# DATA MINING METHODS FOR DIABETES PREDICTION

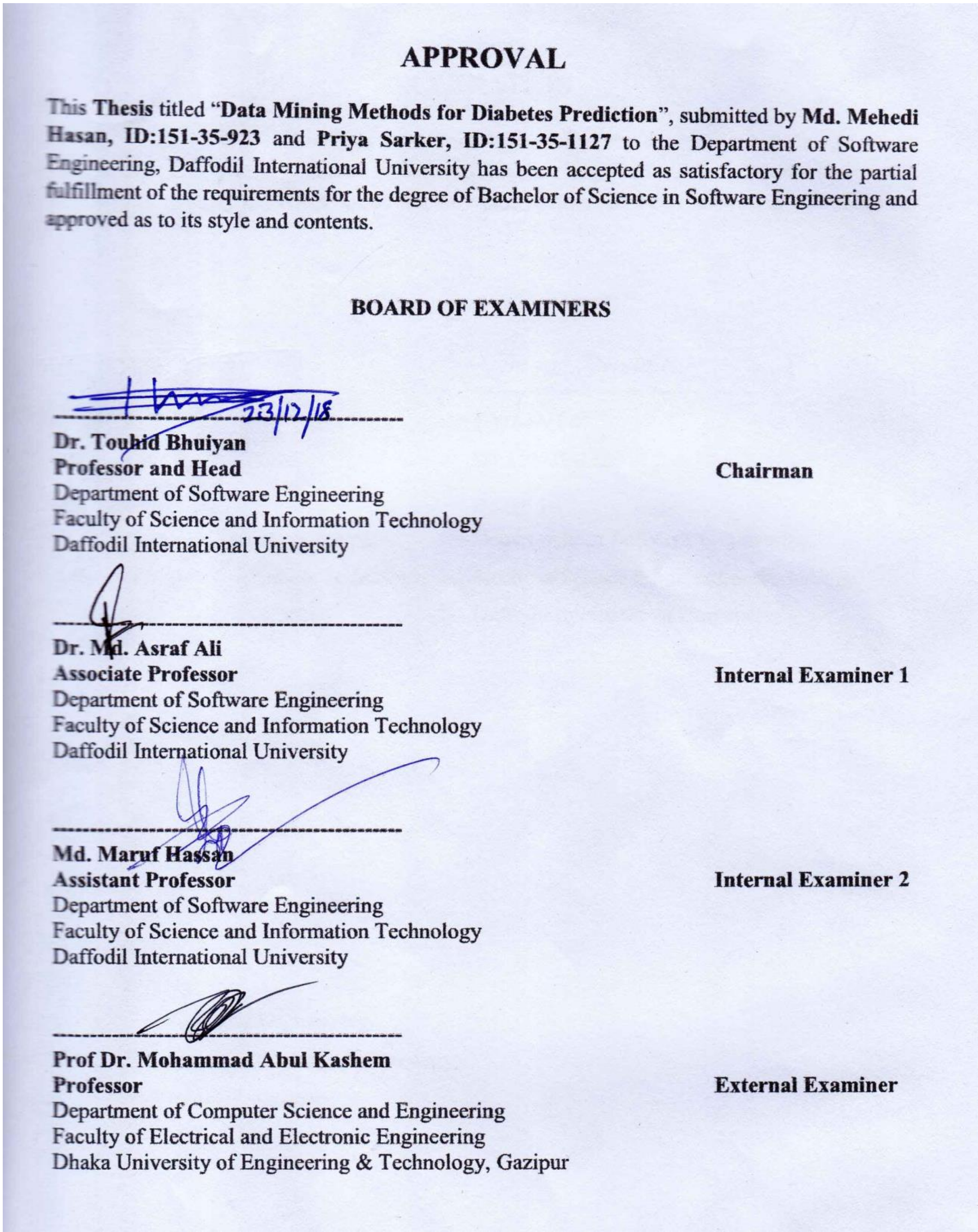
By

**Md. Mehedi Hasan (151-35-923)**
**Priya Sarker (151-35-1127)**

A thesis submitted in partial fulfillment of the requirement for the degree of Bachelor of Science in Software Engineering

**Department of Software Engineering**
**DAFFODIL INTERNATIONAL UNIVERSITY**

Fall – 2018



# APPROVAL

This **Thesis** titled "**Data Mining Methods for Diabetes Prediction**", submitted by **Md. Mehedi Hasan, ID:151-35-923** and **Priya Sarker, ID:151-35-1127** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approved as to its style and contents.

