it would be normalized to 0. If it is below the mean of the feature, it would be normalized to a negative number, otherwise, it would be a positive number.

8. Data mining techniques

8.1. Renowned Classification techniques used in data mining

8.1.1. Naïve Bayes Algorithm

As everyone can perceive, Naïve Bayes algorithm is based on Bayes theorem. It represents a supervised learning strategy which could be used as a statistical method for classification. Naïve Bayes algorithm is named by Thomas Bayes who was one of the most renowned theologians during 18th century. The specific classification method is claimed an analytical classifier which has the chance to categorize without any dependency among attributes and it begins with the most simplistic probabilistic classifier. Consequently, it makes some assumptions and learn the Naïve Bayes Classifier. Furthermore, Naïve Bayes classifier is regarded effective besides the continuous demand of several parameters during the learning phase [8]. Given training data B, strategy of the Bayes theorem is the below one:

$$\textbf{P(A/B) = (P(B/A) P(A))/P(B)}$$

8.1.2. ANN Algorithm

An Artificial Neural Network (ANN) is claimed to be a mathematical model which emulates the biological neural networks. It is composed of an interconnected group of artificial neurons which replicate the way that humans learn. Neural networks consist of input and output layers and, commonly, feature hidden layers which are responsible to transform the input into something that output layer can make good use of it and provoke desirable results. Frequently, an artificial neural network is an adaptive system as it can alter its structure because of external factors and, for this reason, it is assumed as non-linear statistical data modeling tool [23]. As detailed below, the basic structure of an ANN is as follows:

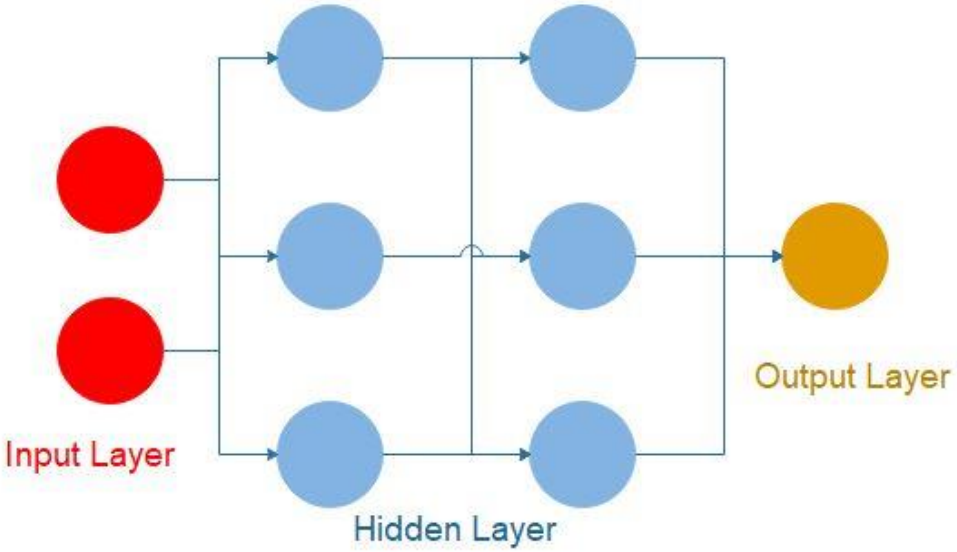


Figure 6: Structure of ANN

8.1.3. Decision Tree Algorithm

All in all, decision trees are regarded easy to understand and modify as the model developed could be expressed as a set of decision rules. It belongs to the category of supervised learning and the dissimilarity with other algorithms is the fact that can be used for solving regression and classification problems as well. Although there are unlimited numbers of training examples and attributes in various databases, decision trees scale in a great level. The output of the decision tree algorithm is a binary tree-like structure which gives the opportunity to interpret conveniently and recognize the important variables. Simultaneously, a decision tree model is composed of rules in order to predict the desirable variable and features with a description of the distribution of the data. Last but not least, we should refer to two phases of a decision tree classifier:

1. Growth phase
2. Prune phase

During the growth phase, the initial aspect of the tree has been built. Afterward, a sub-tree is built, which is the prune phase, in order to eliminate small and deep nodes of the tree provoked by the procedure of the training data. The aim of this measure is to restraint the risk of overfitting and make more effective classification of different data [24] [25]. As detailed below, the basic structure of a Decision Tree is as follows:

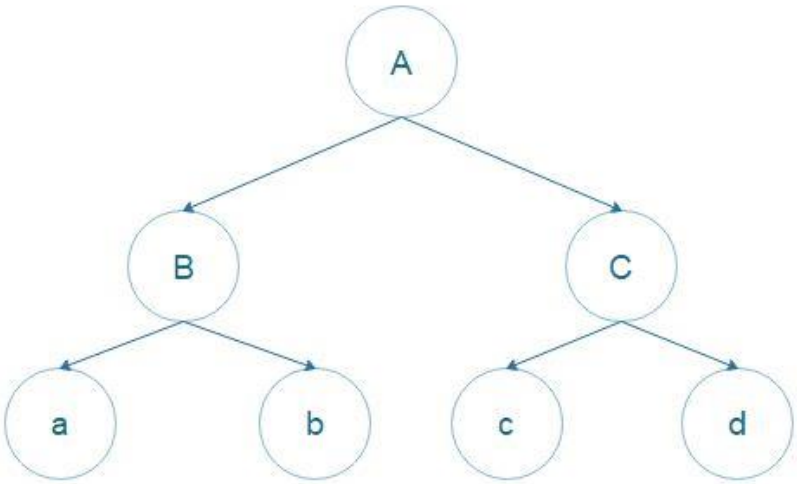


Figure 7: Structure of Decision Tree

8.2. Benefits and limitations of classification algorithm

	Advantages	**Disadvantages**
[26] Naive Bayes Algorithm	• Easy implementation • Small amount of training data to calculate the parameters • Good results	• No class conditional independence which means loss of accuracy • Dependencies between variables
[27] ANN Algorithm	• Storing information on the whole network • Capacity to manipulate incomplete knowledge • Fault Tolerance • Distributed memory • Parallel processing capability	• Hardware dependence • Unknown behavior of the network • No particular network structure • Austerities with the display of the problem to the network • Unknown duration of the network
[28] Decision Tree Algorithm	• Easy implementation • Required modest data training • Manipulate numerical and categorical data • Time efficiency in large data	• No assurance of returning the optimal decision tree • Creation of complex trees which could fail to generalize data in the proper way • Decision trees, for categorical variables,give a preconception about the information gain on the side of the attributes.

9. Past research

In recent years, several researches have been conducted in order to develop better and accurate models using heart disease data set. In the first stage, in 1979, Dr. Robert Detrano utilized Logistic Regression and managed to reach 77% accuracy [5]. In 2005, Polat et al. created a technique which utilizes artificial immune system (AIS) and reached 84.5% accuracy during the classification. Four years ago, Newton Cheung implemented C4.5, BNND, BNNF and Naive Bayes algorithm and the percentage of their accuracies are 81.11%, 81.11%, 80.96% and 81.48% correspondingly. Also, in the majority of the researches were utilized ToolDiag and WEKA Tools. A high percentage of researchers have opted in favor of 10-fold cross validation on the data and mentioned the result for the heart disease and, generally, the medical diagnosis. My study has implemented Python for the development of the algorithm and the appropriate machine learning libraries in order to have the desirable results. Furthermore, with the aim of reliable heart disease predictions, I utilized test-train split method and cross-validation for optimal parameters choice [11].

10. Brief Introduction to model

Support vector machine (SVM) is a supervised machine learning algorithm utilized for classification purposes, based on a state-of-the-art maximum margin classification algorithm rooted in statistical learning theory which was created by Vladimir Vapnik. SVM is a method which is used both linear and non-linear data. In general, with the aim of non-linear mapping, we should modify the original training data into a higher dimension. In the specific algorithm, every attribute will be regarded as a dimension which means that if it is possible to exist n attributes, then it would be created a plotted result with n dimensions. In case of two classes, data could be separated by a hyperplane. Hyperplane identifies the differentiation of the classes. Once we discovered it, then the whole procedure is extremely easy. SVM algorithms have the chance to consider the hyperplane implementing both support vectors and margins. Moreover, the algorithm attempting to classify maximizes the margin which separates both classes and minimizes the error rate of the classification process [5].

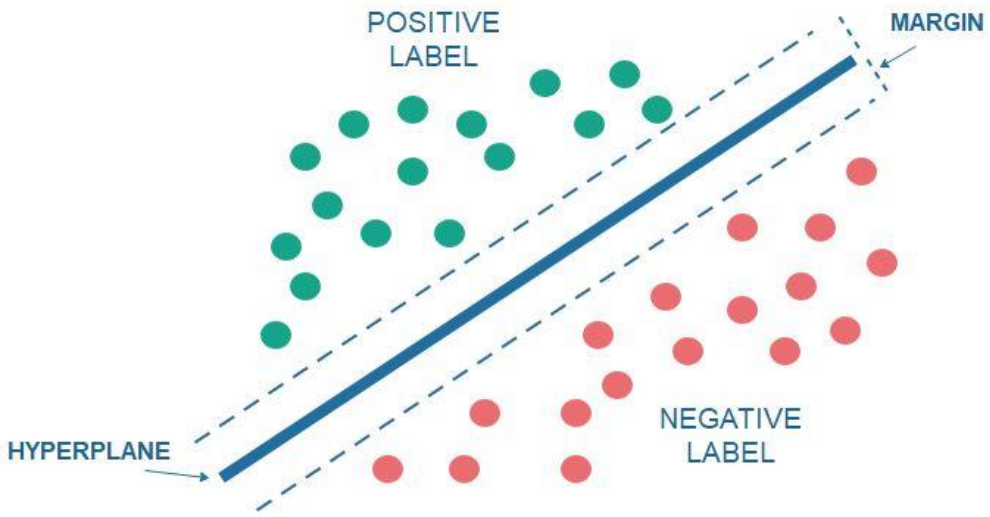


Figure 8: Separating data set with the maximal margin

11. Overview of SVM

In mathematical problems, the SVM algorithm chooses a hyperplane in order to reduce structural risk. By means of implementing this measure, we are able to create an agreement among to intricacy of the decision making and the functions which are responsible for the fitness of the model to the training data set. The specific procedure gives the chance to trainer classifier to predict outcomes with more effective ways to every individual beside the training samples.

