REPUBLIC OF TURKEY
YILDIZ TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

PREDICTION OF CARDIAC FAILURE USING ARTIFICIAL INTELLIGENCE METHODS

KINAN MORANI

MSc. THESIS
DEPARTMENT OF CONTROL AND AUTOMATION
PROGRAM OF CONTROL AND AUTOMATION

ADVISER
ASSOC. PROF. DR. ŞEREF NACİ ENGİN

ISTANBUL, 2018

REPUBLIC OF TURKEY
YILDIZ TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

PREDICTION OF CARDIAC FAILURE USING ARTIFICIAL INTELLIGENCE METHODS

A thesis submitted by Kinan MORANI in partial fulfillment of the requirements for the degree of **MASTER OF SCIENCE** is approved by the committee on 01.06.2018 in Department of Control and Automation, Control and Automation Program.

Thesis Adviser

Assoc. Prof. Dr. Şeref Naci ENGİN

Yıldız Technical University

Approved By the Examining Committee

Assoc. Prof. Dr. Şeref Naci ENGİN

Yıldız Technical University ______________________

ACKNOWLEDGEMENTS

The author thankfully acknowledges the support of Prof. Gyorgy Eigner and Prof. Tamas Ferenci in the Electric and Informatics faulty at Óbuda University, Budapest Hungary. I would like to also present my appreciation and gratitude to Assoc. Prof. Dr. Seref Naci Engin for his advisory at Yildiz Technical University, Istanbul Turkey.

May, 2018

Kinan MORANI

TABLE OF CONTENTS

LIST OF SYMBOLS

ω	The weight
Δ	The error
o	The output
ξ	The extra error
∅	The separable function

LIST OF ABBREVIATIONS

AUC	Area Under Curve
ANN	Artificial Neural Network
CV	Cross Validation
DT	Decision Trees
PCI	Percutaneous Coronary Intervention
RF	Random Forest
ROC	Receiver Operating Characteristic
SVM	Support Vector Machine

LIST OF FIGURES

Page

LIST OF TABLES

ABSTRACT

PREDICTION OF CARDIAC FAILURE USING ARTIFICIAL INTELLIGENCE METHODS

Kinan MORANI

Department of Control and Automation

MSc. Thesis

Adviser: Assoc. Prof. Dr. Şeref Naci ENGİN

The thesis work presents the study carried out on a relatively small dataset -with 1099 samples and 20 attributes- obtained from hospital records in Hungary. Various data analysis methods have been applied and their advantages and shortcomings have been presented. It goes to prove that using a tuned Support Vector Machine model brought in better predicting results in terms of accuracy and compatible cost to a classification problem when compared to Neural Nets, Random Forest or the Decision Tree models.

The thesis makes use of data analysis methods developed in the software package R and compares the forecasting and performance of vastly used technologies in prediction problems including ANNs, RF and SVM models. The results show that taking SVM into consideration while doing predictions for medical diagnoses and other types of applications has been effective. Generally depending on the dataset and the task in hand, one must try various methods before settling on the model that serves the task best.

With the results obtained from this thesis work and further similar work, hospital staff can be equipped better to deal with admitted patients as they can be more informed about the patients' conditions and can have some predictions about their conditions, which can play a vital role in the patient's survival and recovery. As it is about human lives, a small mistake can be hideously risky, thus more and more accuracy and cost effectiveness are required when modelling and predicting.

In this thesis work highly accurate predictions were obtained when using the SVM model. Moreover, an effective performance measured by the ROC has been reached with the SVM model. So for the given task the SVM model proved to be better than ANNs, RF or even DT models.

There is another important issue to check before jumping into the modeling which is the software package that should be used. The thesis work has been done using the R language on R Studio interface. However, Matlab and Python can be used. The thesis work presents R as an adequate language and RStudio as an effective interface.

Key words: Artificial Neural Networks, Random Forest, Support Vector Machine, Decision Tree, Area Under Cover.

ÖZET

YAPAY ZEKA YÖNTEMLERİ KULLANILARAK KALP YETERSİZLİĞİNİN TAHMİNİ

Kinan MORANI

Kontrol ve Otomasyon Mühendisliği Anabilim Dalı

Yüksek Lisans Tezi

Tez Danışmanı: Doç. Dr. Şeref Naci ENGİN

Bu tez çalışmasında, Macaristan'daki hastane kayıtlarından elde edilen 1099 örneklem ve 20 özellikli nispeten küçük bir veri üzerinde yapılan çalışmayı sunmaktadır. Çeşitli veri analiz yöntemleri uygulanarak, avantaj ve dezavantajlar sunulmuştur. Ayarlanmış bir SVM modelinin, ANNs, RF veya karar ağacı ile kıyaslandığında bir sınıflandırma probleminin doğruluğunu maliyet açısından daha iyi bir tahmin sonuçları olduğunu göstermektedir.

Tez, R yazılım paketinde geliştirilen veri analiz yönteminden faydalanarak ve Yapay Sinir Ağları, RF ve SVM modelleri de dahil tahmin problemlerinde en çok kullanılan teknolojilerin tahminlerini ve performanslarını karşılaştırmaktadır. Sonuçlar, tıbbi teşhisler ve diğer uygulama türleri için tahminler yaparken SVM'nin ne kadar önemli olduğunu göstermektedir.

Bu çalışmadan ve benzer çalışmalardan elde edilen sonuçlara göre hastane personeli daha iyi ekipmanlar kullanarak kabul edilen hastalara daha kaliteli hizmet sunabilirler. Hastaların hayatta kalma ve iyileşmesinde hayati bir rol oynar. Küçük bir hatanın çok riskli olabileceği, insan hayatıyla ilgili olduğu için, bu tür tahminler yapılırken doğruluk ve maliyetin çok daha üst seviyelerde olması gerekmektedir.

Bu çalışmada SVM kullandığında büyük tahminlere ulaşılmıştır. Ayrıca, SVM modelini kullanarak ROC tarafından ölçülen üst düzey bir performans elde edilmiştir.

Eldeki veri kümesinin en iyi modelini seçmenin ne kadar önemli olduğunu yukardaki paragraflardan görebiliriz. Kullanılması gereken R yazılım paketini, modellemeye geçmeden önce kontrol edilmesi gereken bir konudur. Tez çalışması R stüdyosunda R dili kullanarak yapılmıştır. Ancak Matlab ve Python da kullanılabilir. Tez verilen veri kümesi için R’yi yeterli bir yazılım adayı olarak sunmaktadır.

Anahtar Kelimeler: Yapay Sinir Ağları, Destek Vektör Makineleri, Karar Ağaçları, Kapak Altında Alan, Rastgele Tahmini.

CHAPTER 1

INTRODUCTION

1.1 Literature Review

Reported by Barrados, M. et al. [1] and many other researches, it is safe to state that today’s health care systems are highly automated and producing more and more data about the actual status of the patients each year. Due to such high amount of data, advanced analysis and soft computing methods needed be used to provide prompt views about newly admitted patients’ conditions and forecast the possibly occurring changes in their status in the future. These forecasting techniques are quite beneficial, since via these techniques and decision support serious issues and suddenly occurring acute problems in patients’ conditions can be prevented [2].

The potential benefit of the use of these methods is unequivocal especially regarding cardiac investigations and similar problems in which the odds for the patient’s death are quite high [3].

1.2 Objective of the Thesis

The aim of this thesis is to inspect and find some solutions to a given problem. In this thesis work, a piece of research was conducted for the purpose of getting more accurate results in predicting the statues of admitted patients to a hospital based on their medical conditions and diagnoses such as the patient’s sex, smoking habits, and other related diagnoses.

The thesis work presents many of the software packages and methodologies that one might encounter as they work out data analysis and forecasting problems starting from issues to be considered to the main modeling methods that could be used for prediction. The thesis presents an experiment that was conducted on a medical dataset to solve a problem of predicting the statue of newly admitted patients to a hospital with the help of analyzing data records of previous patients in that hospital. With such predictions certain actions and cautions can be taken on newly admitted patients and thus more appropriate treatments can be offered to the patients.

1.3 Hypothesis

Different software packages can be applied in healthcare systems to work out classification problems. There are different programming languages and software packages which can help predict outcomes including Matlab, R and Python. However, the programming language chosen for the given dataset in this thesis was the R language. Chapter 2 introduces the advantages and disadvantages of the R language in terms of predicting problems and shows why R was preferred for forecasting on the dataset of 1099 samples and 20 attributes.

There are quite a lot of database management and design software packages including Microsoft Excel, SQL, and Oracle and so on. The second part of chapter 2 points out the benefits of using the Microsoft Excel program for managing the data. The chapter presents the Matlab as a software tool for data analysis and then presents Python as another tool explaining why R language was favorite over Matlab and Python.

On the other hand, different predicting models can be applied to the datasets to predict the outcomes. Choosing the right model is quite crucial to help predict with more accuracy and with as less time consumption as possible. Chapter 3 introduces four of the frequently used methods for classification problems, which are Artificial Neural Networks, Support Vector Machine, Decision Trees and Random Forest Models. These models were introduced theoretically in chapter 2 before being used for the hospital dataset in chapter 4.

In chapter 4, an experiment was conducted on one specific dataset obtained from a Hungarian hospital. Chapter 4 provides explanations and details of the attributes used in the dataset and how the data was prepared and manipulated before modeling.

Chapter 4 also shows the prediction results obtained from using four different modeling methods, ANNs, RF, SVM and Decision Trees, with useful tuning applied to the models. Looking at the results, Chapter 4 explains that the SVM model for the given dataset gave relatively more accurate results with fairly less time consumption in comparison to the other three models.

Last chapter summarizes the results and theories with further discussions regarding how to improve the diagnosis for the given dataset. The chapter presents several ideas to advice the given research with regard to using other modeling methods or obtaining longer datasets. It ends with further projections as follow-ups to the thesis work.

CHAPTER 2

THE SOFTWARE PACKAGES USED

There are many databases, Excel, SQL, and Oracle and so on, which can be used to manage and design the collected data. On the other hand, there are programming languages to prepare and make use of the datasets including R, Python and Matlab. These programs serve statistical analysis purposes.

Matlab is one choice that could be considered for data analysis. R and Python are other significant choices for statistical data analysis. However, it is fair to say that the chosen programming language depends on the background of the user and the datasets in hand [4] [5].

In this work, the chosen programming language was R. R Studio interface was particularly used for data analysis.

Microsoft Excel was chosen for data management.

2.1 R and RStudio

Some of the general advantages of using R are:

1. Free and growing in popularity; R is more or less number one software package for statistical analysis in the world [6].
2. It is available in a range of formats including Windows, Mac and Linux
3. There is a growing library of packages - now more than one thousand- that have been developed for all sorts of applications including libraries for loading data such as 'RMySQL' and 'xlsx', libraries to manipulate data such as 'dplyr', libraries to visualize data such as 'ggplot2', libraries to model the data such as

'car', 'randomForest' and 'caret', others to report results such as 'shiny', libraries to write high R performance codes, to work with the web and still others to write your own R packages [7].

4. Excellent built-in graphics and analysis functions [8].
5. There are lots of online manuals and tutorials to help learn programming with the R language.

Some of the disadvantages of using R are:

1. It is largely - but not exclusively - command lines rather than pull down menus, in that R is less of a windows milieu.
2. R has limited symbolic analysis, in that it could be hard to do proper math such as symbolic integrations or symbolic differentiations [9].
3. R is rather slow in executing complex codes. For very detailed simulations one might want to consider another programming language [10].
4. When errors are made in R, they can be difficult to spot.

R is adequate for the data set in hand for the following reasons:

With all the previous advantages stated in the R programming language and being working on a relatively small and less complex dataset in which many of the stated disadvantages can be overcome, R language was chosen for the presented thesis work.

One of the significant things about the R language is the thousands of users who developed packages and tools in various disciplines as stated in the previous section; analyzing everything from weather or financial data to the human and analyzing computer security-breach data as well. With all the library packages and functions that come with the R software, many users can quite easily prepare the data and solve math problems in any scripts.

R is consistent and strict on punctuation or what is called the fussy property of R. At first beginners of the R might struggle with it but it could be quite useful as users get used to it. However, it helps make the programmer more precise and detailed while typing their codes.