## BOARD OF EXAMINERS

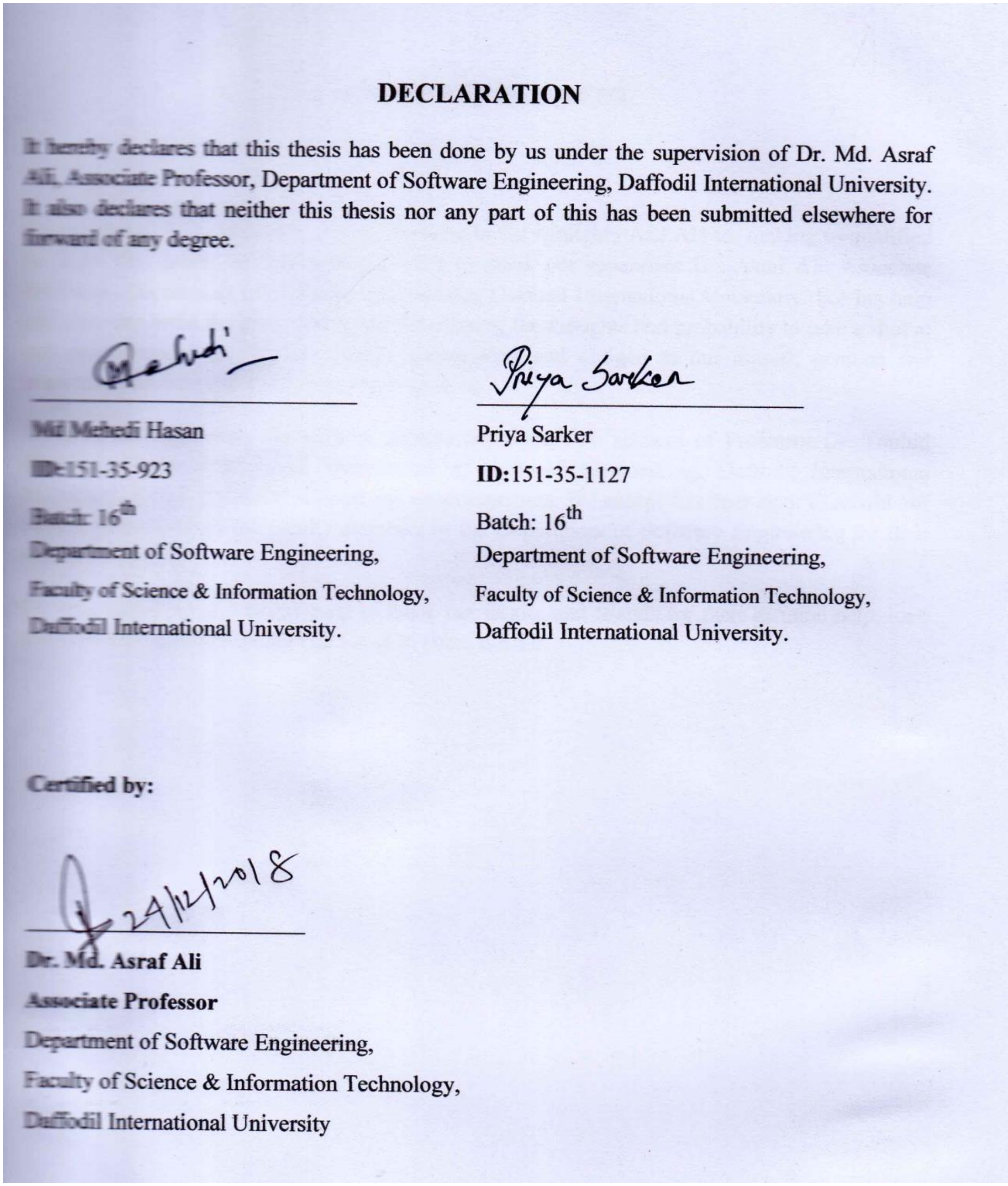
23/12/18

**Dr. Touhid Bhuiyan**
**Professor and Head** — **Chairman**
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

**Dr. Md. Asraf Ali**
**Associate Professor** — **Internal Examiner 1**
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

**Md. Maruf Hassan**
**Assistant Professor** — **Internal Examiner 2**
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

**Prof Dr. Mohammad Abul Kashem**
**Professor** — **External Examiner**
Department of Computer Science and Engineering
Faculty of Electrical and Electronic Engineering
Dhaka University of Engineering & Technology, Gazipur



# DECLARATION

It hereby declares that this thesis has been done by us under the supervision of Dr. Md. Asraf Ali, Associate Professor, Department of Software Engineering, Daffodil International University. It also declares that neither this thesis nor any part of this has been submitted elsewhere for forward of any degree.