The SVM model connected with a kernel function gives the permission to non-linear classifiers to be formed by establishing the original data into a space which has multiple dimensions than the initial one. In other words, whether a linear classifier is utilized in higher dimension space, then it is equivalent with a non-classifier in the initial space.

SVM includes three processes, the creation of the maximal distance among points appeared, the support vector formation and the right-angled decision limit. This model is called linear SVM. Furthermore, we should relate that the margin of the hyperplane intends to be maximized in order to make better results during the classification process and, at the same time, to eliminate the classification errors. The SVM algorithm attempts to predict the appearance of heart disease in a large number of dimensions and classify the data into different labels with the formation of the margins among data groups and clusters [9] .

There are a lot of advantages about the features of the SVM algorithm:

1. The procedure of training and the quality of generalization are extremely simple in comparison with different prevalent techniques.
2. SVM algorithm has the chance to establish models which are effective and complex, and respond to human requirements.
3. One of the most significant advantages is the fact that SVM algorithm is capable of great performance on data sets with a high number of attributes in the unlikely case that exists a limited number of cases during the training phase. The above argument is based on the fact that neural networks should have a great quantity of cases in order to perform well.
4. SVM models are very efficient where there is not information about the data.
5. In case of high dimensional data, algorithm scales very well.

On the other hand, there are some disadvantages about the features of the SVM algorithm:

1. In case of large training data sets, the training time is extremely increased.
2. The selection of an effective kernel function is not an easy procedure.
3. There is an adversity to comprehend the final model, individual effect and variable weights [6] [29].

12. Components of the algorithm

12.1. Linear SVM

Linear SVM is one of the most renowned machine learning algorithms which is responsible to solve classification problems with multiple classes and large data sets in order to develop an initial version of a cutting plane algorithm with the aim of implementing a Linear SVM. The specific algorithm is claimed as a linearly scalable procedure as the CPU Time depends on the size of the training data set.

Some of the most significant features of SVM algorithm are the below ones:

1. It can solve classification problems with multiple classes.
2. It is able to function with data in high dimension (many attributes).
3. In case of large data sets data sets, it achieves high effectiveness.

12.2. Linear SVC

One of the most relevant machine learning algorithms is called Linear SVC. The aim of a Linear SVC is to fit to the data and calculate the optimal hyperplane in order to separate and classify the data. Once the hyperplane is selected, it is essential to give some features to the classifier with the aim of determining the class of the prediction. As a result, this algorithm is the appropriate one for our objective to predict heart disease [30].

12.3. PCA

One of the most significant methods used in this research is PCA. PCA aims to decrease the dimension of the feature space in order new variables to be independent between them and orthogonal to each other.

However, the initial step before implementing PCA is to scale the available data set for the purpose of determining each feature to has unit variance. This is crucial step because fitting algorithms should scale the features in order to make good results. In this research, we selected to utilize the StandardScaler module in order to scale the features of the available data set and eliminate the mean from every feature before applying the scaling to unit variance. Finally, we should transform the scaled data.

On the other hand, it is critical to highlight that we are able to determine the number of principal components. In the specific case, we have selected only 2 components in order to reduce the dimension of the data set. Based on the above, one of the disadvantages of PCA is the fact that it is not reachable to provide the original features (in this case there are 13 features - attributes) so that create the principal components.

Taking everything into consideration, we able to proceed to one of the most significant aspects of PCA which is called data visualization. Whether we run the algorithm and based on the fact that components are orthogonal among them, then we are able to discern the differences between the heart disease classes or not [31]. The plot looks like as the below figure:

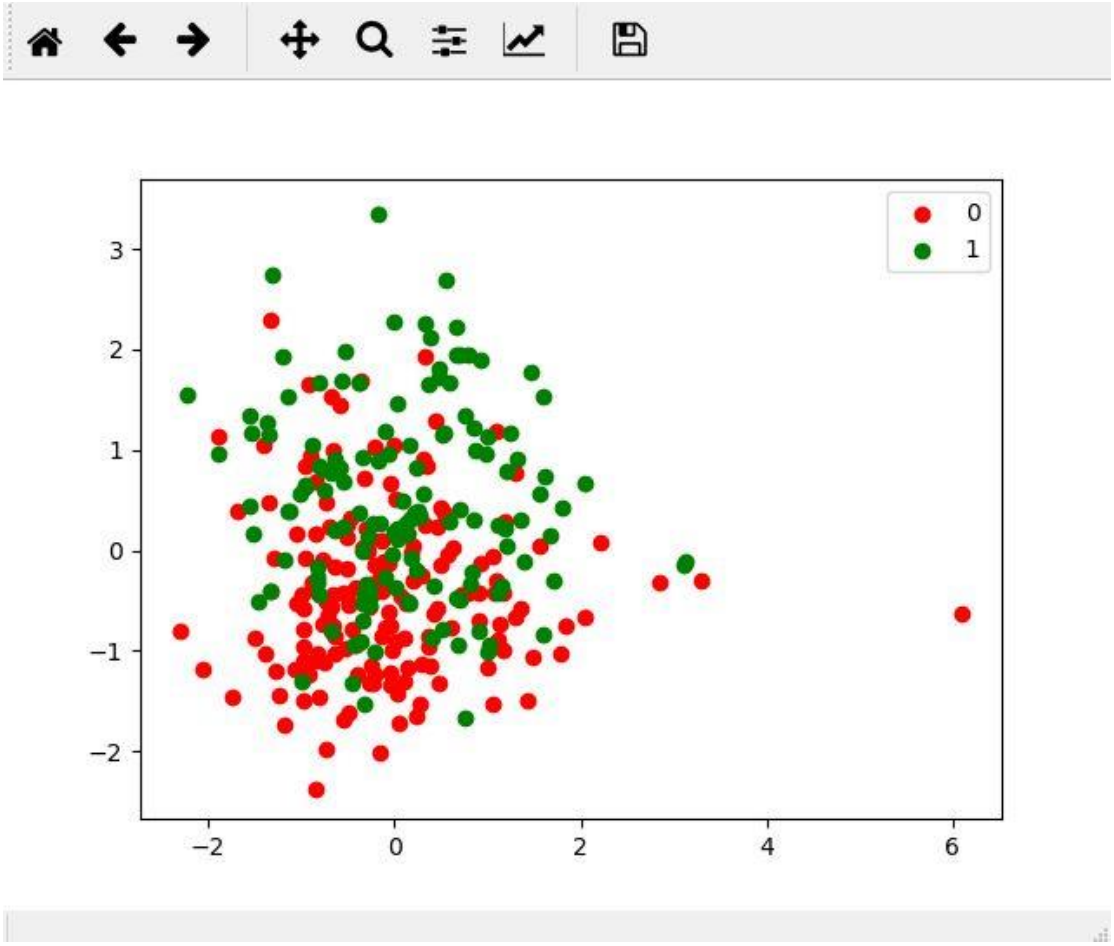


Figure 9: Plot of the heart disease classes based on 2 principal components of the heart disease features - attributes

As we can see above, the two classes (heart disease or not) are separated successfully with the first 2 principal components. Furthermore, we could conclude that PCA is an unsupervised method and it is extremely complicated to perceive the two axes of the plot due to the complexity combination of the original features.

12.4. DataFrame

Pandas DataFrame has the chance to display data in two-dimensional structure and labeled in a form of table with rows and columns. DataFrame is composed of three parts:

- Data
- Rows
- Columns

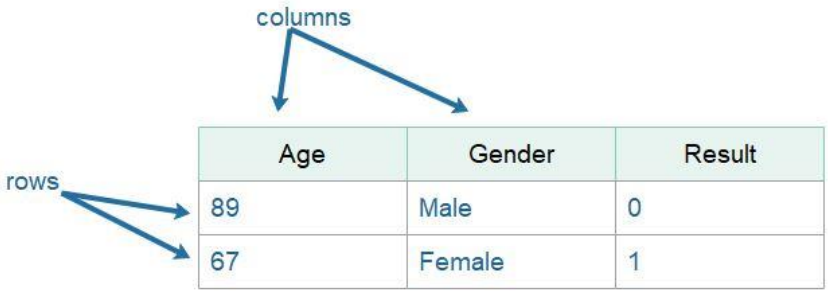


Figure 10: Pandas DataFrame

In the scientific field of technology, with the aim of loading the datasets, Pandas DataFrame is able to be created from lists and dictionaries which are stored in various forms such as CSV, Excel file and SQL Database. Some of the most common methods to create a Pandas DataFrame is as follows:

- DataFrame from List
- DataFrame from dictionary of ndarray or lists.

On the other hand, in order to select columns in Pandas DataFrame, we could just call them using their columns name. Concurrently, in the selection of rows, there is a method which can retrieve rows from DataFrame and it is called DataFrame.loc().

As it is known, one of the most prevalent problems in information systems and data is the fact of missing values. Missing values can be observed in the case of absence of some information for one or further items. In Pandas DataFrame, missing values are declared as Not Available (NA) values. For this reason, in order to examine the possibility of missing values, Pandas DataFrame utilizes various functions such as isnull(), notnull() etc [35].

12.5. Pipeline

As we can assume from its name, pipeline class is responsible to combine a lot of processes into a single estimator of the scikit-learn which is a term defined above. Pipeline class is composed of fit, score and predict methods which are considerably renowned in scikit-learn.

Initially, in order to employ pipeline, we should separate data from their features and labels. This could occur utilizing StandardScaler which eliminates the mean from every feature and, afterward, scale to unit variance [36].