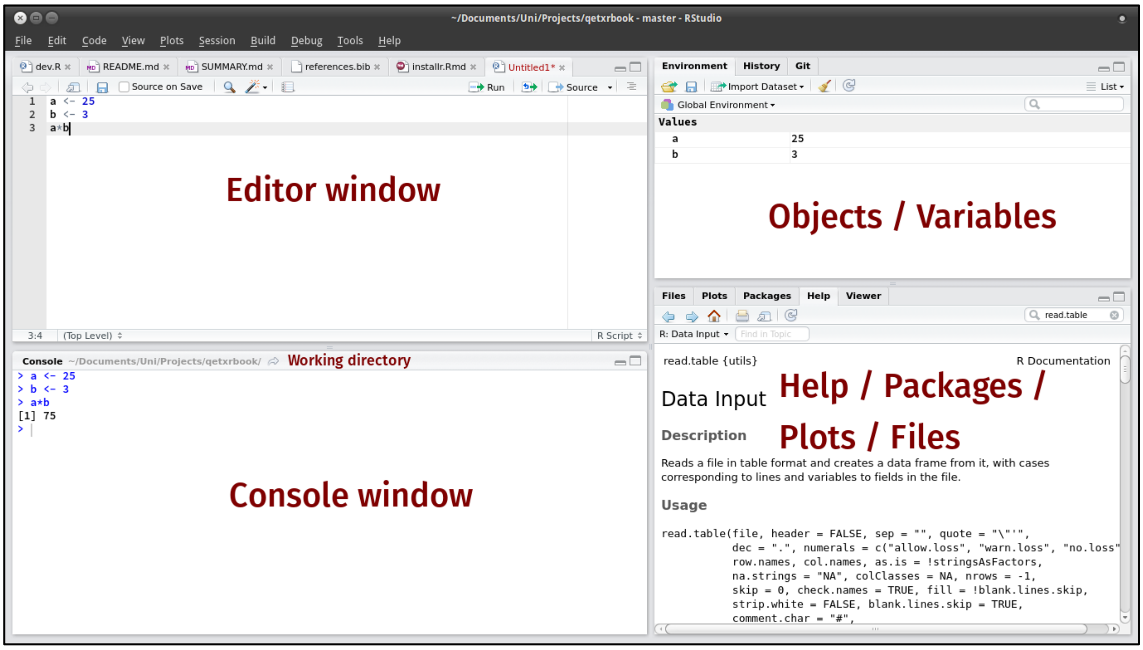
RStudio interface does not only have work space but also a script editor which is handy, because the script can be built up before running it and more importantly R offers the availability of sharing the scripts with others to show how you have conducted certain analysis or a process.

R programming language offers a verity of interfaces. The more often used one and the one used for the thesis work is the R Studio. Figure (2.1) shows the R interface with all its parts.

2.2 Excel for Data Management

Excel is quite a powerful instrument often used in biomedical research as a support for datasets. It contains some basic statistical analysis and could serve for some simple

Interfaces [11] [12].


~/Documents/Uni/Projects/qetxrbook - master - RStudio
File Edit Code View Plots Session Build Debug Tools Help
dev.R README.md SUMMARY.md references.bib installr.Rmd Untitled1*
Source on Save
Run
Source
Editor window
3:4
(Top Level)
R Script
Environment History Git
Import Dataset
List
Global Environment
Values
Objects / Variables
Files Plots Packages Help Viewer
read.table
R: Data Input
Find in Topic
Help / Packages /
Plots / Files
Console ~/Documents/Uni/Projects/qetxrbook/
Working directory
Console window

```
a <- 25
b <- 3
a*b
```

a	25
b	3

```
> a <- 25
> b <- 3
> a*b
[1] 75
>
```

read.table {utils} R Documentation

Data Input

Description

Reads a file in table format and creates a data frame from it, with cases corresponding to lines and variables to fields in the file.

Usage

```
read.table(file, header = FALSE, sep = "", quote = "\"'",
           dec = ".", numerals = c("allow.loss", "warn.loss", "no.loss"
           row.names, col.names, as.is = !stringsAsFactors,
           na.strings = "NA", colClasses = NA, nrows = -1,
           skip = 0, check.names = TRUE, fill = !blank.lines.skip,
           strip.white = FALSE, blank.lines.skip = TRUE,
           comment.char = "#",
```

Figure 2.1 RStudio Interface

Excel is excellent for data management and preprocessing raw data prior to more demanding data analyses. The graphics and some of the basic statistics such as t-tests, F-tests, correlations and even regressions are quite easily conducted in Excel with little to no data formatting beyond what would be necessary for any statistical package. On the other hand, Excel cannot handle unbalanced designs larger than one-way and cannot handle something larger than a two-way factorial design [13].

In this thesis's experiments, the dataset was provided and managed in Excel. Eventually the data was saved in a 'csv' format file. Thus the data was prepared to be read by the

RStudio software which provides further statistical analysis tools useful for the given task.

2.3 Matlab

With Matlab and its data analysis products it is possible to perform analysis and gain insight into the data in a fraction of the time required by spread sheets or traditional

Programming languages like C, C++ or Visual Basic [14].

Matrix parts combine a powerful numeric engine and programming environment with interactive tools for data analytics allowing development of applications faster and without being an expert in data analysis, statistics or machine learning[15][16].

Matlab also allows visualization of data with a multitude of different plots. There are ways to customize your plot where you can label your plots using rich text label and text support for Greek characters. Matlab has also date time tick labels and Axis tick labels that rescale when zooming [17].

As mentioned in section (2.1) and with comparison to what is briefly stated in this section, it can be concluded that Matlab is quicker and more user-friendly in terms of interfaces for junior programmers but it is less statistics with comparison to R. Thus, in general, R was preferred on the given dataset.

2.4 Python

Python is definitely one of the best common programming languages in the world. Many big companies have been dealing with Python as a programming language for quite a long time; Google uses python on web search and YouTube company is largely conducted by Python and many other companies use Python such as IBM, Facebook, Instagram and so on [18].

Python language is known to be simple and easy to learn as it closely resembles the English language; very English simple lines can sometimes replace many long and machine-like lines in other languages such as java, C++ and others.

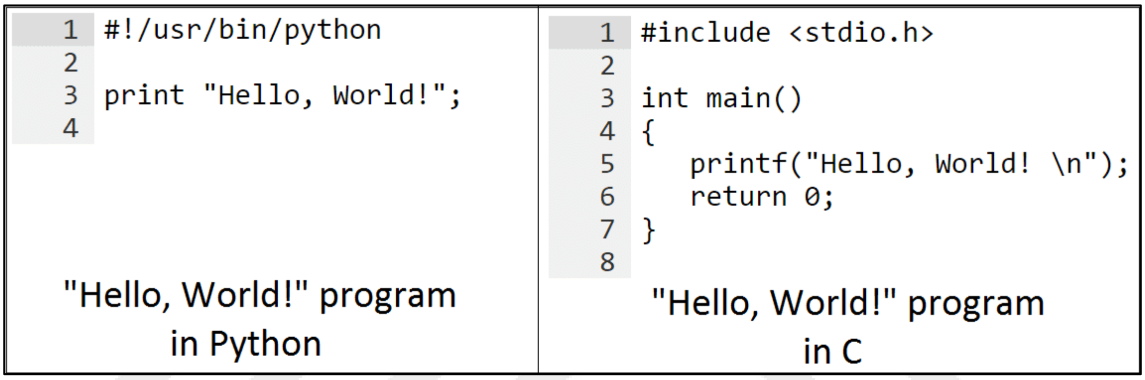
Figure (2.2) shows a comparison between C and Python codes to produce the same outputs, it shows how simple writing a code in Python when compared to the writing it in the C language.

Regarding big data and statistical analysis, Python supports parallel computing and it has many libraries that users can find to handle big data such as PYDOOP, DASK and PYSPARK.

A Map reduce code can be written and processed in Python. For data manipulation and analysis python has developed the numerical engines such as NumPy and Pandas.

Python in those libraries deals with Tabular data and arbitrary matrix data [19].

For this thesis work the properties of the RStudio as stated in section (2.1), mainly the shareable-script property, made the choice of using the R language over Python.



```
#!/usr/bin/python

print "Hello, World!";

```

"Hello, World!" program in Python

```
#include <stdio.h>

int main()
{
   printf("Hello, World! \n");
   return 0;
}

```

"Hello, World!" program in C

Figure 2.2 The simplicity of a python code in comparison to C

CHAPTER 3

THE SOFTWARE TOOLS FOR COMPUTING METHODS

This chapter introduces the main types of machine learning algorithms and presents the main machine learning models including Artificial Neural Networks, Radom Forest, Support Vector Machine and Decision Trees.

The chapter explains the general approaches for all of the models.

3.1 Main Types of Machine Learning Algorithms

Machine learning can be split in to two main types of learning:

The unsupervised learning when input data is used but the outcome points are unknown. In this case, clustering algorithms can be created to make sense of the data and make inferences based on the relative positions on the data points.

The supervised learning is when both the inputs and the outputs are known. In this case regression algorithms are used such as regression algorithms and curve fitting for continuous dataset and classification algorithms can be used for discrete classes or classification [20].

Figure (3.1) shows the supervised and unsupervised learning structure.

In the given task, hospital records already exist and some outcomes of the patients are known and thus in this dataset both the outputs and the inputs are given. Therefore, it is quite clear that we are leading a supervised learning task.

On the other hand, as the outcomes are concrete where they take only two values, which are weather the patient survived or died, it is clear that we are dealing with a classification problem.

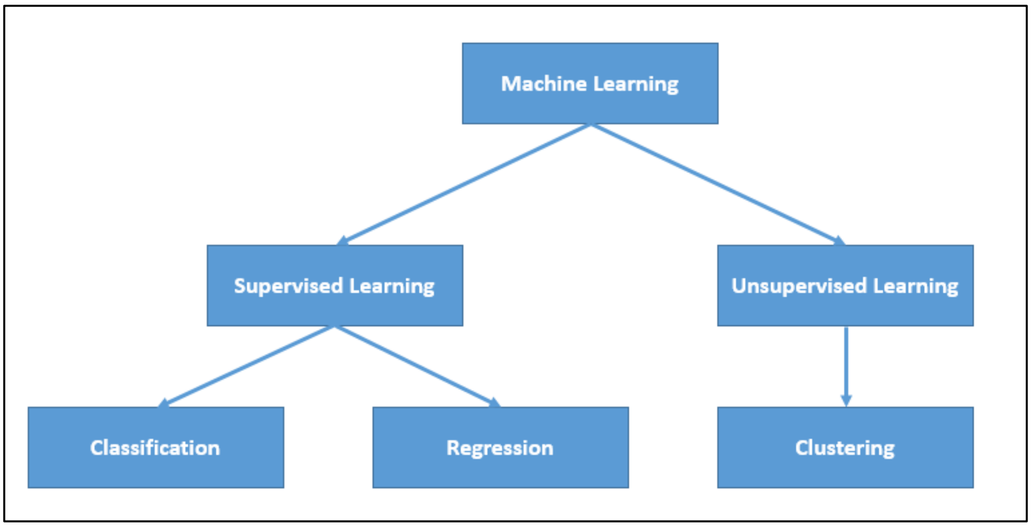


Figure 3.1 Supervised and Unsupervised Learning

3.2 Artificial Neural Networks Modeling

Neural networks are inspired by the brain which has neurons that are vastly and densely connected.

The human brain has about one hundred billion cells called neurons. Neurons are connected to each other through pathways that transmit electrical signals. This connectivity’s allow the neurons to send and receive electrical impulses which internally responsible for the brain's functions [21].

So each cell of the brain can be seen as a mathematical function f(x) with inputs and outputs, when it receives electrical impulses from a cell, it sends it to another cell that is connected to. Fıgure (3.2) shows a general structure of the Artificial Neural Network with one hidden layer.

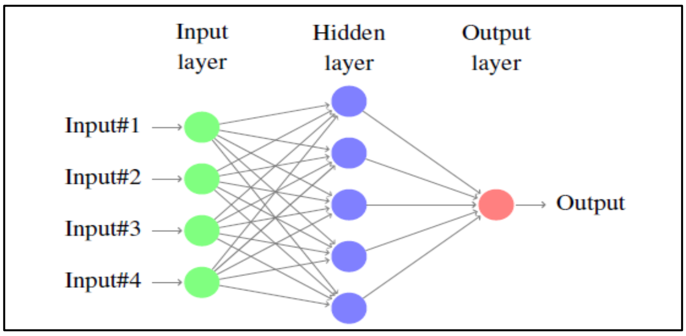


Figure 3.2 A general ANNs structure with one hidden layer

3.2.1 Introduction to Artificial Neural Networks

The ANN is just an attempt to model the brain. The reason is - like the brain- the neural nets can be expected to do some of the things the human brain can do such as pattern recognition. Just as in the brain the neuron in the artificial neural networks takes input and sends output through their connections. But the neural nets should be ordered in a way that they can compute something. In order to do this, it is designated to have 'input' cells and 'output' cells. The hidden cells can be in more than one layer.

The connections between layers can be strong or weak. And the whole neural nets can be seen as a function that takes in a few input numbers and output a few output numbers and in between it does some computations with hidden layers.

Looking at an individual cell - neurons -; each cell is a function by itself that takes some inputs from other cells and it also considers how strong those connections of those cells are, and then it gives an output. For a given cell we will multiply each input signal by that input strength and then all of these will be added together as a final input. In this way, weak inputs are multiplied by a very small connection strength number so they are almost zero, this means that they almost have no effect on the function, and also cells with strong connections strength numbers means that they have their inputs multiplied by high numbers and they have a large effect on the function as expressed in the Equation (3.1)

A lot of different functions can be used. The hyperbolic tanh function in general is used in this thesis work. The hyperbolic tanh function takes the number at the previous sum and then it outputs a number between -1 and 1 as illustrated in the Figure (3.3).