Priya Sarker

| | |
|---|---|
| Md Mehedi Hasan | Priya Sarker |
| ID:151-35-923 | **ID:**151-35-1127 |
| Batch: 16th | Batch: 16th |
| Department of Software Engineering, | Department of Software Engineering, |
| Faculty of Science & Information Technology, | Faculty of Science & Information Technology, |
| Daffodil International University. | Daffodil International University. |

**Certified by:**

24/12/2018

**Dr. Md. Asraf Ali**

**Associate Professor**

Department of Software Engineering,

Faculty of Science & Information Technology,

Daffodil International University

# ACKNOWLEDGEMENT

As a matter of first importance, we are thankful to the Almighty ALLAH for making us qualified to finish this thesis. At that point we want to thank our supervisor Dr. Asraf Ali, Associate Professor, Department of Software Engineering, Daffodil International University. For his help and direction amid the undertaking and for offering the thoughts and probability to take a shot at our recommended us. We are greatly appreciative and obliged to our master, genuine and important direction and support reached out to us.

Close to my supervisor, we wish to express our earnest on account of Professor Dr. Touhid Bhuiyan, Professor & Head, Department of Software Engineering, Daffodil International University, for his consistent support and encouragement. We accept this open door to record our earnest because of all the faculty members of the Department of Software Engineering for their assistance and support.

To wrap things up, we might want to thank our family and friends for their genuine help, love and without this it was outlandish for us to come this far.

# TABLE OF CONTANT

**CHAPTER 4: RESULTS AND DISCUSSION**

**CHAPTER 5: CONCLUSIONS AND RECOMMENDATIONS**

# LIST OF FIGURES

# LIST OF TABLES

## APENDIX A

## APENDIX B

## APENDIX C

# ABSTRACT

## Background

Medical Services is a big commercial aspect in every time. Revenue stream always running in this fields, so many dissatisfied clients and patients and health hanging in the balance, it looks there is an excessive need for one platform for problem solves in healthcare.

## Objectives

The main objective is to examined the performance of different data mining techniques, evaluated the results and compare the models in order to find the best result of diabetes by prediction.

## Methodology

For our study we select some features and algorithms then a suitable is dataset collected from the National Institute of Diabetes and Digestive and Kidney Diseases and checked is there any missing value is available in the dataset. Then redundant data is removed by performing co-relation. Later this three classification algorithms Logistic Regression (LR), Artificial Neural network (ANN), Decision tree (DC) and Support Vector Machine (SVM) are used to find out the accuracy for the prediction. Then compared the accuracy of these three models which one is better for predict diabetes.

## Results

This work uses 4 classification techniques for diabetes prediction. They are, Logistic Regression (LR), Artificial Neural Network (ANN), Decision Tee (DT) and Support Vector Machine (SVM). The performance of different classification technique was evaluated on different measurements techniques. The analysis results show that LR achieved highest performance than the other classifiers.

## Conclusions

Moreover, our current study basically centered on the use of different data mining algorithms for diabetes prediction and discover which of them will convey more accurate prediction.

**Keywords:** Data Mining; Classification; Supervised Learning; Diabetes Prediction

# CHAPTER 1

# INTRODUCTION

## 1.1 Background

Medical Services is a major business perspective in all of time. Income stream continuously running in this field, such a significant number of disappointed customers and patients and wellbeing remaining in a precarious situation, it looks there is an unreasonable requirement for one stage for issue tackles in medical services. Here is our principle thought for enhancing medical services administrations is to put all the more featuring on avoidance and less on treatment. In the event that we consider each medical problem, few out of every odd medical problem is preventable, yet by and large, early perception and discovery can pointer to great wellbeing results and diminish costs. Likewise, the key components of preventive human services are the illness screen. Early observing of ailments can be analyzed while still in the early, repairable stage. Be that as it may, it isn’t feasible for each patient to get screen for each likely sickness. A better arrangement is than have a reasonable, open and dependable methods for computing malady chance and foreseeing ailment presence.