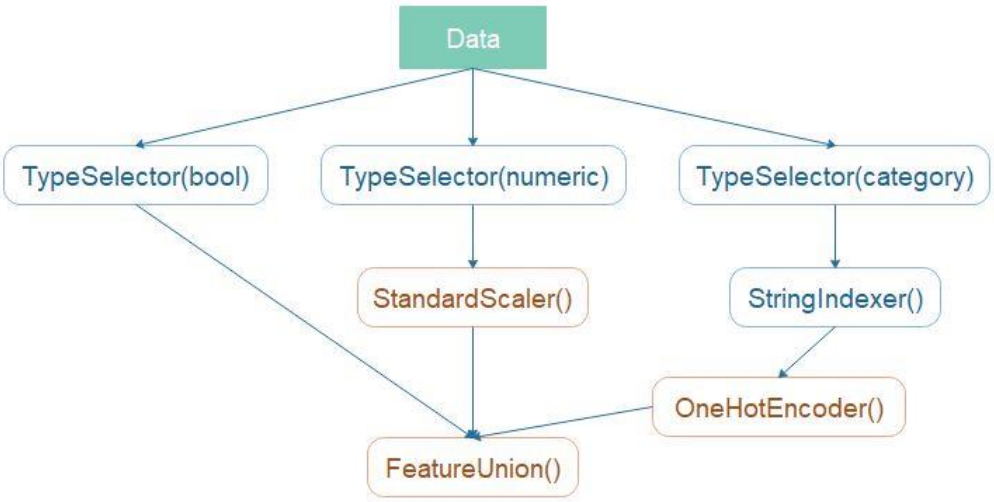


Figure 11: scikit-learn Pipeline

12.6. Cross validation

One of the most significant methods of scikit-learn is called cross validation. It is composed of several similarities to train – test split, but there are some differences as it is applied to more subsets. This means that, we split data into x subsets and train on x – 1 of those subsets. The distinctiveness is the fact that we hold the last subset for the testing phase and we are able to implement it for each available subset [37].

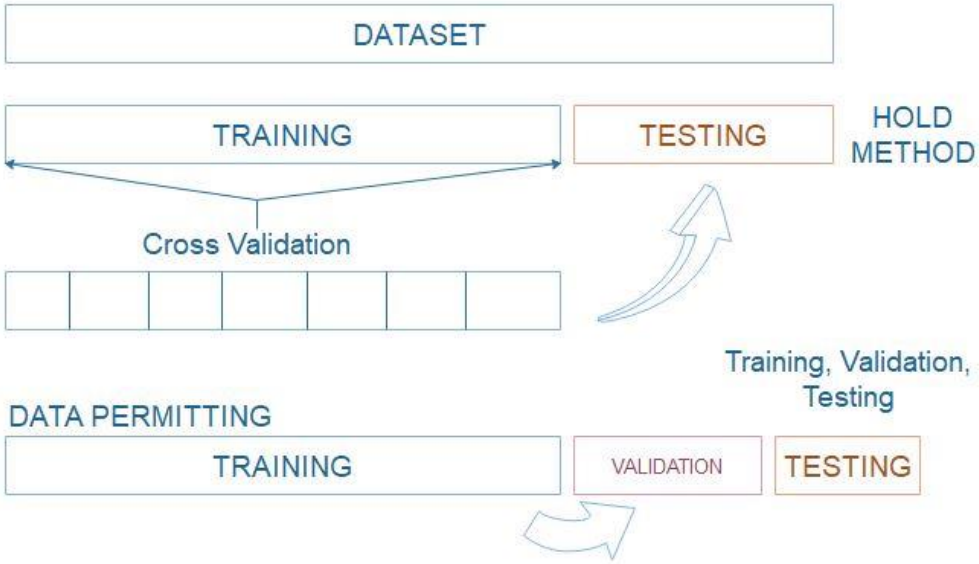


Figure 12: Train/Test Split & Cross Validation

12.7. Pandas

Pandas is one of the most important parts of data analysis field for Python programming language. It is an open-source library which could provide high performance in procedures and convenience to data structures and analysis. Python is claimed as a programming language responsible for data preparation, but less for data analysis and modeling. In comparison with other programming languages, the environment for implementing data

analysis in Python provides a great productivity, performance and flexibility. In the below section, we refer to the major features of the library:

- Effective DataFrame object in order to handle data using indexing characteristic.
- Integrated tools for data procedures (for example reading and writing data) which could utilize different formats such as CSV and Excel files.
- Readjustment of dataset shape.
- Slicing, subsetting and indexing of large datasets.
- Axis Indexing based on hierarchy. This means that Pandas is able to manipulate high-dimensional data in a lower-dimensional data structure.
- Insertion and deletion of columns with the aim of size changeability [38].

12.8. NumPy

NumPy is one of the most significant packages in data science as it provides an array data structure which has unlimited advantages due to faster access in reading and writing items, further convenience and effectiveness. It is composed of collection of tools and techniques that could be utilized in order to solve mathematical models in science of computer engineering and technology. For instance, manipulation of high-dimensional data in multidimensional array object utilizing mathematical functions with high performance is one of the features [39].

12.9. Scikit-learn

Scikit-learn includes a large variety of supervised and unsupervised learning algorithms. One of the capacities is the BSD license and its Linux distribution which are considerably «attractive» in commercial section. The specific library is created by the SciPy and includes the following:

- SciPy which is the basis of scientific computing
- NumPy which is the basis of n-dimensional array package
- Pandas which is responsible for data structures and analysis for Python programming language.
- Matplotlib features of 2D and 3D plotting
- IPython which is an interactive console of Python programming language.

Furthermore, we should mention that the vision for the library is a mean of strength in the field of production. Based on the above, we could consider that the most significant concerns of the scikit-learn are the high convenience, flexibility, code quality and by extension performance [40] [41] [42].

Finally, some functionalities of scikit-learn are the following:

- Classification
- Clustering
- Regression
- Preprocessing
- Model Selection

12.10. Matplotlib

Matplotlib is a library used in plotting procedure for 2D graphics and it was recommended by John Hunter in 2002. It is commonly utilized in python shell, web application servers and various user interface toolkits and the greatest advantage of visualization is the fact that permits to get access in a large amount of data. Generally, there are a lot of toolkits which provide functionalities and features in matplotlib library. A percentage of them could be a different part of the source code (in form of download) and others are integrated into the source code with the association of external dependencies. Some of them are the following ones:

- Basemap
- Cartopy
- Excel tools
- Mplot3d
- Natgrid

Moreover, some of the major available plots is as follows:

- Bar Graph
- Histogram
- Scatter plot
- Area plot
- Pie plot

Nevertheless, learning matplotlib could be a confusing procedure due to the lack of documentation. For this reason, some of the major challenges to deal with are the following:

- Library's code is considerably huge in respect of total lines
- Matplotlib is associated with a lot of interfaces and it has connection with a few backends [43] [44] [45].

12.11. Itertools

Based on the documentation, itertools is claimed as a module which is capable of creating a number of iterator building blocks. As a whole, they form an iterator algebra that assists in constructing tools which are efficient and briefly using Python programming language. In other words, the purpose of itertools is to process iterators in order to create more complicated iterators. For instance, with the aim of generating a sequence of repeated values, we could implement the itertools.cycle function which takes as argument iterable inputs that and returns an infinite iterator which goes to the beginning point once the end of inputs is provoked [46].

13. Mathematics behind SVM

Before we analyze the SVM algorithm, it is important to include some definitions in order to understand the below mathematical terms.

13.1. Length of a vector

Length of a vector y, in other words, is called its norm and its form is as follows: ||y||. For the purpose of computing the norm of the y, we could use the Euclidean norm formula in the scope y = (y1, y2, y3, …, yn).

$$||y|| = \sqrt{y1^2 + y2^2 + y3^2 + \ldots yn^2}$$

Figure 13: Euclidean norm

13.2. Direction of a vector

The form of the direction of a vector y = (y1, y2) is as follows:

$$w = (\frac{y1}{||y||}, \frac{y2}{||y||})$$

Figure 14: Direction of a vector

At the following Figure 10, we have the opportunity to see the cos(θ) = y1/||y|| and cos(α) = y2/||y||. Therefore, the direction vector, which has the symbol ***w,*** is as follows:

$$w = (cos(\theta), cos(\alpha))$$

Figure 15: Direction vector of cos

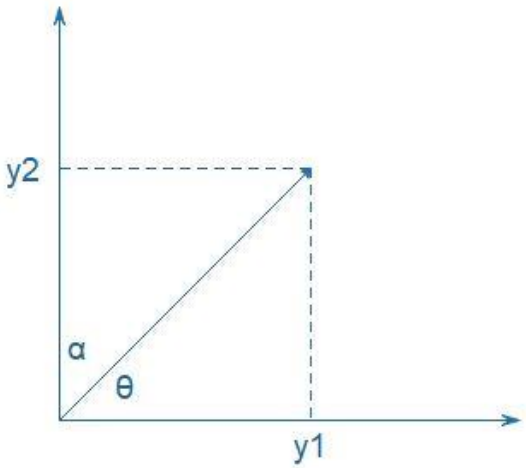


Figure 16: Diagram of the direction vector of cos

Furthermore, we should point out that the norm of a direction vector is always equal to 1. Based on the above, the direction vector, which has the symbol w, is called unit vector.

13.3. Linear separability

Linear separability is claimed as one of the most critical aspects in SVM algorithms. As detailed above in different chapters of the research, the majority of data is not linearly separated, but with the aim of perceiving the specific term we would start with this one and afterward, we would continue with the source of the non-linearly separable cases. In the following diagrams, linear and non-linear cases are depicted:

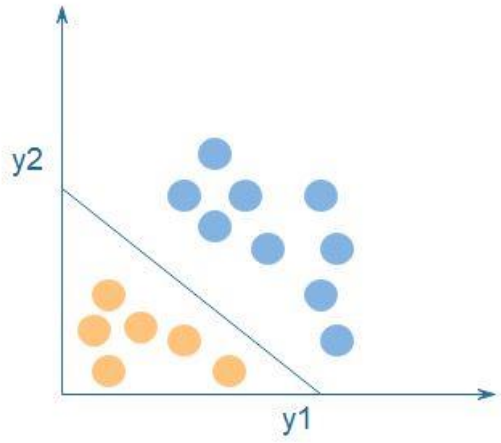


Figure 17: Linear separation

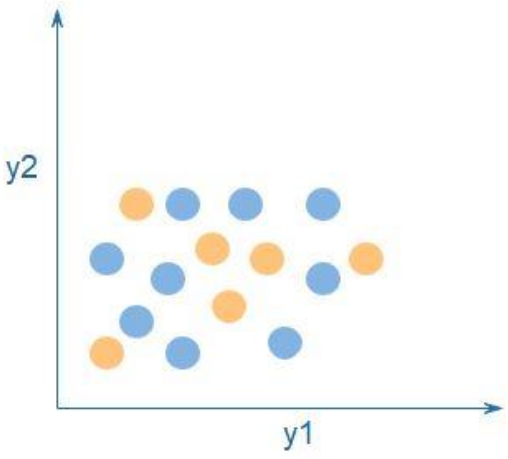


Figure 18: Non-linear separation

13.4. Hyperplane

In the case of hyperplane, we would investigate the two-dimensional one. The two-dimensional linearly separable data could be classified based on the mathematical equation y = ax + b which is a line in space. Therefore, the final equation is going to look like as follows:

$$\alpha x - y + \beta = 0$$

Figure 19: Equation of the line

Whether we declare D = (x, y) and w = (a, -1) then:

$$w * x + \beta = 0$$

Figure 20: Equation from two-dimensional vectors – Hyperplane equation

The specific equation has its source from two-dimensional vectors even it accepts any number of dimensions.