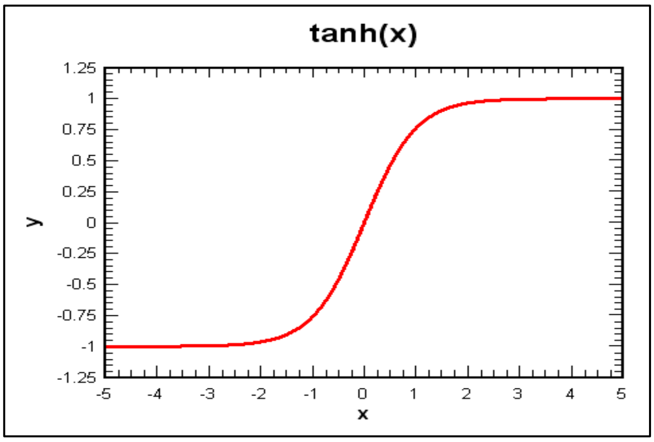


Figure 3.3 The hyperbolic tanh function

So each cell takes an input, multiplies it by its strength, takes the sum of these and then passes it through a function. It sends the sum to all the paths going to its output connections as expressed in Equation (3.1). The computer will do it to all the layers moving left to right. See Figure (3.4).

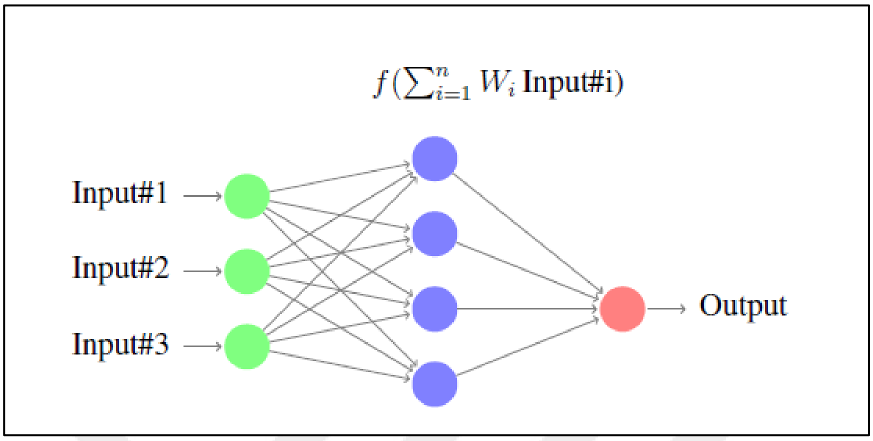


Figure 3.4 Every cell multiplies the inputs with their strengths, sums them up and passes them through a function to the output(s)

3.2.2 How the Neural Networks Learn

One of the best ways for network learning is called "back-propagation" method. This method adjusts the connection strengths until the network learns the dataset and it is measured by means of the error function.

Once the Network is given some inputs and their corresponding outputs -supervised learning-. The machine sets up some random connection weights. Each input is carried down the line to the output as explained in the previous section. And then the output of the neural networks is compared to the right given answers.

However, the output of the neural nets is usually not being the same as the right - real - answers. In that the error is calculated and the Back-propagation process carries on.

So this process includes taking the calculated error and going back to try to find the error in between the cells by multiplying the error at each cell with the connection strengths. As stronger connections are responsible for more errors and by multiplying the error with the strength of the connections, the stronger connections will end up with

higher error and the weaker ones will have very low error - almost negligible-. As the inputs are constant the Δ error is only calculated for the hidden and output layers [22].

In Figure (3.5); There are two inputs - input1 and input2-, one hidden layer with 3 neurons, and one output. The figure shows all weights.

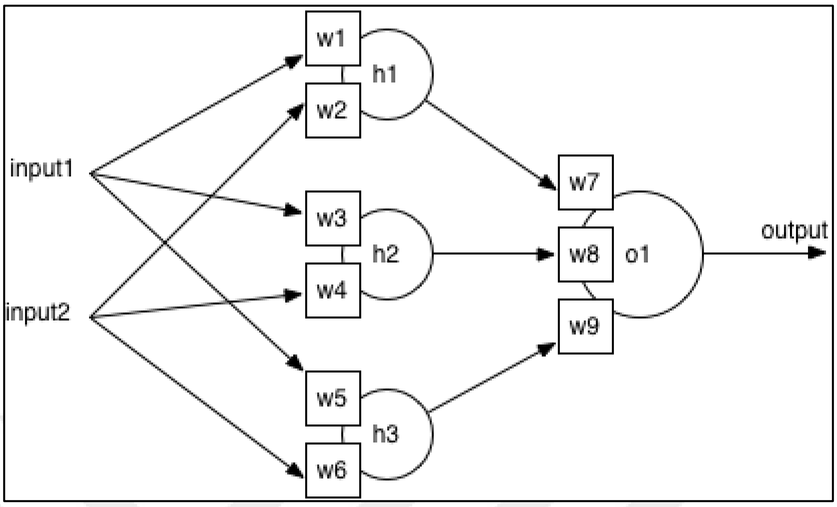


Figure 3.5 ANNs with 2 inputs, 1 hidden layer - 3 neurons and one output

Supposing that f ( ) is the function used for the neural nets and from Figure (3.5) it can be concluded that:

$$h1 = f\,(\,input1\ .\ \omega 1 + input2\ .\ \omega 2) \tag{3.1}$$

$$h2 = f\,(\,input1\ .\ \omega 3 + input2\ .\ \omega 4) \tag{3.2}$$

$$h3 = f\,(\,input1\ .\ \omega 5 + input2\ .\ \omega 6) \tag{3.3}$$

Where h1, h2, h3 are the final outputs of the neurons in the hidden layer as shown in the Figure (3.5).

The summations of the inputs and the passing through the f ( ) function have been done in the previous equations.

$$o1 = f\,(h1\ .\ \omega 7 + h2\ .\ \omega 8 + h3\ .\ \omega 9) \tag{3.4}$$

Where o1 is the output of the final layer as shown in Figure (3.5).

The equation (3.4) indicates the output o1.

If the general error is to be Δ , which is the difference between the actual output and the desired output. Then the three other errors in our example are calculated as follows:

$$\Delta 1 = \Delta\ .\ \omega 7 \tag{3.5}$$

$$\Delta2 = \Delta \, . \, \omega9 \tag{3.6}$$

$$\Delta3 = \Delta \, . \, \omega9 \tag{3.7}$$

Where $\Delta1$, $\Delta2$ and $\Delta3$ are the three calculated errors for the neurons in the hidden layers as in the Figure (3.5).

Those new errors inside the network are propagated back to calculate the new connection during the backpropagation is carried out as follows:

$$\omega1' = \omega1 + c \, . \, \Delta1 \, . \, h1' \tag{3.8}$$

$$\omega2' = \omega2 + c \, . \, \Delta1 \, . \, h1' \tag{3.9}$$

$$\omega3' = \omega3 + c \, . \, \Delta2 \, . \, h2' \tag{3.10}$$

$$\omega4' = \omega4 + c \, . \, \Delta2 \, . \, h2' \tag{3.11}$$

$$\omega5' = \omega5 + c \, . \, \Delta3 \, . \, h3' \tag{3.12}$$

$$\omega6' = \omega6 + c \, . \, \Delta3 \, . \, h3' \tag{3.13}$$

$$\omega7' = \omega7 + c \, . \, \Delta. \, o1' \tag{3.14}$$

$$\omega8' = \omega8 + c \, . \, \Delta \, . \, o1' \tag{3.15}$$

$$\omega9' = \omega9 + c \, . \, \Delta \, . \, o1' \tag{3.16}$$

Where $\omega1'$, $\omega2'$, $\omega3'$, $\omega4'$, $\omega5'$, $\omega6'$, $\omega7'$, $\omega8'$ and $\omega9'$ are the new weights after the backpropagation process is conducted, $h1'$, $h2'$, $h3'$ are the new outputs after the backpropagation process for all the neurons in Figure (3.4) respectively and $o1'$ is the new output of the final neuron after the backpropagation process, and thus all the new connections are calculated.

Note that c is the threshold that the input has to pass so that the input can go as either 0 or 1 to the next phase.

3.3 Support Vector Machine Modeling

After introducing the Artificial Neural Networks model, this section deals with another common tool of modeling. In this section the Support Vector Machine Model is introduced and the math behind the model is explained in details.

The section explains the SVM algorithms and the mathematics used for it.

3.3.1 Introduction to Support Vector Machine

A support vector machine (SVM) is a concept for a set of related supervised learning methods that analyze data and recognize patterns, used for classification and regression analysis. SVM model in a classification problem is basically finding the model that best splits the data. There are many hyperplanes that might classify the data. One reasonable choice as the best hyperplane is the one that represents the largest separation. In more technical phrases; SVM is finding the hyperplane - which is basically a line in two dimensional spaces - that best splits the data by being as far as possible from the support vectors - which are the nearest data pieces to the hyperplane. This is another way of stating that the margin has been maximized the as illustrated in the Figure (3.6) [23].

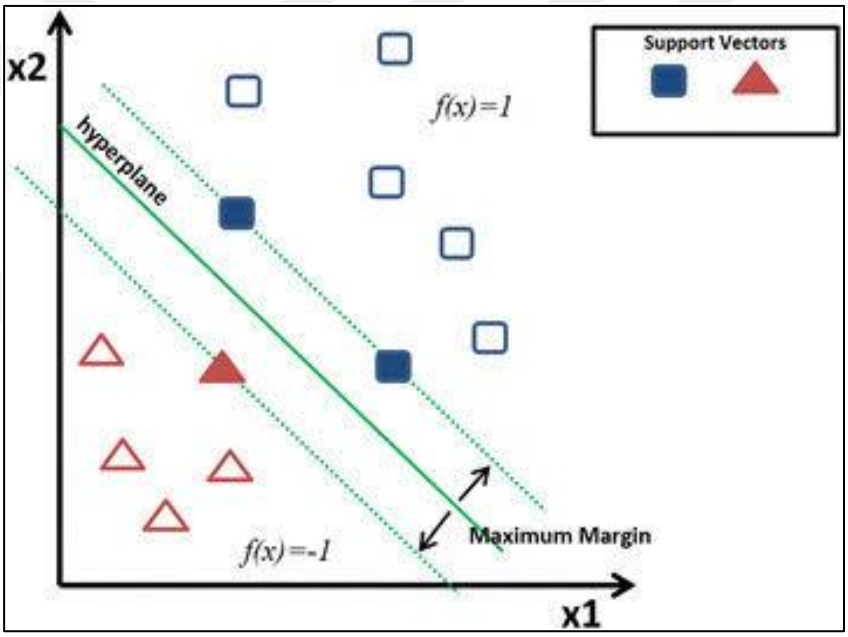


Figure 3.6 SVM showing the best hyperplane, the maximum margin and the support vectors

3.3.2 The Applied Mathematics of the SVM Model

If such a hyperplane - the one that gives the maximum margin -exists, it is known as the

Maximum-margin hyperplane and the defined linear classifier are known as a maximum-margin classifier or, equivalently, the perceptron of optimal stability.

By considering a two-dimensional classification problem as in the figure above. And defining:

$$F\,(1) = 1 \qquad (3.1)$$

$$F\,(1) = -1 \qquad (3.2)$$

The distance of a line from the origin is obtained by $\frac{|b|}{||W||}$ so, the distance between one of the marginal lines and the classifier line is obtained as:

$$\mathrm{d} = |\frac{|b-1|}{||W||} - \frac{|b-1|}{||W||} = \frac{1}{||W||} \qquad (3.21)$$

And therefore the width of the whole margin:

$$\mathrm{D} = 2d = \frac{2}{||W||} \qquad (3.22)$$

Now what is needed is to maximize the margin or minimize $||W||$:

$$\text{But } ||W|| = \frac{1}{2}\, W^T\, W \qquad (3.23)$$

Considering all the point in the dataset S $= x_1, y_{1,} x_2, y_{2,}\, x_m, y_m$ the goal is to minimize

$\frac{1}{2}\, W^T\, W$ Such that $\mathrm{y}_i(W^T\, x_i + b) \geq 1$ for i = 1, 2 ... m

We can also make the support vectors more flexible by taking the error ξ_i or each border line. Then the new optimization problem becomes:

Minimum of $\frac{1}{2}\, W^T\, W + c(x+a)^n = \sum_{i=1}^{m} \xi_i$ where i = 1, 2... m

such that: $\mathrm{y}_i(W^T\, x_i + b) \geq 1 - \ \xi_i$ where i = 1, 2, ...,m

And by inserting the constrains into the final equation, we get the following optimization problem:

Minimum of $\frac{1}{2}\, W^T\, W + \ \sum_{i=1}^{m} \alpha_i [\, y_i(W^T\, X_i + b) - 1\,]$ where α is the multiplier of the constraint.