## 1.2 Motivation of the research

Diabetes is a noticeable ailment that influences tremendous measure of human around the globe. It is a ceaseless malady that happens either when the pancreas does not gather enough insulin or when the body can't viably use the insulin it creates. It is developing colossally, in light of unfortunate way of life, take more rich and shoddy nourishment and absence of physical action. As indicated by report of World Health Association (WHO), In 2014, 8.5% of grown-ups matured 18 years and more seasoned had diabetes. In this way, in 2016, diabetes was the immediate reason for 1.6 million passing’s and in 2012 high blood glucose was the reason for another 2.2 million passing’s **(How Many People Have Diabetes, 08-jun-2018)**. In addition, diabetes is a key reason for visual impairment, kidney disappointment, heart assaults, stroke and lower appendage removal.

With the end goal to find our speculation and to see how we will construct a foreseeing model, we should initially investigate what causes diabetes. Pool of mindfulness about wellbeing and taking undesirable nourishment are the two noteworthy reasons for this ailment. It is sheltered to expect that patients with diabetes are more liable to have higher circulatory strain and higher sugar level. Be that as it may, diabetes may have different impacts on blood.

As referenced previously, it is imperative for specialists to have the capacity to analyze diabetes in its beginning periods and to recommend appropriate drugs that treat this condition. A basic blood test can have a great deal of ramifications of the wellbeing and it is effectively open and generally shabby for open to get analyze. So, we chose to put together our model with respect to distinctive parameters result that can really be actualized, in actuality, thinking about its minimal effort and simply get to. On the off chance that fruitful the application worked in the undertaking could turnover numerous pharmaceutical organizations what's more, in particular, it can spare a large number of lives with simple procedure of analysis.

## 1.3 Problem Statement

Chronic diseases are the noticeable reasons for death and handicap around the world. Infection rates from these conditions are stimulating around the world, progressing crossways every nation and saturating every single financial individual.

Diabetes is an unending, metabolic infection portrayed by raised dimensions of glucose, which alludes to genuine damage to the heart, veins, eyes, kidneys, and nerves. In diabetes, Age don't rely upon this event. There are three sorts of diabetes **("Diabetes" ,08-Jun-2018)**. The most well-known is type 2 diabetes, as a rule it occurs on grown-ups. Consequently, the body ends up impenetrable to insulin or doesn't create enough measure of insulin. Moreover, In the historical backdrop of past decades event of sort 2 diabetes has risen dramatically in nations of all salary levels. Type 1 diabetes, when known as adolescent diabetes or youth diabetes, is a perpetual condition in which is because of the absence of insulin generation. Hence, Type II Diabetes, is an exceptionally famous type of diabetes and it contains enormous measure of individuals in the entire world. What's more, type III is Gestational diabetes. It occurs, in light of changes about hormones when patients consumed pregnancy. In any case, Early identification and expectation can be anticipate it.

## 1.4 Research Questions

Is there any issue on various data mining techniques for diabetes prediction?

## 1.5 Research Objectives

According to the research the objectives of thesis is to examine the performance of different data mining techniques to find out the best models for diabetes prediction.

## 1.6 Research Scope

Diabetes presented to reliably high blood glucose levels can create genuine auxiliary confusions, including coronary illness, stroke, visual deficiency, kidney disappointment and ulcers that require the removal of toes, feet or legs.

With the end goal to anticipate which diabetic patients have a high hazard for these complexities, doctors may utilize numerical models. For instance, the UKPDS Risk Engine ascertains a diabetic patient's danger of coronary illness and stroke — in view of their age, sex, ethnicity, smoking status, time since diabetes determination and different factors but this technique doesn't give the exactness required by specialists.

## 1.7 Thesis Organizations

Chapter 1 presents the all fundamental recourse for this research. First of all, the foundation of the research where it has cleared the whole process of this research and why it ought to be required. Then the motivation the thesis and problem statement are described. After these some research question and research object was declared based on the research scope.