13.5. Classifier

Taking into consideration the hyperplane equation and declare the hypothesis function with the symbol h, we are capable of making reliable predictions:

$$h(ki) = \begin{cases} +1 & w * k + \beta \geq 0 \\ -1 & w * k + \beta < 0 \end{cases}$$

Figure 21: Hypothesis function

On the basis of the above equation, points above or on the hyperplane would be categorized as class +1 and, on the other hand, points which are below the hyperplane would be categorized as class -1. Therefore, the purpose of SVM algorithms in machine learning are to discover a hyperplane which is regarded as the optimal and make better separations among data.

13.6. Dot product

The dot product is used in order to give the association between two vectors. In the following figure, we have the chance to see two vectors x and y and their angle ϕ. According to geometry, the mathematical form is as follows:

$$x * y = ||x|| * ||y|| cos(\phi)$$

Figure 22: Mathematical form of dot product

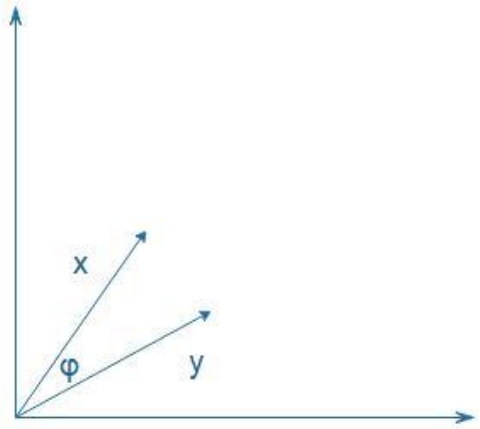


Figure 23: Diagram of mathematical form of dot product

If we define ϕ as ϕ = β – α, then the result looks like as follows:

$$cos(\beta - \alpha) = cos(\beta)cos(\alpha) + sin(\beta)sin(\alpha) = \frac{x1}{||x||}\frac{y1}{||y||} + \frac{x2}{||x||}\frac{y2}{||y||} = \frac{x1y1 + x2y2}{||x||||y||}$$

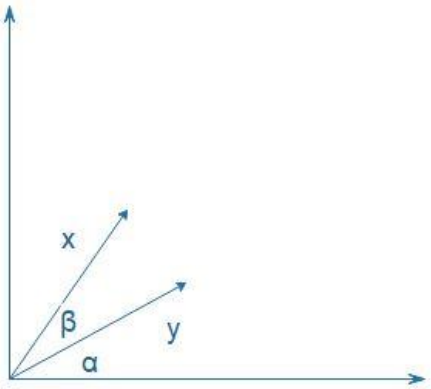


Figure 24: Reformed diagram of mathematical form of dot product

Therefore, the dot product after the above changes is as follows:

$$x * y = ||x||||y||\frac{x1y1 + x2y2}{||x||||y||} = x1y1 + x2y2$$

However, in order to compute dot product for two multiple dimensional vectors, we should implement the following equation:

$$x * y = \sum_{i}^{n} xiyi$$

Figure 25: Dot product for two n-dimensional vectors

13.7. Wolfe dual problem

As it is known, the Lagrangian function is as follows:

$$L(w, b, \alpha) = \frac{1}{2} w * w - \sum_{i}^{m} ai[yi(w * x + b) - 1]$$

Figure 26: Lagrangian function

And the dual problem is the below one:

$$\nabla_w L(w, b, \alpha) = w - \sum_{i}^{m} ai * yi * xi = 0$$

$$\nabla_b L(w, b, \alpha) = -\sum_{i}^{m} ai * yi = 0$$

Therefore, if we combine the two above equations, the result looks like as:

$$W(\alpha, b) = \sum_{i}^{m} ai - \frac{1}{2} \sum_{i}^{m} \sum_{j}^{m} ai * aj * yi * yj * xi * xj$$

And the final dual problem is defined as:

$$max_\alpha \sum_{i}^{m} ai - \frac{1}{2} \sum_{i}^{m} \sum_{j}^{m} ai * aj * yi * yj * xi * xj, ai \geq 0, i = 1...m, \sum_{i}^{m} ai * yi = 0$$

Figure 27: Final form of dual problem

Taking everything into consideration, we are able to conclude that Wolfe dual problem is more practical due to the fact that it depends on the Lagrangian multipliers which are implemented with more ease.

13.8. Kernel Trick

Due to the fact that data are separated non-linearly in two dimensions, we would like to classify it utilizing SVM algorithm. In the first preview, it seems impossible to occur. Nevertheless, whether we modify the two-dimensional data into much more dimensions, such as three, four or five dimensions, we would be able to discover a hyperplane in order to separate the data correctly.

On the other hand, the specific action has some critical drawbacks. For instance, if we have a large number of data, the transformation procedure would outlast and the later computations in order to optimize the performance would affect in a great extent. According to Wolfe dual problem:

$$max_\alpha \sum_i^m ai - \frac{1}{2}\sum_i^m\sum_j^m ai * aj * yi * yj * xi * xj, ai \geq 0, i = 1...m, \sum_i^m ai * yi = 0$$

For the purpose of resolving the specific problem, in essence, we should calculate the dot product xi * xj. Therefore, it is essential to find out a function which transforms the data into higher dimensions, but the result would be the same as the initial one. The particular function is called kernel function and its form is as follows:

$$max_\alpha \sum_i^m ai - \frac{1}{2}\sum_i^m\sum_j^m ai * aj * yi * yj * K(xi, xj), ai \geq 0, i = 1...m, \sum_i^m ai * yi = 0$$

where: *K(xi, xj) = xi * xj*

The most important aspect of this action is to realize that via kernel function we are able to calculate dot product in different space, but with the same result and classify correctly non-linearly data.

Additionally, it is important to point out that there are several kernel functions and the most renowned of them are the below ones:

- Linear kernel
- RBF kernel
- Polynomial kernel

Initially, the *linear kernel*, which is utilized as we aforementioned, looks like as the below one:

$$K(yi, yj) = yi * yj$$

Figure 28: Linear kernel

The *polynomial kernel* is declared as the below one:

$$K(yi, yj) = (yi * yj + c)^d$$

Figure 29: Polynomial kernel

In the specific kernel function, d value depicts the degree of freedom and c a constant. Whether the d value is larger than 1, then the result would be overfitted.

Finally, RBF kernel is declared as the below one:

$$K(yi, yj) = exp(-\gamma||yi - yj||^2)$$

Figure 30: RBF kernel

In essence, the RBF kernel (which is called Gaussian kernel) resolves the problem of overfitting in polynomial kernel. It is composed of a parameter γ which is responsible for the model's behavior. If γ value is small one, then we get a linear SVM. Otherwise, whether the γ value has a numerous value, then model is affected in a great extent [10] [32] [33] [34].

14. Performance Evaluation and comparison

In the following table, it is displayed an analysis of 4 algorithms in order to predict the heart disease based on the accuracy of each algorithm. According to unlimited research papers and the implementation of the algorithm, we have the opportunity to see that Support Vector Machine has higher accuracy in comparison with the other algorithms. Therefore, the performance of the algorithms is shown in the below table:

Algorithm	Accuracy
SVM	90.09%
Naïve Bayes	81.14%
Decision Tree	79.05%
ANN	85.30%

Table 3: Accuracy analysis of Data Mining Techniques

Figure 31:Prediction Performance of Algorithms

Taking everything into consideration, we are able to infer that SVM algorithm is the more effective option in order to predict heart disease [7].

15. Working of the heart disease prediction system

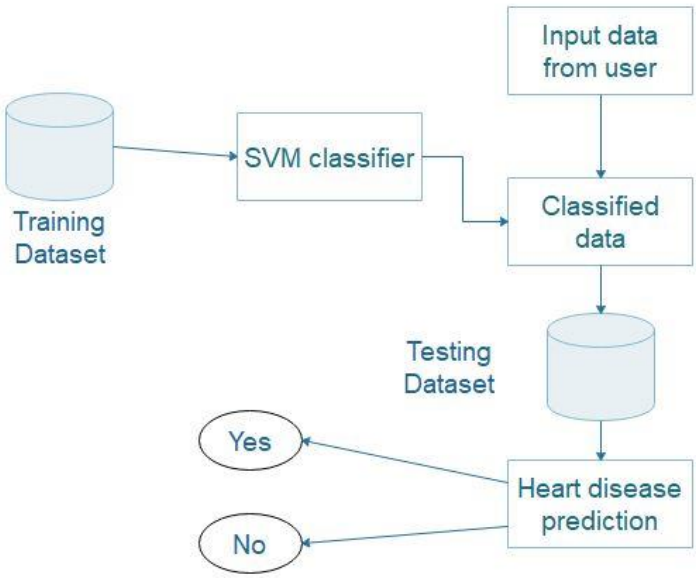


Figure 32: System Architecture

As we can observe to the above figure, the training data set is provided as input to the SVM classifier. We should mention that data, which are classified, are utilized during the testing phase as well. The major system includes the following procedures:

- Training
- Testing

Initially, during the *training* process, it is claimed that classification uses labeled data. This means that we have the chance to know the number of the classes and their definitions, and, simultaneously, a high number of labeled data based on these classes.