Now to solve the optimization problem, we take the derivative of the equation over W and b and get each of the two to equal to zero. In that we get the following equations respectively:

$W - \ \sum_{i=1}^{m} \alpha_i y_i x_i \ = 0$ Or

$$\omega = \sum_{i=1}^{m} \alpha_i\, y_i\, x_i \qquad (3.24)$$

And

$$\sum_{i=1}^{m} \alpha_i\, y_i = 0 \qquad (3.25)$$

Now by inserting the final values of W and b we get:

Maximization of $-\ \frac{1}{2}\ \ \sum_{i=1}^{m}(\sum_{i=1}^{m} \alpha_i\alpha_j y_i y_j\ x_i^T\ x_j) +\ \sum_{i=1}^{m} \alpha_i$ where i, j = 1, 2, ... ,m

such that: $\sum_{i=1}^{m} \alpha_i\ y_i = 0$ where $\alpha_i\ \geq 0$

By mapping data to a higher dimensional feature space, we might get the data to be always linearly separable for the higher dimensional space mapping we use the function $\emptyset$

$$\text{Max} \sum_{i=1}^{m} -\frac{1}{2} \sum_{i,j=1}^{m} \alpha_i\, \alpha_j\, y_i\, y_j - k\left(x_i, x_j\right) \qquad (3.26)$$

Such that $\alpha_i\ \geq 0$ for every i in 1, 2, ...,m

$$\sum_{i=1}^{m} \alpha_i y_i = 0$$

Where

$$k\left(x_i, x_j\right) = (\emptyset\left(x_i, .\left(x_j\right)\right))$$

is called the ’kernel’ function.

The kernel functions commonly used in SVM’s formulations are:

1- Linear:

$$k\left(x_i x_j\right) =\ x_i^T\ x_j \qquad (3.28)$$

2- Polynomial:

$$k\left(x_i x_j\right) = (\alpha\ x_i^T\ x_j + r)^d \quad \text{where}\ \ \alpha\ > 0 \qquad (3.29)$$

3- Radial basis function (RBF):

$$k\left(x_i x_j\right) =\ \exp(\ |-\alpha||x_i||\ ^2\ ) \ \text{where}\ \ \alpha\ > 0 \qquad (3.30)$$

4- Sigmoid:

$$k\left(x_i x_j\right) =\ \tanh(\ |-||x_i||\ ^2\ ) \qquad (3.31)$$

Where $\alpha, r$and d are kernel parameters.

3.4 Decision Trees and Random Forest Algorithms

In medical decision making (classification, diagnosis, etc.) there are many situations Where decisions must be made effectively and reliably. Conceptual simple decision making models with the possibility of automatic learning are the most appropriate approach for performing such tasks. Decision trees are a reliable and effective decision making technique that provide high classification accuracy with a simple representation of gathered knowledge and they have been used in different areas of medical decision making. It was proposed to inject randomization in the learning process in order to create decorrelated trees.

By averaging the predictions, it was demonstrated that random forests achieve greater generalization and thereby superior accuracy [24].

3.4.1 Decision Trees Algorithms

Decision Trees learning is one of the most widely used and practical methods for inductive inference and is an important tool in machine learning and predictive analytics [25].

Some advantages of using decision tree models are that they are simple to understand and interpret, have value even with little data available; some of the decision random forest models are built on even experts describing a certain situation (its alternatives, probabilities and so on) and are robust to noise and capable of learning disjunctive expressions.

However, decision tree model comes with some disadvantages as well, in that if the data is having different levels then information gain in decision tree might be biased in favor of attributes with higher levels. Furthermore, if the dataset had many undefined values or if the outputs are linked to each other, then the structure of the decision trees could get really complex.

Decision tree is a tree-like structure. It generally consists of three types of node: the 'root' node at the beginning of the tree, the 'internal' nodes inside the tree and the Leaves at the end.

Figure (3.7) shows a decision tree structure with 7 nodes and 8 ends.

It can be noticed from Figure (3.7) that a decision tree's structure consists of a root node Q1, internal leaves Q2, Q3, Q4, Q5, Q6 and Q7 and also the leaves which are the ends.

The flow of any decision in the decision trees starts from the root and goes all the way down to the leave.

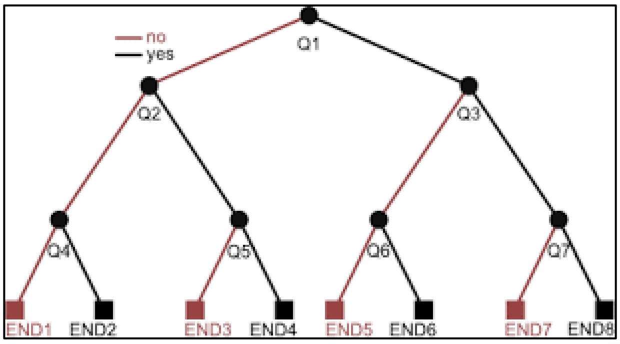


Figure 3.7 Decision Tree general structures with root, internal nodes and leaves

In terms of splitting the data in the decision tree model: a classification tree search through each independent variable to find a value of a variable that best split the data into two or more branches. In general, best split is the one that minimizes the impurity - misclassification - of the outcome in the resulting data subsets. And the splitting process is repeated until it finds any stopping criteria.

There are different splitting creations in decision tree models, some of those are:

The splitting could be based on a minimum number of observations. We could choose a specific number of observations for a split of data to happen.

The splitting could happen through what is called ' information gain', which is impurity based criteria that uses the entropy measure. Entropy, on the other hand, could be thought of as moving from ordered state to very disordered one, and the entropy would be the measure of the increase in disorder.

The Gini index is another criterion to split the data in decision tree models. The Gini index is another impurity based criteria that measures the divergence between the probability distribution of the target attributes' values. Gain ratio criteria are also usable for data splitting, which normalizes information gain by dividing in by the entropy.

Now these stopping criteria should be chosen carefully as it is possible to have under fitting trees in case of rigid stopping criteria or over fitting trees in case of lenient criteria.

The number of branches and nodes that should be used in the trees can be determined as follows:

To determine the tree's depth, we could use the "cost complex" parameter or Cp. In that we try to find the lowest Cp value which can give the appropriate tree size. In Figure

(3.8) Cp = 0.016 gives the best tree depth [26].

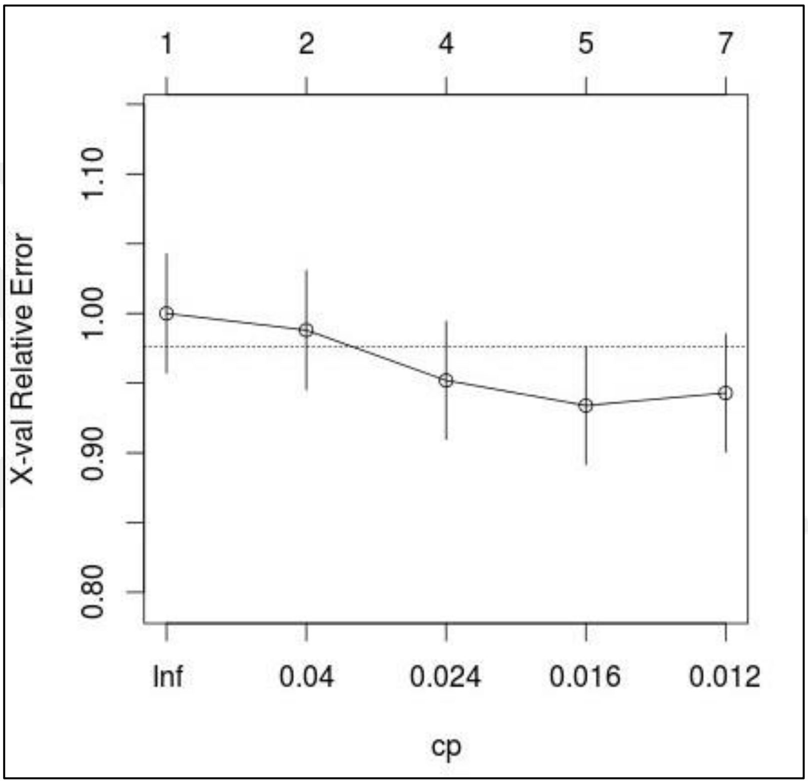


Figure 3.8 Cost Complex Parameter with tree depth = 0.016 fits best

3.4.2 Random Forest Algorithms

The random forest model performs both classification and regression tasks. As the name suggests, the random forest model creates a forest with a number of decision trees. In general, the more trees in the forest, the more robust the model is and thus the higher the accuracy is. The random forest uses the same algorithms as the decision trees such as the information gain, the Gini index approach or others.

So in the decision tree, to classify the data based on the given attributes each tree in the forest gives a vote for a class. The forest then chooses the most votes, and in case of a regression task it takes the average of the votes of the trees [27].

The advantages of using the random forest model:

1. It can perform both classification and regression tasks.
2. It can better handle the missing values and still maintain accuracy for the missing data.
3. It does not overfit the model.
4. It can deal with large datasets.

However, Random forest model still comes with disadvantages:

1. It is good with classification tasks but not as good with regression as it does not give precise continuous nature predictions.
2. Random forest model can be noisy in case of regression tasks and might overfit the data.

Few of the random forest applications are:

1. in the banking sectors, finding loyal and fraud customers.
2. In the stock market, random forest is used to identify the stock behavior as well as the expected profit or loss [28].
3. In computer vision, random forest is used for image classification [29].

The way the Random Forest model works is as follows:

1. If the number of cases in a set is N. Then a sample of these N cases are taken at random and with replacement.
2. If there are M variables, a number of m<M is specified such that at each node, m variables are selected at random out of the M. The best split on this m is used to split the node. The value of m is held constant while the forest grows.
3. Each tree grows to the largest extend possible without pruning.
4. Predicting the new data based on the majority of votes in case of classification and the mean of the results in case of regression.

The Random Forest is called an 'ensemble' model as they are a divide and conquer

approach that is used to improve performance. The main principle behind the ensemble method is that each tree classifier is rather a weak learner but when classifiers come together they are strong learners. In that ensemble method reduces invariance and improves performance.

CHAPTER 4

ANALYZING DATA AND FORECASTING FOR MEDICAL DIAGNOSIS

4.1 The Dataset in Hand

The data was obtained in Hungary of previously admitted patients to the hospitals. It provides historic details of diseases, diagnoses and similar medical records of patients admitted to the hospital including the hospitalization dates, sex smoking habits, the dates of their death and so on.

Our dataset includes the following attributes:

1. The patient's identifier which gives an identifying number to every patient.
2. Weather the patient was recorded as alive or dead at the time, and the value for this attribute is either 'yes' or 'no'.
3. The date and time of death, which could be a normal date including the hours of the patient's death. However, the dates of death could be unrecorded or unknown and they are identified as 'no data' in the dataset.
4. The gender of the patient, as being male or female.
5. The date of birth including the hours.
6. PCI having happened or not; where Percutaneous Coronary Intervention (PCI) is a treatment for a person experiencing myocardial ischemia (inadequate blood flow to the heart) or myocardial infarction (heart attack). The goal of PCI is to open up a coronary artery (blood vessel that brings blood and oxygen to the heart muscle) and restore blood flow [30].
7. The reanimation before the hospitalization, which can be 'yes', 'no' or 'no data' in the hospital records. The goal of the facial reanimation surgery is to create a

smile which is symmetrical, spontaneous and ideally re-innervated by the facial nerve [31].

8. Carcinogenic shock at hospitalization, which could be 'yes', 'no' or 'no data'. Carcinogenic shock (CS) is defined as a state of critical endorganic hypo perfusion due to reduced cardiac output. Notably, CS forms a spectrum that ranges from mild hypo perfusion to profound shock. Carcinogenic shock (CS) remains the most common cause of death in patients with acute myocardial infarction although mortality could be reduced [32].
9. The myocardial infarctus in medical history, which can also be 'yes', 'no' or 'no data'. The term myocardial infarction (MI) should be used when there is evidence of myocardial injury [33].
10. The health failure in medical history, which can take the value 'yes', 'no' or 'no data'.
11. The hypertonia in medical history or during the treatment, which can also be 'yes', 'no' or 'no data'. Hypertonia is a neurological dysfunction associated with a number of central nervous system disorders, including cerebral palsy, Parkinson’s disease, dystonia, and epilepsy [34].
12. The stroke in medical history, which takes the value of' yes', 'no' or 'no data'.
13. The diabetes in medical history and it can be 'yes', 'no' or 'no data'. Diabetes mellitus is a complex, chronic illness requiring continuous medical care with multi factorial risk reduction strategies beyond glycemic control. Ongoing patient self-management education and support are critical to preventing acute problems and reducing the risk of long-term complications [35].
14. The peripheral vascular disease (PVD) in medical history, which can be 'yes', 'no' or 'no data'. The peripheral vascular disease or (PVD) is a blood circulation disorder that causes the blood vessels outside the heart and brain to narrow, block, or spasm. This can happen in the arteries or the veins [36].
15. The hyperlipidemia in medical history, which can be 'yes', 'no' or 'no data'. Hyperlipidemia is a well-established cardiovascular risk factor, and some laboratory and epidemiological evidence suggests it is also a risk factor for cancer [37].
16. If the patient had been a 'smoker' or 'nonsmoker'. Still sometimes there could be no record of any smoking in the dataset.