Chapter 2 is the literature review where discuss about the previous research papers. Literature review includes Article searching procedure, Article inclusion and exclusion benchmark, Data extraction and Article Search Results. Article searching procedure means to find out the articles from different sites. Then article inclusion and exclusion benchmark mean to find out the exact article by using some keywords and Data extraction is to read and analyze the include articles to collect the key information. Article search results is the result of the selected articles that are going to be needed in this work.

Chapter 3 is the research methodology where discuss about the dataset collection process and features of the dataset. For this work we collect the dataset from the National Institute of Diabetes and Digestive and Kidney Diseases.

Chapter 4 discuss about the result of the analysis. All the data has processed to make a final result base on the research goal and it ensure about the final goal that can archive.

Chapter 5 talked about the conclusion and recommendation for the future work of the research.
The final outcome of the analysis and the next step to solve this problem is discuss in this part.

# CHAPTER 2

# LITERATURE REVIEW

## 2.1 Background

We used a symmetric searching method to pick out all of the available articles that discuss about diabetes prediction using data mining technique. In our systematic procedure, we search related keywords from Google scholar, IEEE, Springer, Pubmed, Elsevier and Sciencedirect databases in order to access the article. We used the keywords, “diabetes prediction using data mining” to find journal articles published in English Language. We studied the title, abstract, methodology, results and conclusion of each article. Articles those written in English and that used data mining techniques were considered. We discovered 15 articles that have found in renowned journals and conference papers.

## 2.2 Literature Review Summary

**(Meng, Huang, Rao, Zhang, & Liu, 2013)**, describe the comparison of three data mining techniques, like Logistic Regression (LR), Artificial Neural Network (ANN) and Decision Tree (DT) for diabetes prediction. A standard questionnaire was directed to acquire data on statistic on 1487 individuals in Zhuguang and Liurong communities in Guangzhou, China. Then they developed three predictive models and evaluate the three models in terms of their accuracy, sensitivity and specificity by using12 input factors and one output variables from the questionnaires output. The Decision tree (DT) show the best classification performance and the ANN show the most nominal performance. The logistic regression model gained a classification accuracy of 76.13% with a sensitivity of 79.59% and a specificity of 72.74%. The ANN model reached a classification accuracy of 73.23% with a sensitivity of 82.18% and a specificity of 64.49%; and the decision tree (C5.0) gained a classification accuracy of 77.87% with a sensitivity of 80.68% and specificity of 75.13%. For choosing the optimal predictive models for implementing community lifestyle interventions to decrease the radiation of diabetes their studies would help in future.

**(Concaro, Sacchi, Cerra, & Bellazzi, 2009)**, introduced the features of the fundamental possibilities of the use of temporal association rules for the mining of healthcare databases. The Regional Healthcare Agency (ASL) of Pavia has been gathering and keeping up a central data repository which stores both managerial and clinical medicinal services information about the number of inhabitants in Pavia region. They just centered around the consideration conveyance stream of Diabetes Mellitus, and demonstrate the use of an algorithm for the extraction of Temporal Association Rules on sequences of hybrid events. This procedure grants to suitably manhandle the mix of healthcare information sources, and can be utilized to assess the relevance of the consideration conveyance stream for explicit pathologies, so as to reassess or refine the unseemly practices which lead to inadmissible results.

**(Aljumah, Ahamad, & Siddiqui, 2013),** concentrates upon predictive analysis of diabetic treatment utilizing a regression-based data mining technique. A software mining tool Oracle Data Miner (ODM) was employed for predicting modes of treating diabetes. For experimental analysis the support vector machine algorithm was used. Datasets of Non-Communicable Diseases (NCD) risk factors. This study, predictions on the effectiveness of various treatment strategies for young and maturity teams were elucidated. The particular orders of treatment were found to be different.

**(Perveen, Shahbaz, Guergachi, & Keshavjee, 2016),** the prevalence of Diabetes Mellitus is increasing at a fast pace, deteriorating human, economic and social fabric. From the Canadian Primary Care Sentinel Surveillance Network (CPCSSN) database the dataset was acquired for this analysis. Over three age groups in the Canadian population, using bagging adaboost as well as J48 decision tree to classify patients with diabetes mellitus using diabetes risk factors, the study formed gradually good models with higher performance to classify diabetic patients. Result of this analysis showed that, overall performance of adaboost ensemble method is better than bagging as well as standalone J48 decision tree.