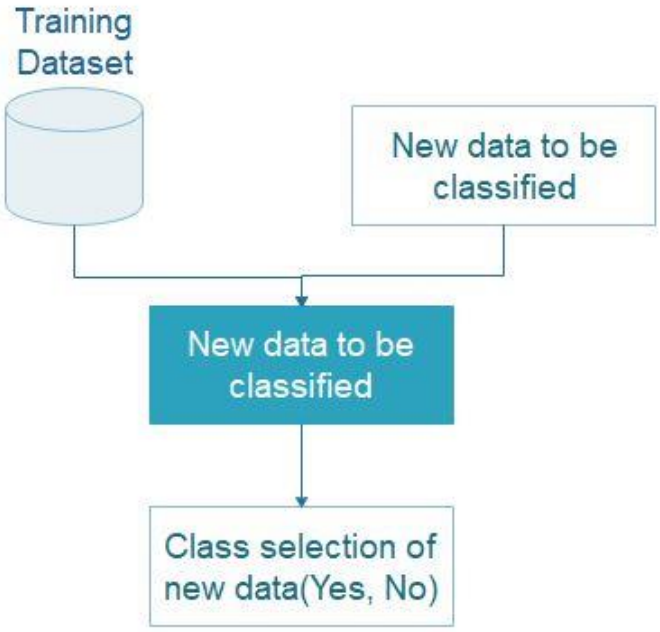


Figure 33: Classification

As it is known, classification is claimed as a supervised learning process. In essence, it creates a model which is responsible to classify data taking into account the training data set and the equivalent class label. This action assists to classify new data of the user.

On the other hand, during the testing process, we attempt to make correct predictions using unknown input data:

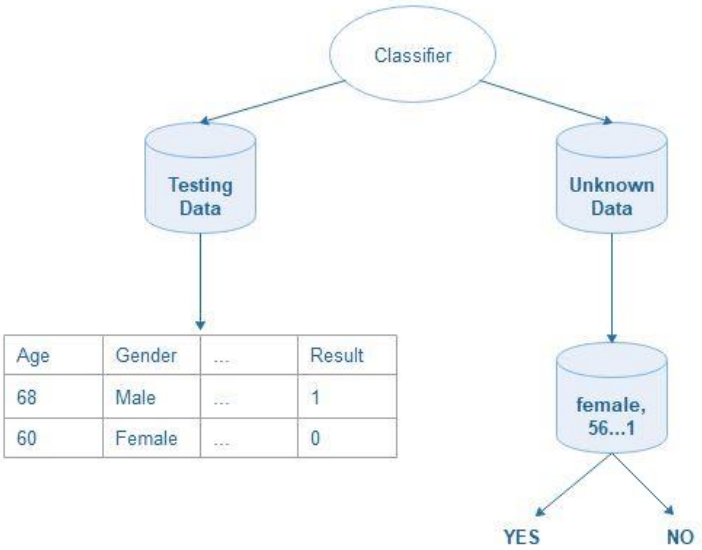
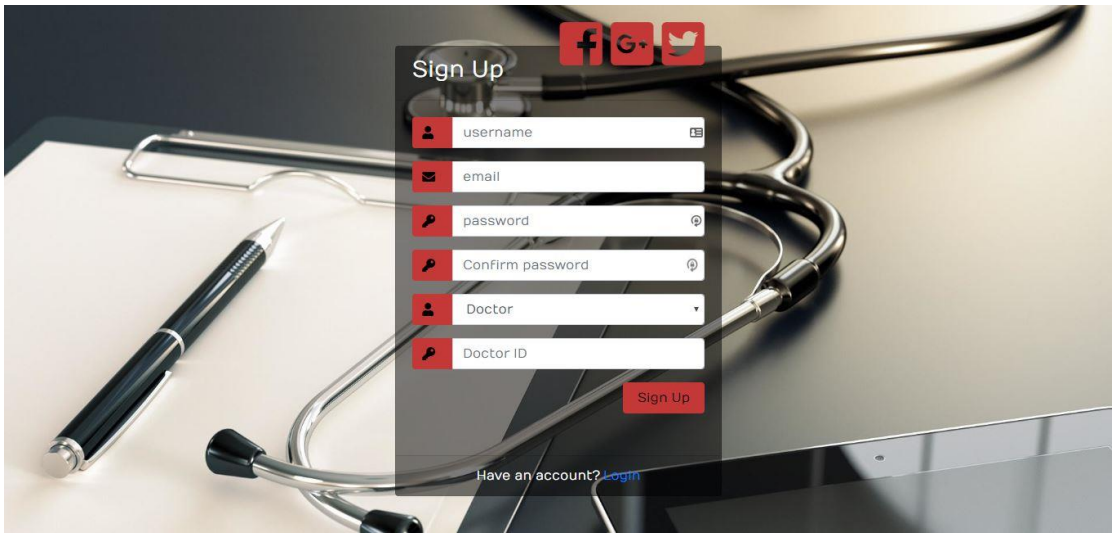


Figure 34: Prediction

16. User Manual

Initially, the first page of the project is the Login Page. Users are requested to insert their usernames and passwords. Login Page is similar to patients and doctors.

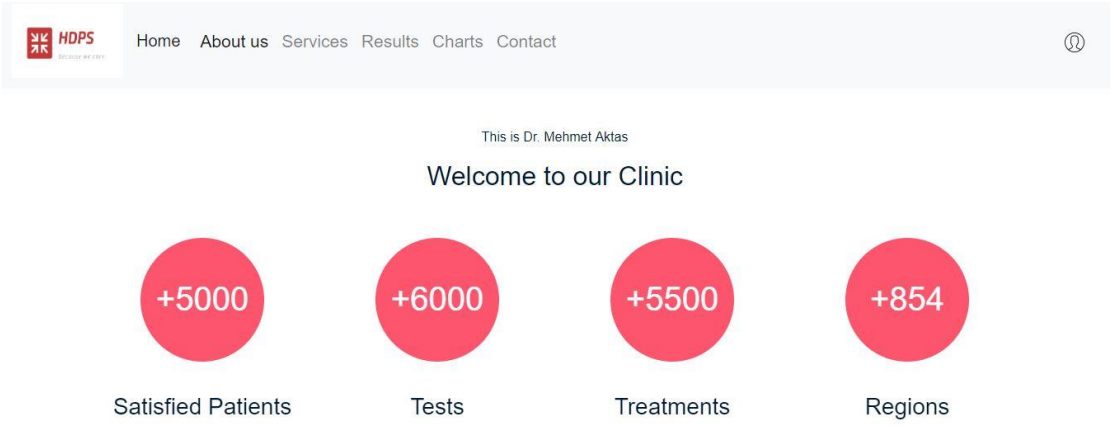
Furthermore, users have the chance to sign up to the system via the Sign Up Page. They should insert username, email, password, confirm password. Simultaneously, in order to distinguish doctors and patients, doctor needs to insert his/her Doctor ID to has a successful registration to the system. Also, patients must register to the system in order doctors to complete the prediction process.



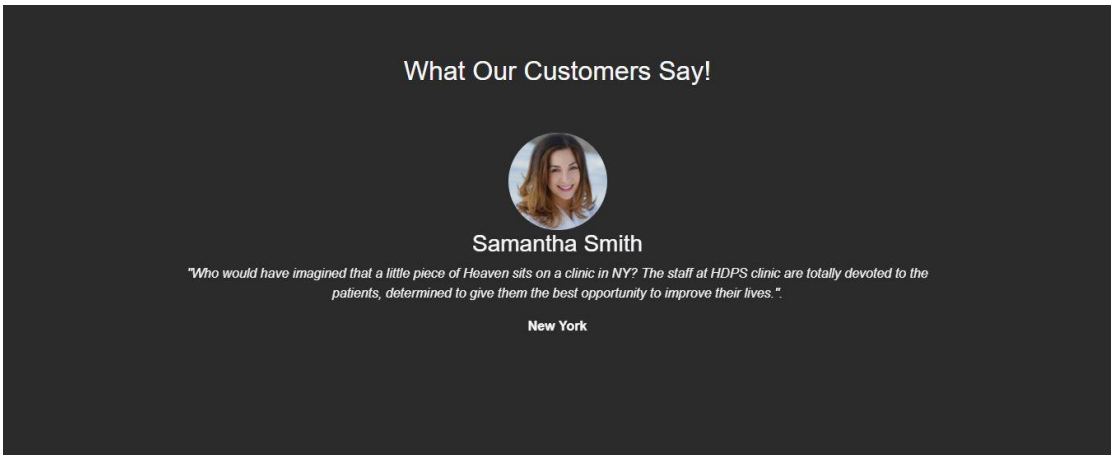
As detailed below, the next pages are similar to doctors and patients. Concretely:

In Home Page, users are able to learn the major features of HDPS in a form of a slideshow.

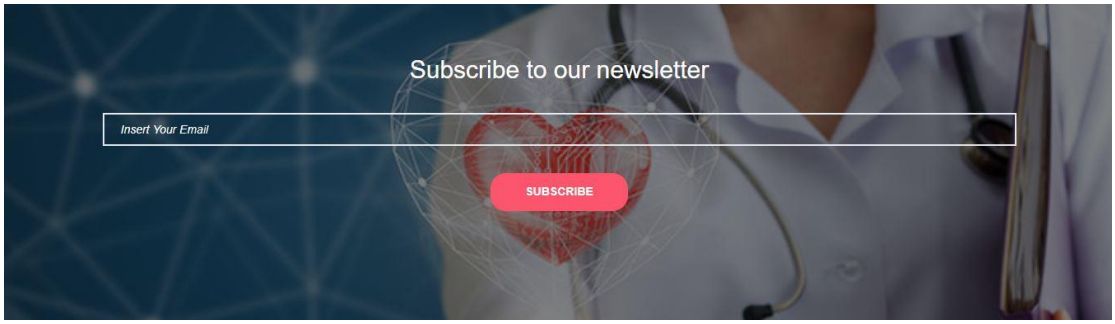
Afterward, users could observe the number of satisfied clients – patients, the conducted examinations etc.