17. The creatinine level, which signifies impaired kidney function or kidney disease. Creatinine is a chemical waste molecule that is generated from muscle metabolism. It is produced from creatine, a molecule of major importance for energy production in muscles. As the kidneys become impaired for any reason, the creatinine level in the blood will rise due to poor clearance of creatinine by the kidneys. Abnormally high levels of creatinine thus warn of possible malfunction or failure of the kidneys. In the dataset the creatinine levels are given either a numeric vale that indicates the level of the creatinine in the blood or a 'no data' value. As the data was prepared, the creatinine level was assigned three vales; 'normal' when the creatinine level was between 74.3 to 107 micro moles per liter, 'not normal' for all the other numeric values or 'no data' when there is no record of the creatinine level in the blood [38].
18. The diagnosis, which is either 'STEMI' or 'NSTEMI' [39]. STEMI and NSTEMI share the release of specific myocardial necrosis markers which define them clinically as acute MI. Ever since the redefinition of myocardial infarction (MI) in the year 2000, 1 those entities have entered the field: non-ST-elevation MI (NSTEMI) or STEMI [40].
19. The treatment identifier, that takes numeric values to list the treatment being done on the patient.
20. The data of the patient's hospitalization including the hours.

Table 4.1 shows 6 attributes for the first 18 patients in the dataset

Patient identifier	**Is the patient alive or not**	**Date of death**	**Sex**	**Date of birth**	**PCI is happened or not**
9695	Yes	No data	Female	21-06-60 0:00	Yes
9699	Yes	No data	Female	09-08-64 0:00	No
9700	No	18-02-14 0:00	Male	27-10-23 0:00	No
9701	No	15-01-15 0:00	Male	09-05-28 0:00	No
9795	Yes	No data	Male	14-12-53 0:00	Yes
9797	No	29-12-15 0:00	Male	14-08-58 0:00	Yes
9799	No	25-05-14 0:00	Male	06-01-43 0:00	Yes
9800	Yes	No data	Female	25-07-47 0:00	No

Table 4.1 (cont'd)

9801	Yes	No data	Female	26-04-65 0:00	No
9801	Yes	No data	Female	26-04-65 0:00	No
9472	Yes	No data	Male	25-06-55 0:00	Yes
102	Yes	No data	Male	28-05-41 0:00	Yes
9891	Yes	No data	Female	30-07-34 0:00	Yes
9891	Yes	No data	Female	30-07-34 0:00	Yes
9895	No	04-02-14 0:00	Female	22-03-39 0:00	Yes
9897	Yes	No data	Female	10-02-35 0:00	No
9898	Yes	No data	Male	04-09-45 0:00	Yes
9900	No	10-07-14 0:00	Female	28-09-26 0:00	No

It is possible to check some of the attributes in the original dataset.

Figure (4.1) shows the numbers of 'males' and 'females' in the dataset. Figure (4.1) shows that the given dataset has relatively more male patients than females. The number of the recorded male patients in the dataset is almost 600 whereas the number of females is less than 500. Summing up the numbers of the females and the number of males in the dataset adds up to 1099 which is the number of the samples in the dataset. Please note that this particular table has always a value of either 'male' or 'female' and there is no input as 'no data'.

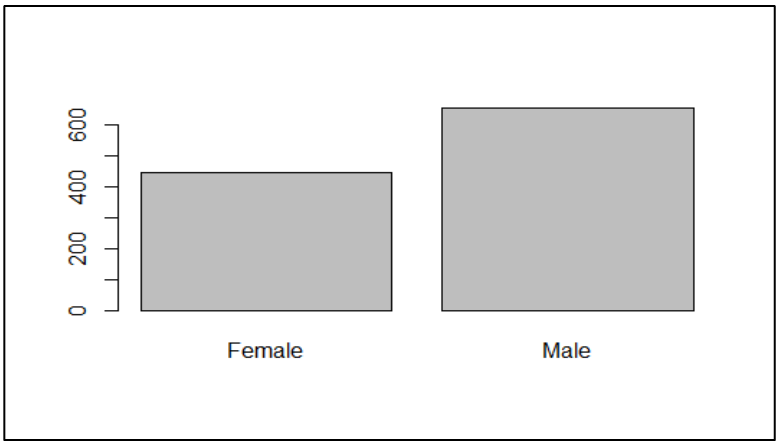


Figure 4.1 The sex of the patient in the given dataset

Figure (4.2) plots the stroke happening in the medical history of the patients in the dataset. The figure shows that many of the patients in the dataset had strokes before their hospitalization dates with a number higher than 600, whereas less than 200 of them did not have a stroke in their medical records before the date of their hospitalizations.

The number of the unknown data about the stroke in medical history is almost 300 patients in the given dataset.

Furthermore, Figure (4.3) shows the sex of the patients –male or female- against their health failure in their medical histories. Figure (4.3) shows the percentage of males

who did not have health failure before their hospitalization dates is about 20 percent of

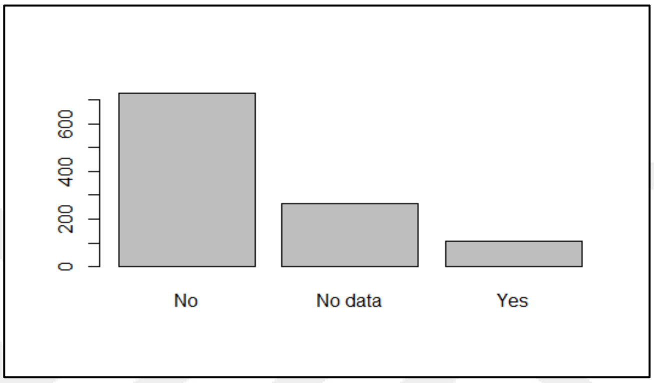


Figure 4.2 Stroke in medical history

All the males in the dataset and the percent of the males who had failure before their hospitalization dates is less than 10 percent of all the males, whereas the percentage of unknown data regarding the health failure in males is more than 60 percent.

For the health failure in the females in the dataset, Figure (4.3) shows that also less than 20 percent of the females showed no health failure in their medical records, less than 10 percent showed health failure, yet more than 60 percent of the female data about the health failure in females is still unknown.

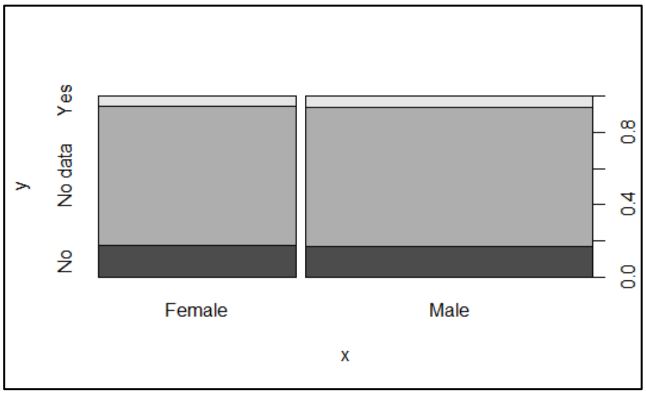


Figure 4.3 The Sex (male/female) vs. Health Failure in Medical History

4.2 Preparing the Data

A dataset of 1099 variables and 20 attributes was to be prepared and modeled to help better treat admitted patients.

As we are interested in predicting whether newly admitted patients are more likely to survive after 30 days of their hospitalization. The more accurate the results are, the higher the likelihood of new patients' situations to be conceived.

We had to refer back to the previous records of patients in the hospital.

In order to determine the 30 days period of the previous patients, the difference -in days- between the dates of every patient's admission and their dates of death in those records was calculated. Afterwards a new column was added to the dataset to include the calculated difference.

Before jumping into modeling, data preparation process carried out as follows:

1. Eliminating the unnecessary columns: all columns in the data set that were unnecessary for the final predictions were eliminated from the final data set. The eliminated columns - attributes- are the patient's identifier, the data of birth, the treatment identifier, the date of hospitalization and the date of death.
2. Digitizing the data: the new data frame was digitized by assigning the numbers 0,1 and 2 to the values yes, no and no data in the data set, and in the case of the creatinine level the 0, 1 and 2 were assigned to the values of ‘not-normal’, ‘normal’ and ‘no data’ respectively.
3. Scaling the data: the data was scaled so that it took values between 0 and 1 and in that equal emphasis was given to each attribute in the dataset.
4. Partitioning the dataset: The partitioning consists of two types :

   First it was train-test partitioning with 30% given to the test data and 70% to the training data from the whole dataset. In that the samples -variables- are about 770 in the trainset and about 330 in the test set with all the attributes equally included in both sets.

   Second it was the 10 fold partitioning for cross validation where at first only the training set, which is 70% of the whole dataset, was divided into 10 folds. Then the whole dataset was divided into 10 folds to add more validity to the achieved results. The mentioned division of the data allows to get different misclassifications, which validates the results and secures them and then the average misclassification of all the 10 calculated misclassifications was calculated and used.

4.3 Preparing the Data

This section presents the commonly used methods for modeling including the Artificial

Neural Networks, The Random Forest, the Support Vector Machine and the Decision Trees models and it presents how those models have been applied to the given dataset.

4.3.1 Artificial Neural Networks

As explained in section (3.1) a sample ANN architecture for a supervised learning multi-layer perceptron is composed of:

1. A layer of input elements also called the input vector, representing independent variables.
2. Optionally, one or more hidden processing layers.
3. Weighted connections between nodes in adjacent layers
4. An output layer of one or more elements, representing the dependent variable(s).

Tuning the Neural Nets can be conducted based on the variable rep or – the repetitive variable- which is the number of times the model would be run, it might be that the second or third time of running the model gives better results, and so neural nets model’s results could be improved by adding the repetitive variable. However, the neural nets model may not converge to many of the repetitive values.

The ANNs nodels that were used are

In one hidden layer Neural Nets models: 1 ,3 and 5 neurons were tried.

In two hidden layers Neural Nets model: 2-1 neuron was tried.

It was found that the 5neurons-1layer Neural Nets had the best results in terms of getting the lowest misclassification in the prediction when applied to the training set and to the test set as well.

Further work was conducted by tuning the 5neurons-1layer Neural Nets with 4 and 5 repetitions in order to help with the converge of the model. Table (5.1) in the Results section shows all the results of applying the Neural Net models and Figure (4.4) shows the 5-neurons 1-layer neural net model. Figure (4.4) is not so clear since the figure was saved directly from a dense plot in RStudio.

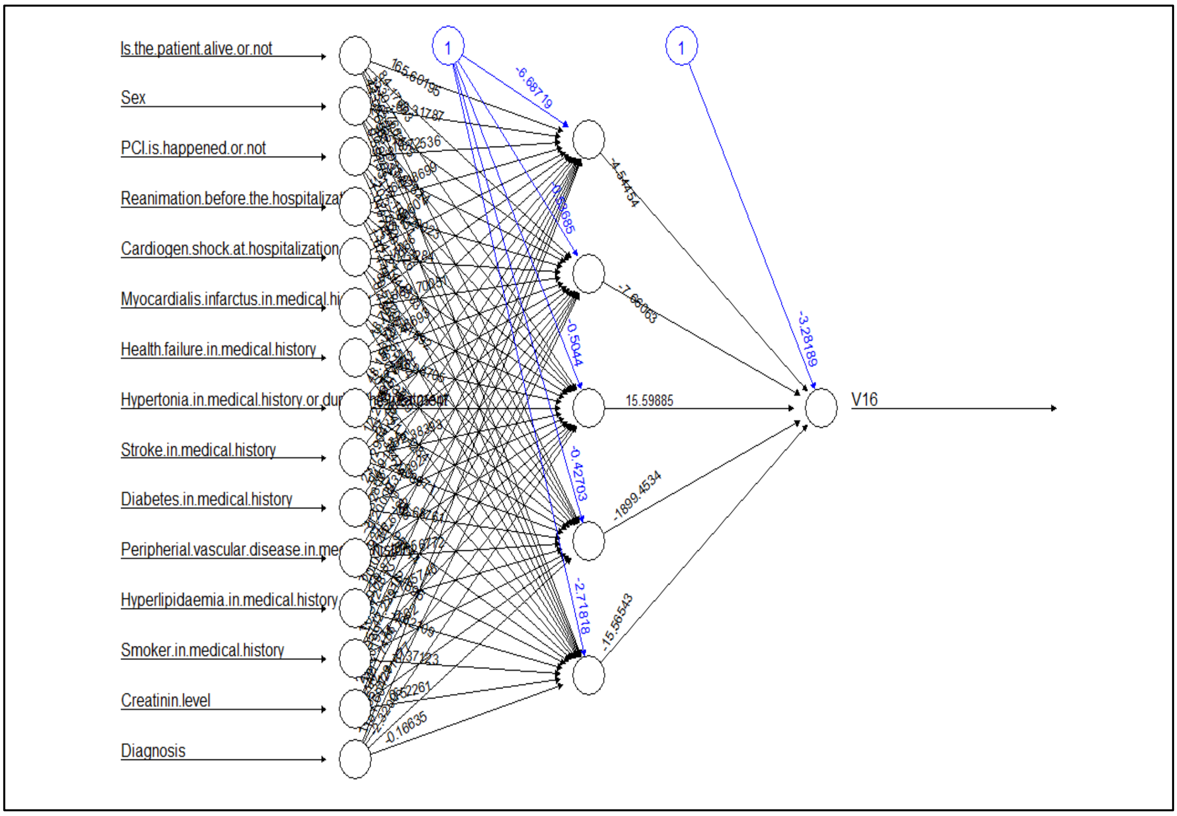

Is.the.patient.alive.or.not
Sex
PCI.is.happened.or.not
Reanimation.before.the.hospitaliza
Cardiogen.shock.at.hospitalization
Myocardialis.infarctus.in.medical.hi
Health.failure.in.medical.history
Hypertonia.in.medical.history.or.dur
Stroke.in.medical.history
Diabetes.in.medical.history
Peripherial.vascular.disease.in.me
Hyperlipidaemia.in.medical.history
Smoker.in.medical.history
Creatinin.level
Diagnosis
1
1
165.60195
-6.68719
-0.59685
-0.5044
-0.42703
-2.71818
-3.28189
-4.54454
-7.68063
15.59885
-1899.4534
-15.56543
-0.16635
V16