**(Concaro, Sacchi, Cerra, Stefanelli, et al., 2009),** the main potentials of the application of temporal association rules for the mining of healthcare databases is highlighted in this study. A compact repository including both administrative and clinical data related to a sample of diabetic patients is turned out through a temporal data mining technique by the application of to extract temporal association rules. This process could be thoroughly absorbed to assess the overall standards and quality of care, while reducing costs considered by the perspective of a Regional Healthcare Agency.

**(Sa-ngasoongsong & Chongwatpol, 2012),** in this study,Akkarapol and Jongsawas proposed a framework so that data mining techniques can predict better and explain the causes of increasing diabetes. They provide the in-depth analysis on how data mining approach can be a great in diabetes prediction. They followed the CRISP-DM Model (Cross Industry Standard Process for Data Mining). CRISP-DM Model used as a comprehensive data mining methodology and process model for conducting this data mining study. At first, they analyze the relevant data set and discover the insights into the data. A large number of data was collected. After that they preprocess the raw data and finalize a dataset with 50,788 records, consists of 43 variables. Then they apply analytical data mining techniques to predict and explain factors that increase the prevalence of diabetes in the patient samples. Different models are constructed and compared in order to predict patients with diabetes. Logistic regression produces the best results with overall misclassification rate of 22.89%. Although Artificial Neural Networks (ANN) has the lowest false negative rate of 20.55%.

**(Joshi & Borse, 2017)**, it is critical to guess device which can be utilized to decide if somebody has diabetes or not. There are a few techniques which create precise expectation and Artificial neural system utilizing Back engendering neural system is one of them. This neural system having an info layer with 8 parameters, one concealed layer with 10 neurons and one yield layer is actualized to create great outcomes. The GUI is created so specialist can stack input parameter perusing and can apply preparing at whatever point required and results for a solitary or various patient can be gotten. There is expansion alternative is accommodated single and numerous patients. This paper outlines the usage and advancement of the product device buIlt in MATLAB which will foresee whether somebody is diabetic or not.

**(Jayalskshmi & Santhakumaran, 2010),** this paper shows the effect of missing value technique and pre-processing technique. A classifier has connected to Pima Indian Diabetes dataset and the outcomes were enhanced hugely when utilizing certain mix of preprocessing and missing value techniques. It demonstrates that a few mixes of missing values and pre-processing the precision was colossally made strides. This demonstrates preprocessing and missing values assume a noteworthy job in characterization. In future this can be connected for various preparing techniques and systems.

**(Georga, Protopappas, Mougiakakou, & Fotiadis, 2013)**, endless consideration of diabetes accompanies expansive measures of information concerning oneself and clinical administration of the infection. In this paper, it is proposed to treat that data from two alternate points of view. Right off the bat, a predictive model of short-term glucose homeostasis depending on machine learning is given the point of averting hypoglycemic events and prolonged hyperglycemia on a day by day premise. Second, information mining approaches are proposed as an instrument for clarifying and predicting the long-term glucose control and the frequency of diabetic inconveniences.

**(Sisodia & Sisodia, 2018)**, one of the essential true restorative issues is the identification of diabetes at its beginning time. In this analysis, precise endeavors are made in planning a framework which results in the expectation of ailment like diabetes. Amid this work, three machine learning classification algorithms are contemplated and assessed on different measures. Investigations are performed on Pima Indians Diabetes Database. Exploratory outcomes decide the sufficiency of the structured framework with an accomplished exactness of 76.30 % utilizing the Naive Bayes classification algorithm. In future, the structured framework with the utilized machine learning classification algorithms can be utilized to predict or diagnose other diseases. The work can be broadened and enhanced for the robotization of diabetes analysis including some other machine learning algorithms.

**(Kavakiotis, Tsave, Salifoglou, Maglaveras, el al, 2017),** in this analysis, a precise exertion was made to distinguish and survey machine learning and data mining approaches connected on DM research. DM is quickly developing as one of the best worldwide wellbeing difficulties of the 21st century. To date, there is a huge work done in almost all parts of DM research about and particularly biomarker identification and prediction-diagnosis. The approach of biotechnology, with the immense measure of information delivered, alongside the expanding measure of EHRs is anticipated that would offer ascent to assist top to bottom investigation toward finding, etiopathophysiology and treatment of DM through work of machine learning and data mining techniques in improved datasets that include clinical and organic data.