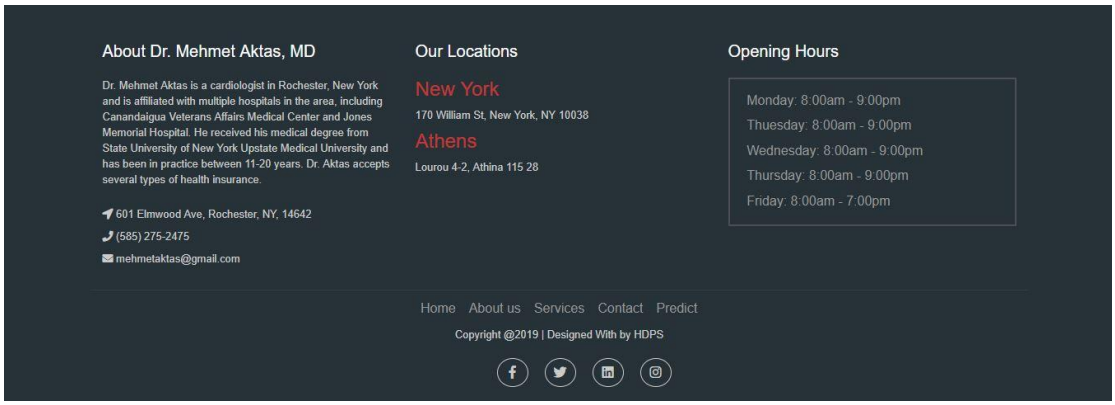
In the specific part of Home Page, users could get informed about the testimonials and the clients' comments.



The input below gives the opportunity to users to insert their emails in order to get notified about HDPS and various information.



Finally, in footer section, doctors and patients are able to learn several information about HDPS such as operation hours.



In About Page, users have the opportunity to get to know details about HDPS and the most important medical practitioners and doctors via a concise resume.

Dr. Mehmet Aktas

Heart Surgeon

Dr. Michael Smith is a cardiologist in New York and is affiliated with multiple hospitals in the area, including HDPS and St. Francis Hospital-Roslyn. He received his medical degree from University at Buffalo, School of Medicine and

Samantha Doe

Cardiologist

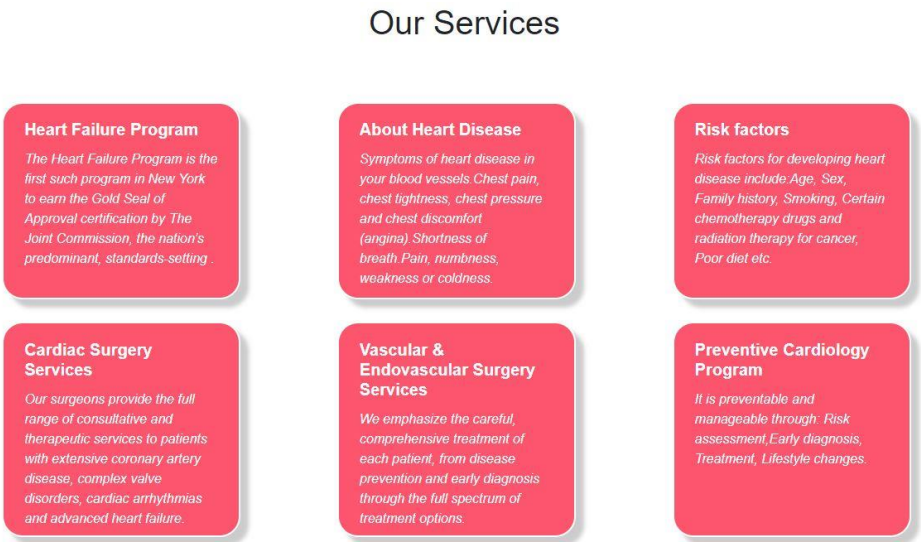
Dr. Samantha Doe is a cardiologist in New York, New York and is affiliated with multiple hospitals in the area, including HDPS and Mount Sinai Hospital. She received her medical degree from Albert Einstein College of Medicine of Yeshiva

Michael Aaron

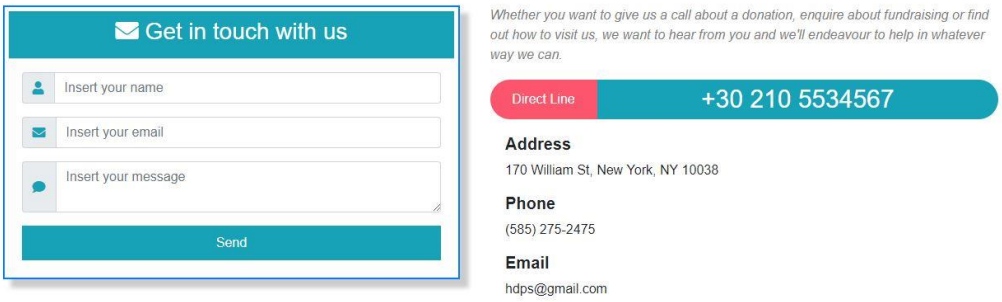
Cardiologist

Dr. Michael Aaron is a cardiologist in New York and is affiliated with multiple hospitals in the area, including HDPS and Community Medical Center-Toms River. He received his medical degree from New York College of Osteopathic

In Services Page, users have the opportunity to get notified about the services provided by HDPS via cards.



Further down, in Contact Page, it is displayed the communication way in order users to get in touch with HDPS and ask any questions.

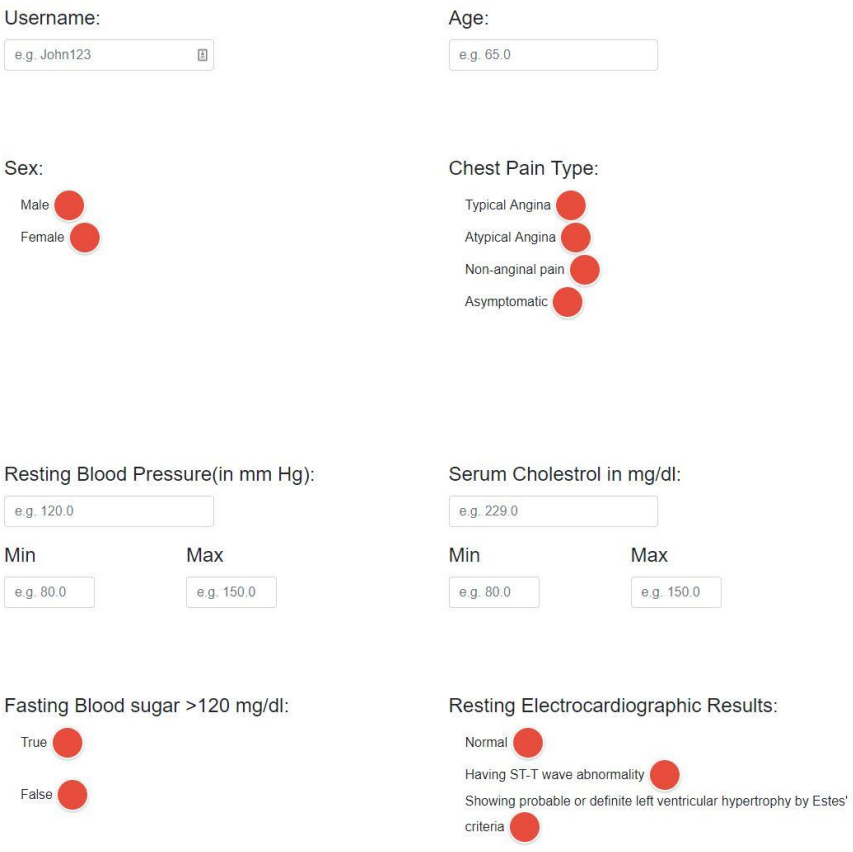


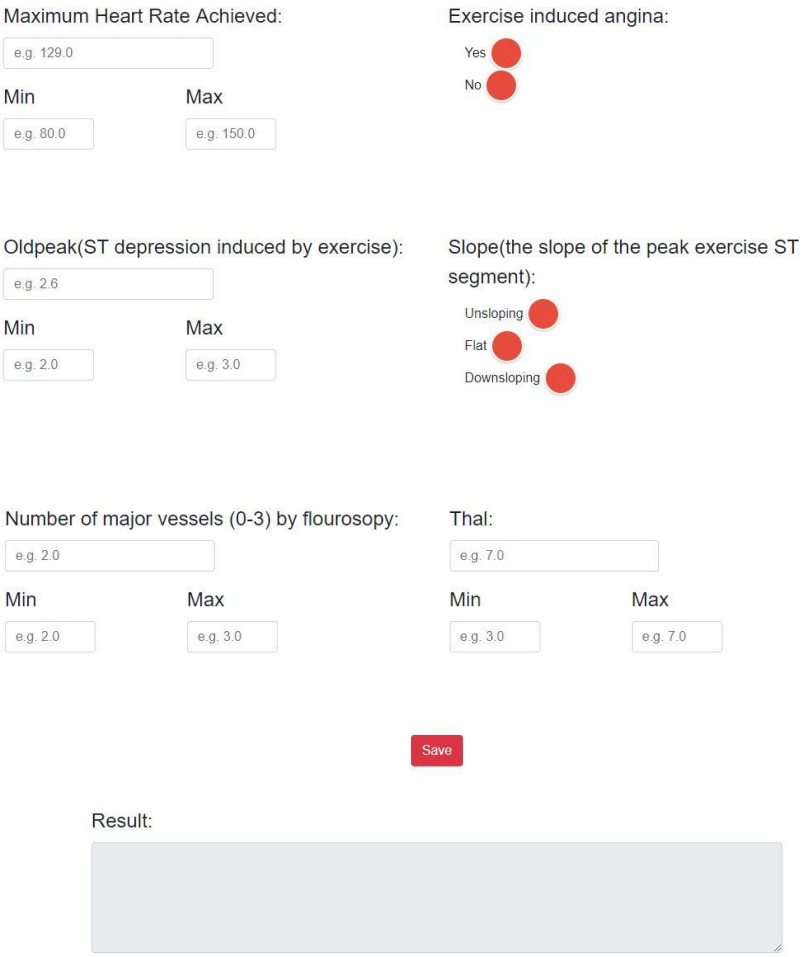
In case of doctor, the navigation bar includes the Predict option. In the specific part, doctors have the capability to insert patients' data to equivalent input fields to predict whether they would appear heart disease or not. The input data are the following ones:

- Username (patient)

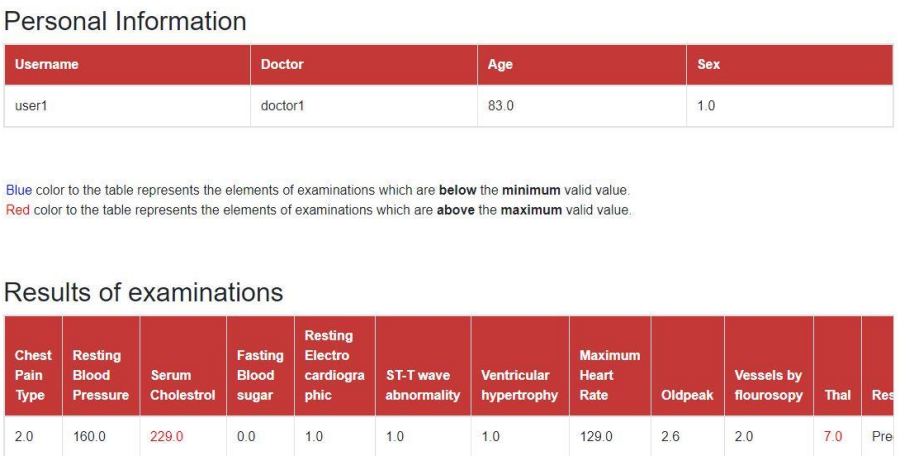
- Age
- Sex
- Chest Pain Type
- Resting Blood Pressure
- Serum Cholesterol
- Fasting Blood Sugar
- Resting Electrocardiographic Results
- Maximum Heart Rate Achieved
- Exercise Induced Angina
- Oldpeak
- Slope
- Number of Major Vessels
- Thalach

Input fields which require a number, users should insert a decimal number, etc. Age: 67.0





In case of patient, the navigation bar includes the Results option. In the specific part, patients have the opportunity to see their examinations, doctor's username and the result of the prediction.