Figure 4.4 Artificial Neural Network Model with 5-neurons 1-layer applied on the hospital records dataset

Figure (4.5) shows the ANNs model with two hidden layers 2 and 1 neuron in each respectively. Figure (4.5) is not so clear since the figure was saved directly from a dense plot in RStudio.

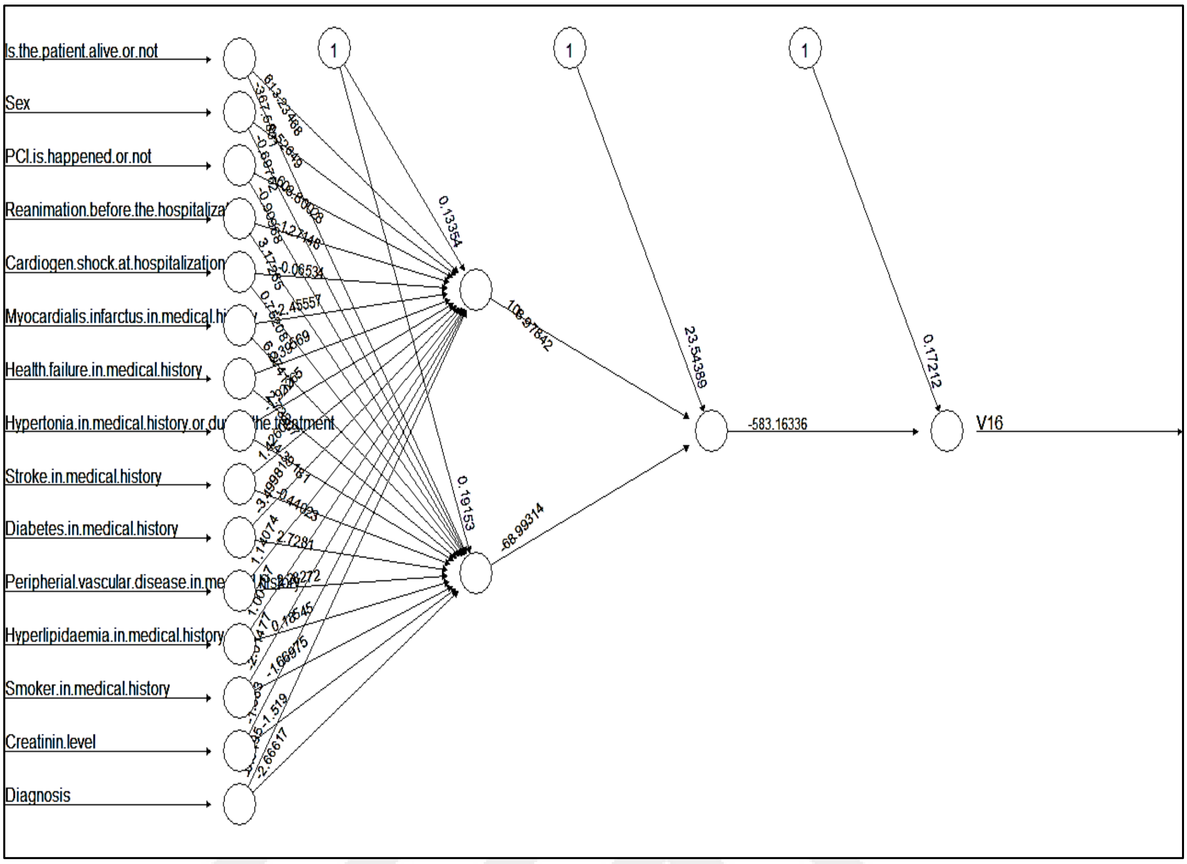

Is.the.patient.alive.or.not
Sex
PCI.is.happened.or.not
Reanimation.before.the.hospitaliza
Cardiogen.shock.at.hospitalization
Myocardialis.infarctus.in.medical.hi
Health.failure.in.medical.history
Hypertonia.in.medical.history.or.du the.treatment
Stroke.in.medical.history
Diabetes.in.medical.history
Peripherial.vascular.disease.in.me history
Hyperlipidaemia.in.medical.history
Smoker.in.medical.history
Creatinin.level
Diagnosis
1
1
1
613.23468
-367.6
-0.69762
-0.90968
3.17265
0.75208
6.9
52649
608.80028
1.27148
-0.06534
2.45557
0.39569
1.42
-3.49981
1.14074
1.00
-2.01477
-1.
0.13354
0.19153
0.44023
2.7281
0.28272
0.18545
-1.66975
-1.519
-2.66617
108.97842
-68.99314
23.54389
0.17212
-583.16336
V16

Figure 4.5 Artificial Neural Network Model with 2,1-neurons 2-layers applied on the hospital records dataset

4.3.2 Random Forest

A Random Forest is an ensemble of tree-structured classifiers. Every tree of the forest gives a unit vote, assigning each input to the most probable class label. One of the major advantages of using the Random Forest is that it does not suffer from over fitting, even if more trees are appended to the forest.

Turning the random forest models depend on two values: ntree – the number of trees in the forest- and mtry, which is the number of variables per level.

The random forest modeling was tried on our dataset, and by tuning the random forest model it was found that by assigning mtry to be 3 and the number of threes to 250 trees we got a good tuned model in terms of accuracy and time consumption.

So having more than 250 trees in the forest would give only slight improvements with more time consumption to the results and number of tries was better at 3 and 6 with less error rate at 3.

Figure (4.6) shows the results where mtry = 3 was the best choice for the tree's depth. Figure (4.7) shows the error against the depth of the tree. Table (5.1) in the Results section shows the results of applying the Random Forest.

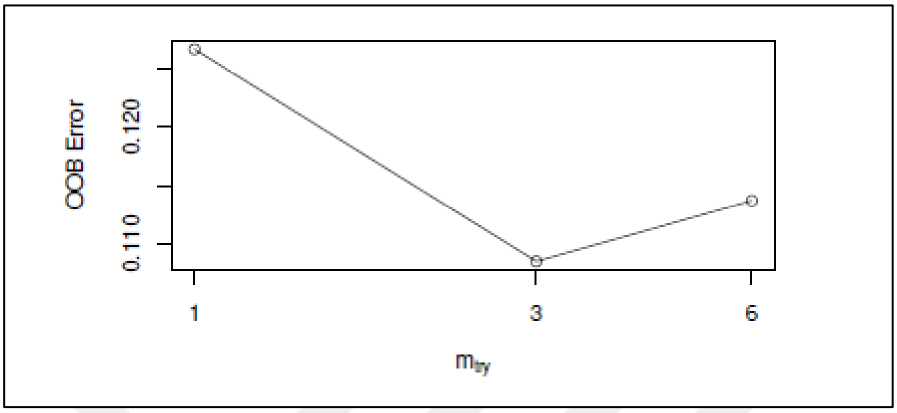


Figure 4.6 RF 230 / 3 showing that we get less error at mtry = 3

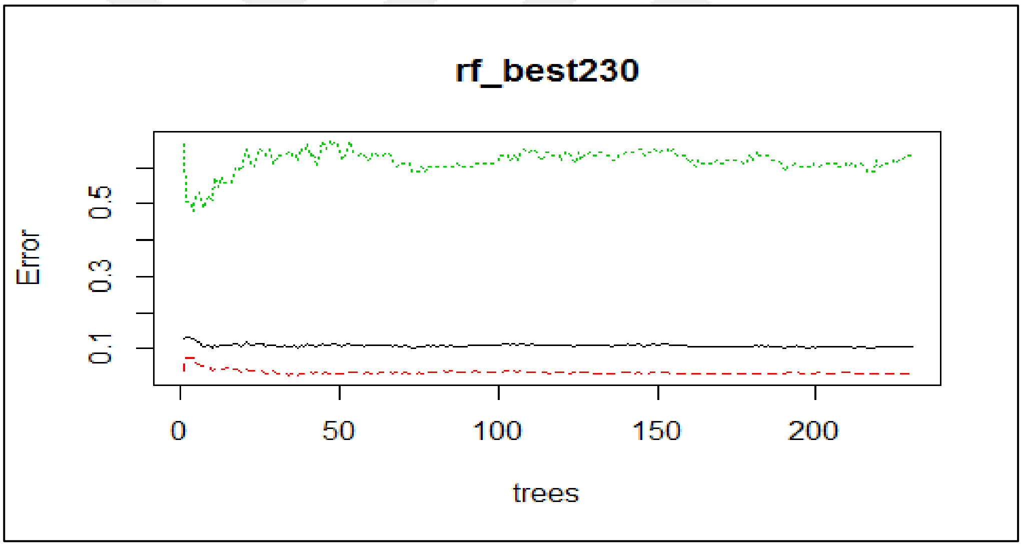


Figure 4.7 RF 230 / 3 showing the error vs the depth of the tree

4.3.3 Support Vector Machine

In general and based on the experience we can indicate that the SVM model works well for data which does not have many attributes. If there are thousands of attributes, there may be a need to choose a subset of them before giving the data to SVM.

In tuning the SVM models, we can use the epsilon and the cost value in such a way that they are not so high as to overfit vectors and not too low either.

In the given dataset, SVM gave good results especially after tuning and getting the best model. As the epsilon was tuned with values between 0 and 1 with an increment of 0.1 and the cost was tuned between 4 and 2 to the power of seven so that we use sufficient number of combinations. By adjusting all the tuning, the SVM model provided more accurate predictions. Table (5.1) in the Results section shows the results of applying the SVM models.

4.3.4 Decision Trees

In the supervised classification, large training data is very common, and decision trees are widely used in that case. However, as some bottlenecks such as memory restrictions, time complexity or data complexity, many supervised classifiers including classical tree cannot directly handle big data.

For our relatively small dataset the decision tree model was also tried and some good results were obtained. Table (1.5) in the Results section shows the results of applying the decision tree model.

Figure (4.8) shows part of the decision tree starting from node 15 all the way to the final nodes - fitting all the nodes in one figure would make it difficult to read. We can notice that the splitting criteria was based on impurity, and in that the 'Sex' attribute came later - node 36 - and 'smoker in medical history' came 25th. All the final nodes are presented in the Figure (4.8).

Figure (4.8) is not so clear since the figure was saved directly from a dense plot in RStudio.