**(Wu, Yang, Huang, He, & Wang, 2018),** this paper meant to set up a proper forecast display for the high-risk T2DM group. In view of some of analysts' encounters, a novel model is proposed, which comprises of double-level algorithms, the enhanced K-means and logistic regression algorithms. So as to make a substantial correlation with others' outcomes, it was important to direct this model utilizing the WEKA toolkit and utilize a similar Pima Indian Diabetes Dataset. Appropriate channels were used to enhance the legitimacy and reasonability of the dataset. The proposed model that comprised of both group and class strategy guaranteed the upgrade of prediction accuracy. In Section 4, another sensible dataset given by Dr. Schorling was used to test and confirm the model. The proposed model has ended up being fitting for predicting T2DM. One of the proposed model's advantages was that it avoids deleting overmuch original data. It guarantees the high caliber of test information. The other advantage was that the model can apply in the Pima Indian Diabetes Dataset and additionally different datasets. While the confinement was that it consumes more time during the part of preprocessing.

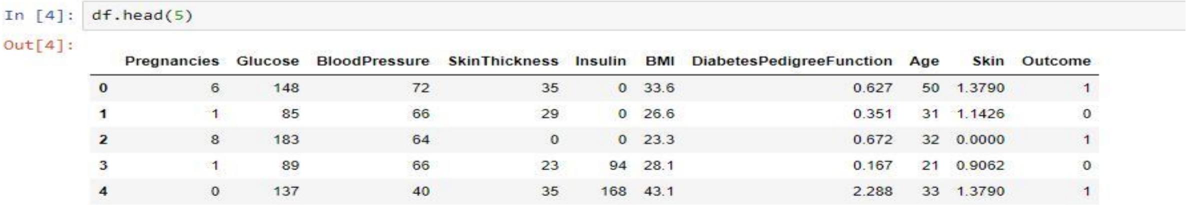
# CHAPTER 3
# RESEARCH METHODOLOGY

## 3.1 Data collection

For this work, the dataset that has been used was collected from National Institute of Diabetes and Digestive and Kidney Diseases **(Research Summary | National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), 2018)**. Some medical separate variables are given in the dataset, for example, pregnancy record, BMI, insulin level, age, glucose fixation, diastolic circulatory strain, triceps skin crease thickness etc. This dataset consists of 768 patient's information where every one of the patients are female and not less than 21 years of age. The quantity of true cases is 268 (34.90%) and the quantity of false cases are 500 (65.10%), separately, in the dataset. In this work we picked eight particular parameters for data processing, for example,

a) Pregnancies: Number of times pregnant
b) Glucose: Plasma glucose concentration 2 hours in an oral glucose tolerance test
c) Blood Pressure: Diastolic blood pressure (mm Hg)
d) Skin Thickness: Triceps skin fold thickness (mm)
e) Insulin: 2-Hour serum insulin (mu U/ml)
f) BMI: Body mass index (weight in kg/ (height in m) ^2)
g) Diabetes Pedigree Function: Diabetes pedigree function
h) Age: Age (years)
i) Outcome: Class variable (0 or 1)

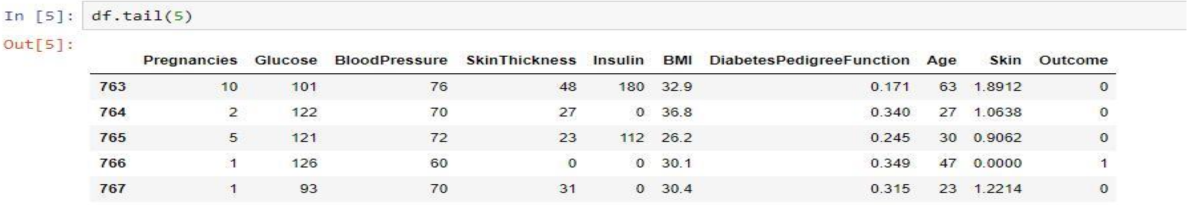
Figure 1 show the top 5 rows of the dataset and figure 2 show the bottom 5 rows of the data set.



```
In [4]: df.head(5)
Out[4]:
```

| | Pregnancies | Glucose | BloodPressure | SkinThickness | Insulin | BMI | DiabetesPedigreeFunction | Age | Skin | Outcome |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 6 | 148 | 72 | 35 | 0 | 33.6 | 0.627 | 50 | 1.3790 | 1 |
| **1** | 1 | 85 | 66 | 29 | 0 | 26.6 | 0.351 | 31 | 1.1426 | 0 |
| **2** | 8 | 183 | 64 | 0 | 0 | 23.3 | 0.672 | 32 | 0.0000 | 1 |
| **3** | 1 | 89 | 66 | 23 | 94 | 28.1 | 0.167 | 21 | 0.9062 | 0 |
| **4** | 0 | 137 | 40 | 35 | 168 | 43.1 | 2.288 | 33 | 1.3790 | 1 |