Personal Information

Username	Doctor	Age	Sex
user1	doctor1	83.0	1.0

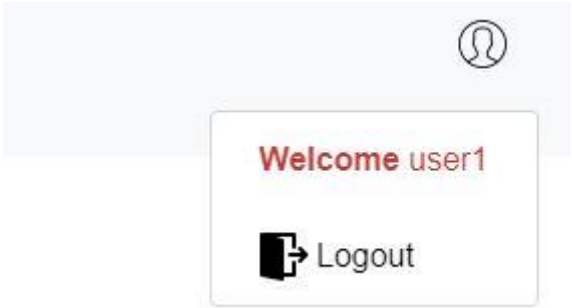
Blue color to the table represents the elements of examinations which are **below** the **minimum** valid value.
Red color to the table represents the elements of examinations which are **above** the **maximum** valid value.

Results of examinations

Chest Pain Type	Resting Blood Pressure	Serum Cholestrol	Fasting Blood sugar	Resting Electro cardiographic	ST-T wave abnormality	Ventricular hypertrophy	Maximum Heart Rate	Oldpeak	Vessels by flourosopy	Thal	Res
2.0	160.0	229.0	0.0	1.0	1.0	1.0	129.0	2.6	2.0	7.0	Pre

Furthermore, the navigation bar includes the Charts option. In the specific part, patients have the opportunity to see the results of the latest (up to 5) examinations in form of charts.

Also, doctors and patients are able to logout via the navigation bar in the right side.



17. Installation Instructions

17.1. Install Python

Initially, users should browse to the specific url:
https://www.python.org/downloads/windows/

in order to find the file which is called: Windows x86-64 executable installer (64-bit system) or Windows x86 executable installer (32-bit system) and download it.

Note that Python 3.5.0 *cannot* be used on Windows XP or earlier.

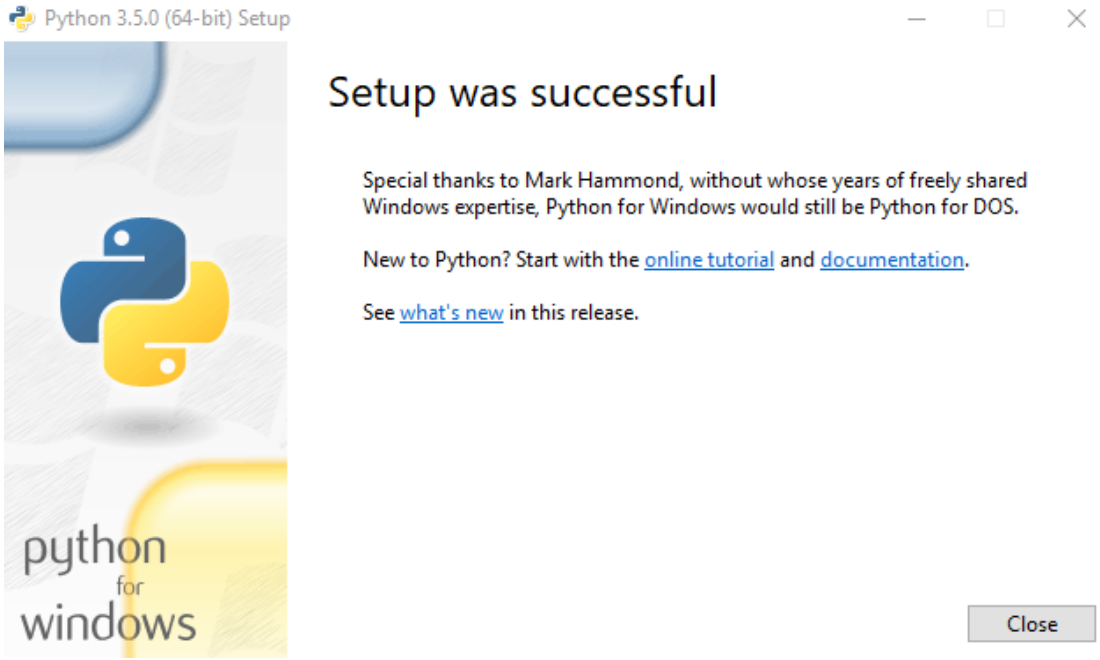
- Download Windows help file
- Download Windows x86-64 embeddable zip file
- Download Windows x86-64 executable installer
- Download Windows x86-64 web-based installer
- Download Windows x86 embeddable zip file
- Download Windows x86 executable installer
- Download Windows x86 web-based installer

Once it is downloaded, click the executable file and in the popup window click the checkbox: **Add Python 3.5 to PATH** and afterward, **Install Now**.

IMPORTANT! Maintain the below path during the installation phase:

C:\Users\Admin\AppData\Local\Programs\Python\Python35

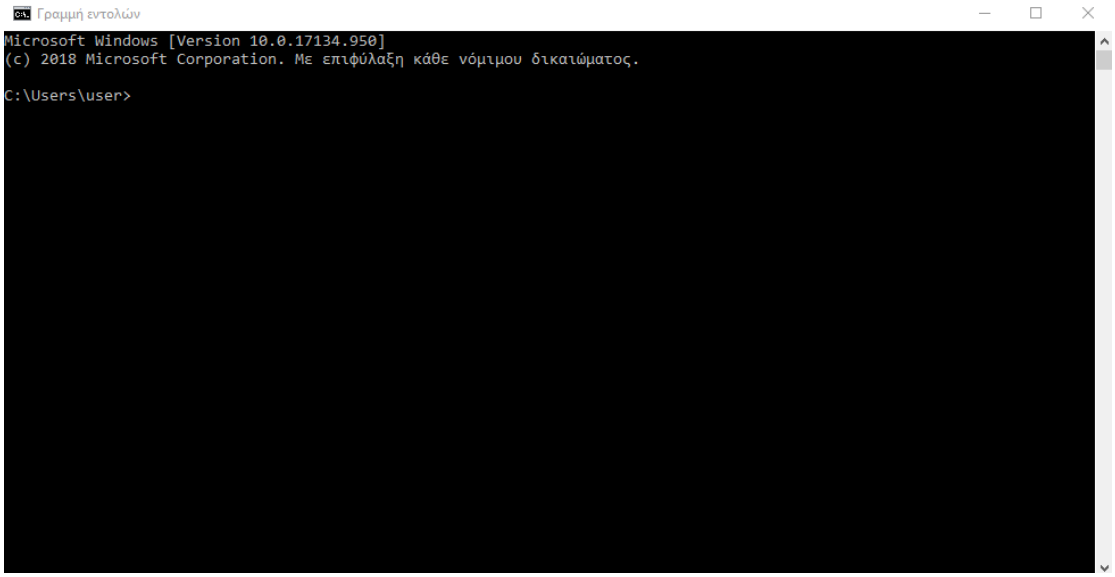
After the successful Setup, it would be displayed the below popup window:



and click, **Close**.

17.2. Install Packages

The first step is to open the **command prompt**.



Afterward, users should insert the below commands in sequence in order to install the required packages:

- python -m pip install --upgrade pip
- pip install pip==19.0.3
- pip install setuptools==40.8.0
- pip install numpy==1.16.2
- pip install -U scikit-learn==0.20.3
- pip install scipy==1.2.1
- pip install pyparsing==2.4.0
- pip install kiwisolver==1.0.1

- pip install pandas==0.24.2
- pip install pytz==2018.9

And close the command line.

17.3. Install Database – Locally

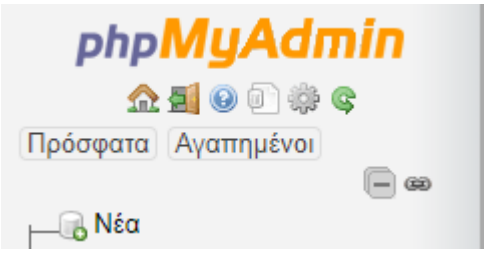
Initially, users should browse to the specific url in order to download XAMPP based on their systems (Windows, Linux e.t.c).

https://www.apachefriends.org/download.html

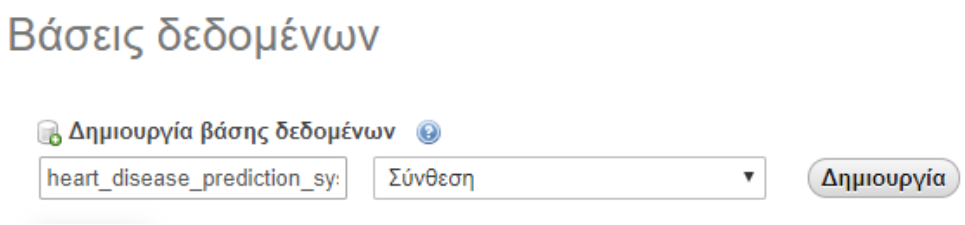
After the completion of the XAMPP installation, users should browse to the specific url:

http://localhost/phpmyadmin/

and click on the left side of the screen **New.**

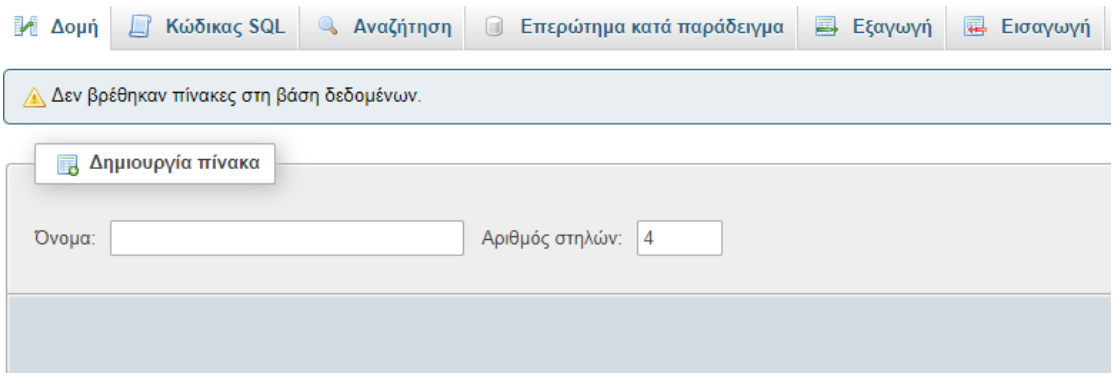


In the next page, users should write **heart_disease_prediction_system** to the input field.