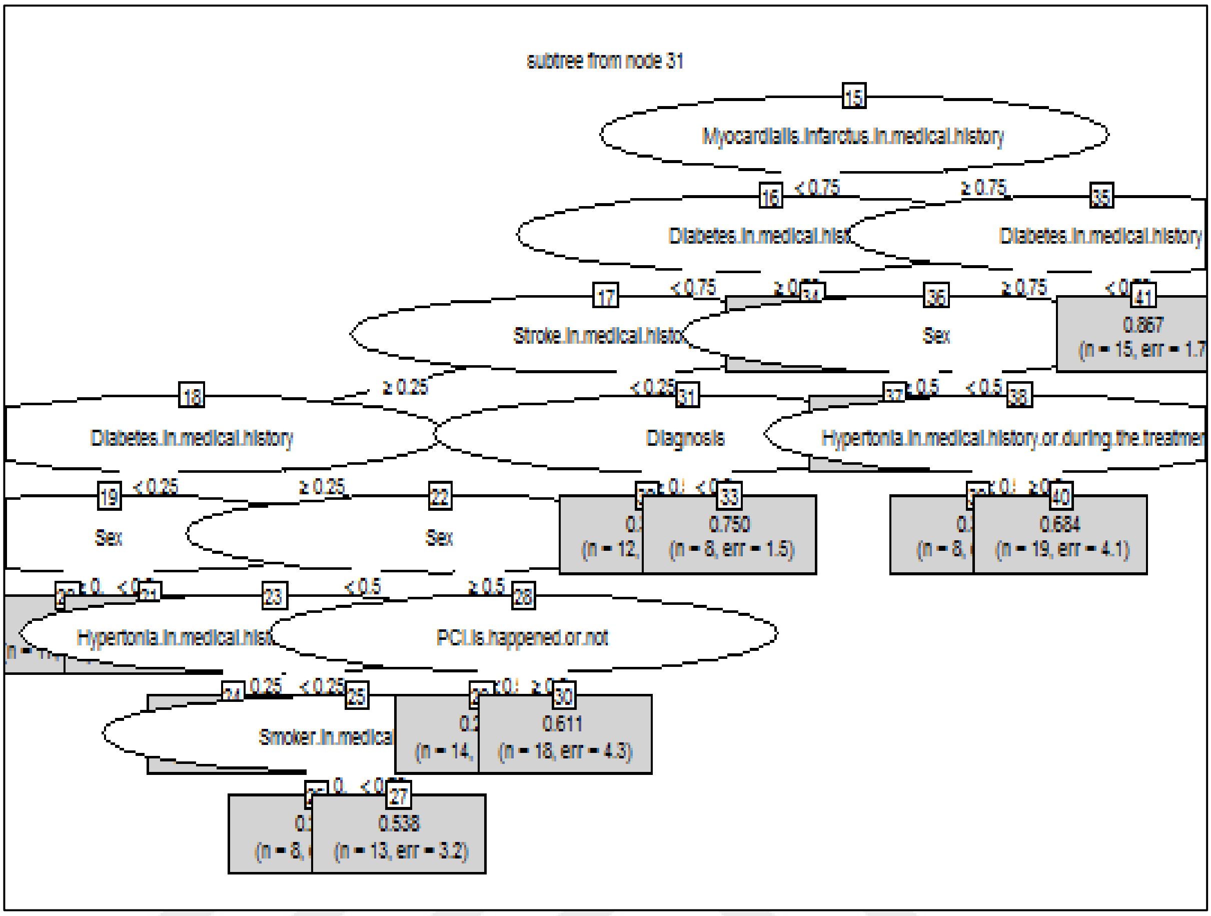


Figure 4.8 Decision Trees model on the train set showing nodes starting from node 15

CHAPTER 5

RESULTS AND DISCUSSIONS

This chapter presents two types of results obtained from applying the aforementioned models on the given dataset.

1. The Misclassification results: they show how accurate or misclassified the obtained results are, it shows the misclassifications resulted from using each of the aforementioned methodologies: ANNs, RF, SVM and Decision Trees.
2. The performance results: after getting more satisfying outcomes from the misclassifications results, the performance results show how effective the performance of the Support Vector Machine model based on the ROC and the AUC criteria.

1. The Misclassification Results

To calculate the misclassifications of any model first find the confusion matrix should be found. The confusion matrix for a binary classifier contains of four identifying numbers. One of them indicates the number of true predictions that the patient had died, the second number in the confusion matrix indicates the right predictions that the patient had not died, the third number in the confusion matrix is the mis-prediction that the patient had died and the forth number is the mis-prediction that the patient had not died.

The hits -right predictions- are divided by all the predictions -hits and misses. The number of hits or the right predictions that the model gave for the death of the patient and the number of the hits or the right predictions the patient died are added up and then divided by the same number plus the number of misses in the survival and the number of misses in the death.

Finally by subtracting one from the accuracy we get the misclassification of the whole model.

Table (5.1) shows the obtained results from all the four modeling types done to the dataset of 1099 samples -variables-:

The table shows the misclassifications.

The accuracy can always be calculated for each model as:

Accuracy = 1 – misclassification

Table 5.1 Misclassifications on training and test sets

THE MODEL	**THE TUNED MODEL**	**ON TRAIN SET**	**ON DATA SET**
ANNs	1-neuron 1-layer	0.382428	0.058461
	3-neuron 1-layer	0.065891	0.058461
	5-neuron 1-layer	0.045219	0.015384
	2-1 neuron 2-layer	0.108527	0.04
RF	Non-tuned	0.034883	0.027692
	3 / 200	0.036175	0.08923
	3 / 230	0.042635	0.095384
	3 / 500	0.042635	0.095384
SVM	Non tuned	0.086563	0.043076
	Tuned	0.031007	0.015384
Decision Trees	Non-tuned	0.126614	0.12

Next the cross validation methods were used to better assess the models. Table (5.2) shows the results in terms of misclassifications when applying the following models:

Random Forest model with 200 tress and mtry=3, Random Forest model with 230 trees and mtry=3 and the tuned SVM model using 10-fold Cross Validation on the training set and on the whole set.

Notice that in Table (5.1) and Table (5.2) the misclassification results are the average of the 10 misclassification resulting from each of the 10 folds.

In Table (5.2) the average of the resulted misclassifications was considered.

As we are having 10 misclassifications resulting from 10 Cross Validation folds, the average of those was put in the Table (5.2) to be considered as the final results of the 10-fold Cross Validation.

Again the results in table (5.2) prove that SVM is relatively a better model for the given data set with average misclassification being around 0.039045 on the training set and average misclassification being 0.036396 on the whole set.

Table 5.2 Misclassifications with CV-10 Folds on the train and whole sets

THE MODEL	**THE TUNED MODEL**	**CV ON TRAIN SET**	**CV ON TEST SET**
RF	3 / 200 3 / 230	0.03876 0.037322	0.039024 0.038823
SVM	TUNED	0.039045	0.036396

As you can tell from the results above, the best accuracy in general was achieved with the SVM model -after tuning the epsilon and the cost of the model-. The misclassifications are as follows:

On the train and test sets: 0.028423 and 0.024615 respectively.

On the 10 fold cross validation on the train set : the mean of the misclassification 0.039045

On the 10 fold cross validation on the whole set: the mean of the misclassification 0.036396

2. The Performance Results

One way to measure the performance of a model is to draw the Receiver Operator Curve ROC.

The ROC is a graphical block to evaluate a performance of a binary classifier and ROCs can be used to compare models and compare classifiers.

The ROC contains two important variables - the sensitivity and the specificity.

The sensitivity is the probability that the test is positive -the patient died within 30 days of hospitalization- given that the patient had actually died within the 30days. Thus correct predictions of the model. The specificity is the probability that the test is negative -the patient did not die within 30 days of hospitalization- given that the patient had not actually died. Thus correct predictions of the model. For accurate predictions of the model the sensitivity and the specificity are needed to be high.

Now the Receiver Operation Curve is a curve where the x-axis is (1-specificity) and the y-axis is (sensitivity). Ideally the curve should climb quickly toward the top-left meaning the model correctly predicted the cases. And the bigger the area under this curve AUC the better the model's performance is. In other words we need the AUC to approach one for a better performance of the model [41] [42].

In our models and in terms of the tuned SVM performance; The ROC curve was drawn and the area under that curse AUC was calculated for the SVM model in three parts. See Figures (5.1) Figure (5.2) and Figure (5.3).

First the curve was drawn for the SVM model on the training set as in Figure (5.1). Second the ROC was drawn for the SVM model on the test set as in Figure (5.2) and last the curve was drawn for the SVM model on the whole dataset as in Figure (5.3).

The following results were reached:

AUC1 = 0.889823 on the train set

AUC2 = 0.992399 on the test set

AUC3 = 0.945945 on the whole set

The area under the curve in the SVM modeling is quite close to 1 thus the model's performance is quite good.

Figure (5.1) shows how close the ROC curve from covering up the whole area and becoming 1, which indicates great performance on the train, set in the given dataset when Support Vector Machine model was applied.

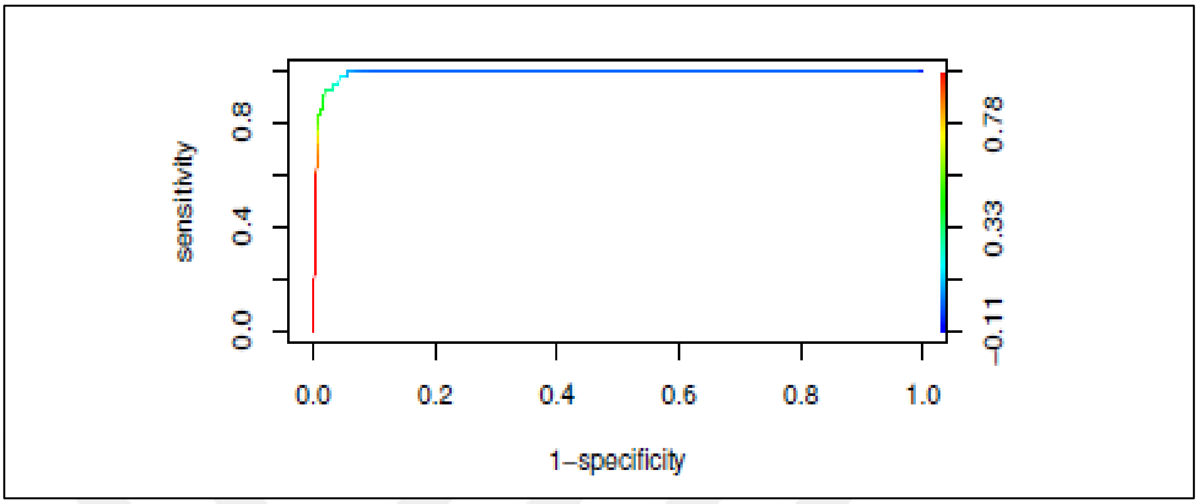


Figure 5.1 Performance - Area Under Curve for the SVM model on the training set, (1-specificity) vs. sensitivity

Figure (5.2) again shows how close the ROC curve from covering up the whole area and becoming 1, which indicates great performance on the test set in the given dataset when Support Vector Machine model was applied.

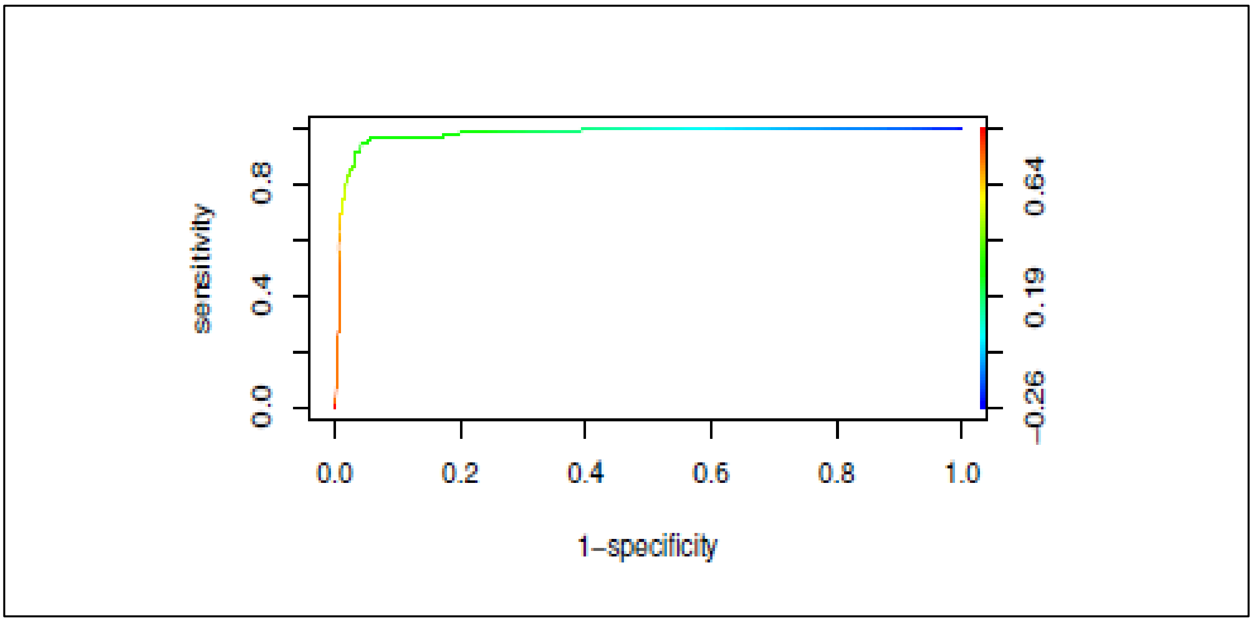


Figure 5.2 Performance - Area Under Curve for the SVM model on the test set, (1-specificity) vs. sensitivity

Figure (5.3) shows how close the ROC curve from covering up the whole area, which indicates great performance on the whole given dataset when Support Vector Machine model was applied.