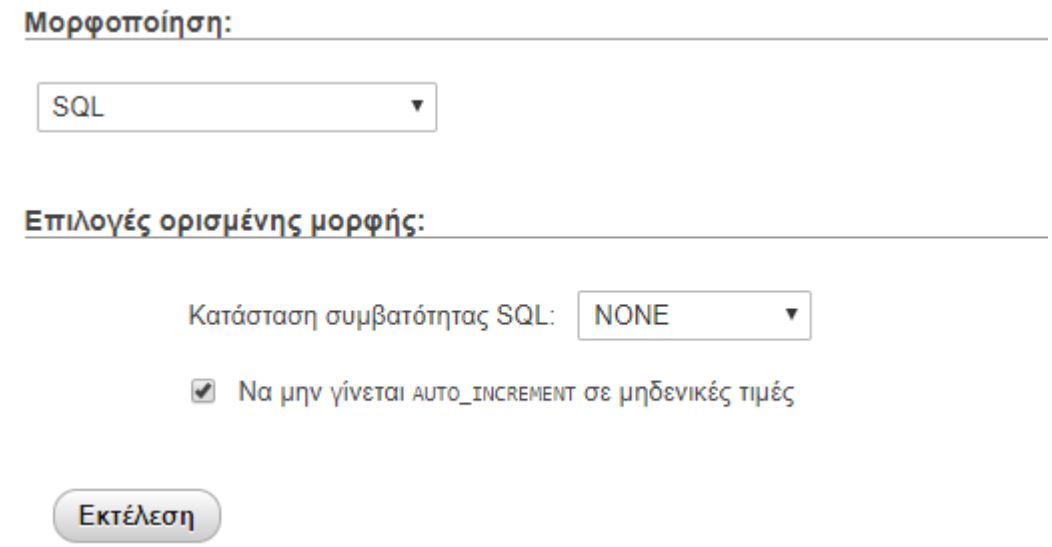


And click **Create.**

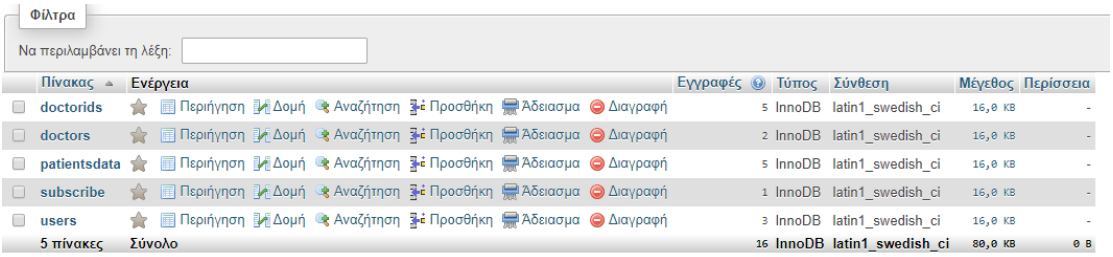
Click on the navigation bar **Insert**



In the next page, click **Choose file** and select the file which is called: **heart_disease_prediction_system.sql** and click **Execute**.



After the successful insertion, users would see the following page:



This means that everything was completed successfully.

Conclusions and Future Work

As we can perceive, prediction of heart disease is one of the most major challenges in the health care field and systems. The purpose of this research and its implementation attempts to provide a perception about a data mining classification algorithm which is called SVM classifier, along with its advantages and drawbacks. It is utilized a data set of 303 records which is called Cleveland Dataset and we have selected 13 attributes in order to make predictions. According to the results and based on various researches, it is proved that SVM classifier has a great accuracy in comparison with Naive Bayes, Decision Tree and ANN in the training and testing phase. Furthermore, we should mention that only Support Vector Machine (SVM) has been implemented in technical part in the specific paper. There are several improvements which could assist to enhance the scalability and accuracy of the specific prediction system. However, this study could be further extended in order to make breakthroughs for more accurate systems in hospitals. The performance of heart disease's diagnosis is likely to be improved to a great extent by estimating many class labels during the prediction procedure and make more reliable results. Therefore, it is able to improve the capabilities of traditional techniques, eliminate drastically the error rate of the human interference and, in general, provide unlimited better circumstances in medical diagnosis and prediction.

Acknowledgements

The specific research was supported by my professor Dimosthenis Kyriazis and my guide Argyro Maurogiorgou who provided excellent guidance, inspiration and recommendations and they played a primary role in order to obtain extraordinary experiences throughout the project and broaden my mind in the health care field. Their professional occupation with this research has provoked to be engaged actively with health care systems and develop my intellectual maturity which is important during my career in the foreseeable future. Finally, I would like to express my gratitude to the University of Piraeus and the Departure of Digital Systems which assisted to obtain knowledge about technology, enhance soft and hard skills and meet the magnitude of the specific scientific field.

References

[1] https://hal.archives-ouvertes.fr/hal-01826700/document
[2] http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0101-74382016000200321
[3] https://www.dovepress.com/ensemble-approach-for-developing-a-smart-heart-disease-prediction-syst-peer-reviewed-fulltext-article-RRCC#
[4] https://www.ijariit.com/manuscripts/v2i3/V2I3-1141.pdf
[5] https://pdfs.semanticscholar.org/db02/e8dd8edc69786af0cf969f8eca2b7bdb0426.pdf
[6]https://www.academia.edu/5692677/Heart_Disease_Prediction_System_Using_SVM_and_Naive_Bayes_

[7]https://www.researchgate.net/publication/326200298_A_NOVEL_METHOD_TO_PREDICT_HEART_DISEASE_USING_SVM_ALGORITHM

[8]https://www.ijcaonline.org/archives/volume156/number2/kanikar-2016-ijca-912368.pdf
[9]http://www.jatit.org/volumes/research-papers/Vol12No1/1Vol12No1.pdf
[10]https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3092139/
[11]http://www.ijmlc.org/vol5/544-C039.pdf
[12]https://www.geeksforgeeks.org/basic-concept-classification-data-mining/
[13]https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/classify.htm#i1005746
[14]https://searchenterpriseai.techtarget.com/definition/supervised-learning
[15]https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/

[16]https://towardsdatascience.com/supervised-machine-learning-classification-5e685fe18a6d

[17]https://towardsdatascience.com/unsupervised-machine-learning-9329c97d6d9f

[18]https://whatis.techtarget.com/definition/unsupervised-learning
[19]https://www.datarobot.com/wiki/unsupervised-machine-learning/
[20]https://www.codecademy.com/articles/normalization
[21]https://arxiv.org/ftp/arxiv/papers/1503/1503.06462.pdf
[22]https://t4tutorials.com/decimal-scaling-normalization-in-data-mining/
[23]https://www.digitaltrends.com/cool-tech/what-is-an-artificial-neural-network/
[24]https://www.ibm.com/support/knowledgecenter/en/SSEPGG_9.7.0/com.ibm.im.model.doc/c_decision_tree_calssification.html

[25] http://dataaspirant.com/2017/01/30/how-decision-tree-algorithm-works/
[26] https://www.slideshare.net/ashrafmath/naive-bayes-15644818
[27] https://www.linkedin.com/pulse/artificial-neural-networks-advantages-disadvantages-maad-m-mijwel

[28] https://learning.maxtech4u.com/decision-trees/
[29] https://statinfer.com/204-6-8-svm-advantages-disadvantages-applications/
[30] https://pythonprogramming.net/linear-svc-example-scikit-learn-svm-python/

[31] https://towardsdatascience.com/dive-into-pca-principal-component-analysis-with-python-43ded13ead21

[32] https://scikit-learn.org/stable/tutorial/basic/tutorial.html
[33] https://shuzhanfan.github.io/2018/05/understanding-mathematics-behind-support-vector-machines/
[34] https://towardsdatascience.com/demystifying-maths-of-svm-13ccfe00091e

[35] https://www.geeksforgeeks.org/python-pandas-dataframe/

[36] https://towardsdatascience.com/a-simple-example-of-pipeline-in-machine-learning-with-scikit-learn-e726ffbb6976

[37] https://towardsdatascience.com/train-test-split-and-cross-validation-in-python-80b61beca4b6

[38] https://pandas.pydata.org/
[39] https://www.datacamp.com/community/tutorials/python-numpy-tutorial
[40] https://machinelearningmastery.com/a-gentle-introduction-to-scikit-learn-a-python-machine-learning-library/

[41] https://www.codecademy.com/articles/scikit-learn

[42] https://towardsdatascience.com/hands-on-introduction-to-scikit-learn-sklearn-f3df652ff8f2

[43] https://www.geeksforgeeks.org/python-introduction-matplotlib/
[44] https://realpython.com/python-matplotlib-guide/
[45] https://www.edureka.co/blog/python-matplotlib-tutorial/
[46] https://realpython.com/python-itertools/

[47] BERIKOL GB 2016. Diagnosis of acute coronary syndrome with a support vector machine. Journal of Medical Systems