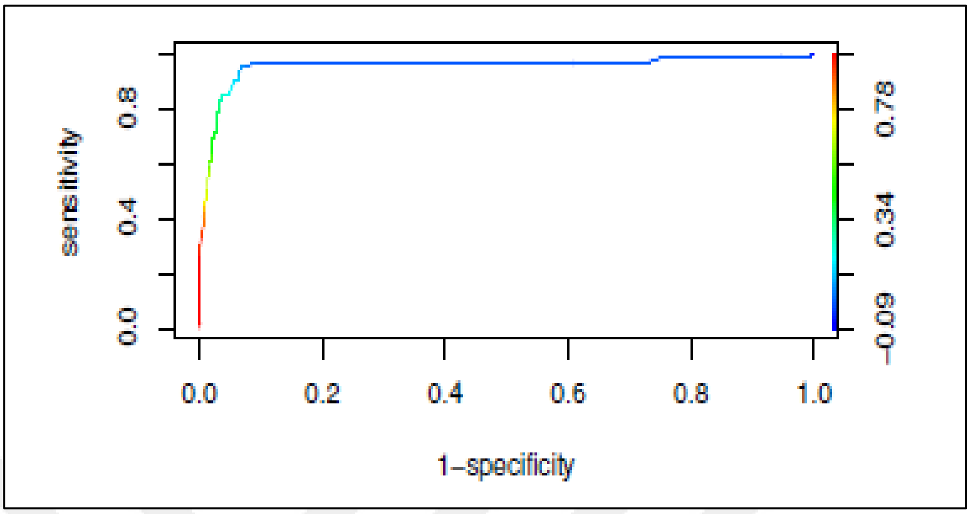


Figure 5.3 Performance - Area Under Curve for the SVM model on the whole dataset, (1-specificity) vs. sensitivity

As the performance of the SVM was relatively good, we can conclude that on the given dataset using the SVM model would give better predictions than using any of the other three models with the tuning applied.

CHAPTER 6

CONCLUSIONS

This work can add to the fact that on relatively small datasets the SVM model is quite convenient to try and to use mainly for classification problems. Yet the Random forest which can be applied not only to a classification problem but to a regression one can also result in good accuracy and less time consumption, unlike the use of the Artificial Neural Nets which in our case took relatively longer time to converge and somehow less accurate results. Furthermore, the usage of one decision tree model gave quite bad results on the trained dataset.

This thesis work shows that in terms of accuracy, the tuning of the SVM or the RF models slightly improved the results but not quite much. However, the tuning of the models still consumed relatively more time, for that further work is still required to see about whether tuning those model would be sufficient to the problem in hand.

In order for the modeling to be more practical for health care systems, especially with an increasing number of data in every dataset, three improvements can be suggested for the analysis:

1. Developing the data: A similar study could be done on a bigger dataset with more variables and attributes. Even merging correlated attributes or dropping some of them as viable could make the modeling less time consuming and more efficient. On the other hand, gathering more details about the patients in the hospital records would guarantee to give more accurate prediction for the newly admitted patients. An example would be having more details about the patient’s frequent smoking habit, not just weather they had smoked or not, but also how

frequent they had smoked; assigning words such as ‘rarely smoked’ , ‘occasionally smoked’ , ‘smoked sometimes’ or ‘smoked always’.

2. Developing the preparation of the data: while the aforementioned study did not take into consideration the ‘date of the birth’ while doing the data preparation, further study can be done that includes the age of the patient. Adding such attributes to the preparation process and later to the modeling could result in more accurate predictions.
3. Much as the results on the given dataset are quite satisfying when using the tuned SVM model, more trials can be done and tested. An example would be, trying various types of neural net modeling, with 7 neurons in one hidden layer or 9 neurons, even trying a mix such as 3-2 neurons in two hidden layers and so on. For the Random Forest model, trying more than 500 trees and probably mtry to be 6 could give good results.

REFERENCES

[1] Barrados, M. and Mitchell, J. I., (2017). “Getting started with big data: The promises and challenges of evaluating healthcare quality,” Cyber Society, Big Data, and Evaluation: Comparative Policy Evaluation Journal, 24(1):179-192.

[2] Wang, Y., Kung, L. and Byrd, T. A., (2018).“Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations,” Technological Forecasting and Social Change, 126(7): 3–13.

[3] Chaurasia, V., (2017). “Early prediction of heart diseases using data mining techniques”, Carib.j.SciTech Journal, 1(8): 208-217.

[4] Bronzino, J. and Peterson, D., (2015). The Biomedical Engineering Handbook, 4th ed. Boca Raton, Florida, USA: CRC Press.

[5] Weng, Y., Wu, C., Jiang, Q., Guo, W. and Wang, C., (2016). “Application of support vector machines in medical data,” in Cloud Computing and Intelligence Systems (CCIS), 2016 4th International Conference 200–204 August 2016.

[6] Lortie, C., J., (2017). “A review of R for Data Science: key elements and a critical analysis.” PeerJ Preprints, 1(5):287-290.

[7] Solanki, H., Patel, Y., Vaghela, D., and Vaidya, M., (2017). R: An Open Source Software Environment for Statistical Analysis. INFLIBNET Centre, Gandhinagar.

[8] Horton, N., J., and Kleinman, K., (2015). Using R and RStudio for data management, statistical analysis, and graphics. CRC Press, New York.

[9] Horton, N., J., and Kleinman, K.,(2016). Graphical Data Analysis with R. Wiley Online Library, Cambridge.

[10] Victor A. B., (2016). Using R for Numerical Analysis in Science and Engineering. CRC Press, New York.

[11] De Levie, R., (2004). Advanced Excel for scientific data analysis. Oxford University Press, USA.

[12] Meyer, D., Z., and Avery, L., M., (2009). Excel as a qualitative data analysis tool. Sage Publications Sage CA: Los Angeles, CA.

[13] Assaad, H., I, Hou, Y., Zhou, L., Carroll, R., J., and Wu, G., (2015). Rapid publication-ready MS-Word tables for two-way ANOVA. Springer, SpringerPlus, China.

[14] Stormy A., (2017). MATLAB (Fourth Edition) Butterworth Heinemann, London.

[15] Harry, Z., (2017). MATLAB function for 3-D and 4-D topographical visualization in geosciences, ELSEVIER, New Delhi.

[16] João M.P.C., João M., F., Miguel P., M., Tiago, C., and Ricardo, N., (2012). “Enriching MATLAB with aspect-oriented features for developing embedded systems”. Journal of Systems Architecture. 59(7):130-146.

[17] Bisták, P., Halás, M., and Huba, M., (2017). Modern Control Systems via Virtual and Remote Laboratory Based on Matlab, ELSEVIER, Berlin.

[18] Shiraiw, S., Greenwald, M. , Stillerman, J.A., Wallace, G.M., London, M.R., and Thomas, J. , (2018). Tools to export published datasets together with metadata from IDL/Python/MATLAB and Scope, ELSEVIER, London.

[19] Stephen, K., Philipp, M., Jim, Y., and Julien J.,H., (2018). A python framework for multi-agent simulation of networked resource systems Environmental Modelling Software, New York.

[20] Oral, M., Oral, E., and Aydın, A., (2012). Supervised vs. unsupervised learning for construction crew productivity prediction, Automation in Construction, Congress, 2-14 August 2012, Washington.

[21] Amato, F. , Lopez, A. , Pe˜na-M´endez, E. M., Vaˇnhara, P. , Hampl, A. and Havel, J., (2013). Artificial neural networks in medical diagnosis., Texas.

[22] Walczak, S. ,(2018). Artificial neural networks. In Encyclopedia of Information Science and Technology, Fourth Edition. IGI Global, New York.

[23] Hsu, C.-W., Chang, C.-C. , Lin, C.-J., et al., (2003). A practical guide to support vector classification., Washington.

[24] Chaudhary, A. , Kolhe, S. and Kamal, R. , (2016). “An improved random forest classifier for multi-class classification,” Information Processing in Agriculture, 3(4): 215–222.

[25] Gharehchopogh, F. S. , Mohammadi, P. and Hakimi, P. , (2012). “Application of decision tree algorithm for data mining in healthcare operations: A case study,” International Journal of Computer Applications, vol. 52, no. 6, 2012.

[26] Podgorelec, V. , Kokol, P., Stiglic, B. and Rozman, I., (2002). “Decision trees: an overview and their use in medicine,” Journal of medical systems, 26(5): 445–463.

[27] D´ıaz-Uriarte R. and De Andres, S. A., (2006). “Gene selection and classificationof microarray data using random forest,” BMC bioinformatics Journal, 7(1):3-8.

[28] Xiao, M., Yan, H. , Song, J. , Yang, Y. and Yang, X. , (2013). “Sleep stages classification based on heart rate variability and random forest,” Biomedical Signal Processing and Control, 8(6):624–633.

[29] Goodfellow, I., Bengio, Y. and Courville, A., Deep Learning, http://www.deeplearningbook.org. 18 February 2016.

[30] Torpy, J. M., Lynm, C. and Glass, R. M. , (2004). “Percutaneous coronary intervention,” JAMA, 291(6):778–778.

[31] Andrews, P., Randhawa, P. S., Joseph, J., Goh, S., Li, Q., Poirrier, A.-L., Leong, S., Lesser, T. and Saeed, S. R., (2016). “A prospective 4-year study of the objective and subjective outcomes of fifteen patients after dynamic facial reanimation surgery,” Clinical Otolaryngology, 41(6):825–829.

[32] Thiele, H., Ohman, E. M., Desch, S., Eitel, I. and de Waha, S., (2015). “Management of cardiogenic shock,” European heart journal, 36(20):1223–1230.

[33] Ibanez, B., James, S., Agewall, S., Antunes, M. J., Bucciarelli-Ducci, C.,Bueno, H., Caforio, A. L., Crea, F., Goudevenos, J. A., Halvorsen, S., et al.,(2017). “2017 esc guidelines for the management of acute myocardial infarction in patients presenting with st-segment elevation: The task force for the management of acute myocardial infarction in patients presenting with st-segment elevation of the european society of cardiology (esc),” European heart journal, 39(2):119–177.

[34] Lee, C. A., Chin, L.-S., and Li, L., (2017). Hypertonia-linked protein trak1 functions with mitofusins to promote mitochondrial tethering and fusion., Berlin.

[35] A.D. Association et al., (2014).“Standards of medical care in diabetes—2014,” Journal of Diabetes Care, 37(1):14–8.

[36] Ulbricht, C., (2012). “Peripheral vascular disease: An integrative approach: A natural standard monograph,” Alternative and Complementary Therapies, 18(1):44–50.

[37] Carter, P., Uppal, H., Chandran, S., Bainey, K. and Potluri, R., (2017). “3106 patients with a diagnosis of hyperlipidaemia have a reduced risk of developing breast cancer and lower mortality rates: a large retrospective longitudinal cohort study from the uk acalm registry,” European Heart Journal, 38(1):133-139.

[38] Hall, J. E., (2011). Guyton and hall textbook of medical physiology., Philadelphia.

[39] Bode, C. and Zirlik, A., (2007). Stemi and nstemi: the dangerous brothers., Philadelphia.

[40] Sim, D. S., Kim, J. H. and Jeong, M. H., (2009). “Differences in clinical outcomes between patients with st-elevation versus non-st-elevation acute myocardial infarction in korea,” Korean Circulation Journal, 39(8):297–303.

[41] Hanley, J. A. and McNeil, B. J., (1982). “The meaning and use of the area under a receiver operating characteristic (roc) curve.” Radiology Journal, 143(1):29–36.

[42] Sahiner, B., Chen, W., Pezeshk, A. and Petrick N., (2017). “Comparison of two classifiers when the data sets are imbalanced: the power of the area under the

precision-recall curve as the figure of merit versus the area under the roc curve”, International Society for Optics and Photonics, 104(2):101-116.

CURRICULUM VITAE

PERSONAL INFORMATION

Name Surname	: Kinan MORANI
Date of birth and place	: 01.03.1989 – Hama, Syria
Foreign Languages	: English, Arabic, Turkish
E-mail	: kinan.morani@hotmail.com

EDUCATION

Degree	**Department**	**University**	**Date of Graduation**
Undergraduate	Electrical Engineering	Albaath University	2014
High School	General Science	Fawaz Hasek High School	2008

PUBLISHMENTS

1. Morani, K. Engin, Ş. N., Eigner, G., Ferenci, T., (2018).“Data Analysis and Forecasting in Hospital Related Application”, 2nd student international congree, Izmir-Turkey, May 2018.

2. Morani, K., Eigner, G., Ferenci, T., Kovacs, L., Engin, Ş. N., (2018). “Prediction of the Survival of Ptients with Cardiac Failure by using Soft Computing Techniques”, SACI conference, Romania, May 2018.