Figure 2.1: Top 5 rows of the dataset



```
In [5]: df.tail(5)
Out[5]:
```

| | Pregnancies | Glucose | BloodPressure | SkinThickness | Insulin | BMI | DiabetesPedigreeFunction | Age | Skin | Outcome |
|---|---|---|---|---|---|---|---|---|---|---|
| **763** | 10 | 101 | 76 | 48 | 180 | 32.9 | 0.171 | 63 | 1.8912 | 0 |
| **764** | 2 | 122 | 70 | 27 | 0 | 36.8 | 0.340 | 27 | 1.0638 | 0 |
| **765** | 5 | 121 | 72 | 23 | 112 | 26.2 | 0.245 | 30 | 0.9062 | 0 |
| **766** | 1 | 126 | 60 | 0 | 0 | 30.1 | 0.349 | 47 | 0.0000 | 1 |
| **767** | 1 | 93 | 70 | 31 | 0 | 30.4 | 0.315 | 23 | 1.2214 | 0 |

Figure 2.2: Bottom 5 rows of the dataset

## 3.2 Description of the Classification Techniques

### 3.2.1 Logistic Regression (LR)

Logistic regression is a measurable strategy for analyzing a dataset in which there are something like one distinct variable that choose an outcome. The outcome is estimated with a dichotomous variable (in which there are only two conceivable outcomes 0 or 1). Logistic Regression (LR) utilize a black box function. This black box function is prominently known as the Softmax function. As of late, LR is a prominent strategy for binary classification problems. Logistic Regression was utilized in the organic sciences in mid twentieth century **(Berkson, 1944)**.

Application of Logistic Regression:

Logistic Regression is being used in Healthcare, Social Sciences, various ML and Data Mining for advanced research & analytics.

Logistic Regression formula:

$$In\ (p/1-p) = b0+b1*x$$

Logistic Function:

Logistic regression is named for the capacity utilized at the center of the technique, the logistic function. The logistic function, additionally called the sigmoid capacity. The sigmoid/logistic function is given by the accompanying condition.

$y = 1 / 1+ e^{-x}$

It is an S-shaped curve that gets closer to 1 as the value of input variable increases above 0 and gets closer to 0 as the input variable decreases below 0. The output of the sigmoid function is 0.5 when the input variable is 0.

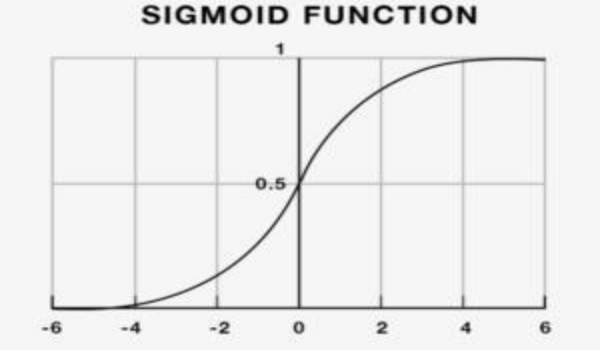


Figure 3.1: Sigmoid Function

Thus, if the output is more than 0.5, we can classify the outcome as 1 (or positive) and if it is less than 0.5, we can classify it as 0 (or negative).

### 3.2.2 Artificial Neural Network (ANN)

The subsequent model from neural computing is frequently called an artificial neural network (ANN) or a neural network **(Gaur, 2012)**. Artificial neural network (ANN) is an imperative machine learning strategy for biological research. In machine learning, ANN is an advantageous computational model which works like biological neurons **(Hecht-Nielsen, 1992)**.

ANN mainly organized into three layers

i. Input layer
ii. Hidden layer or processing stage
iii. Output layer.

In addition, the result of output layer at each node is called its activation or node value **(van, & Bohte, 2018)**.

The data mining dependent on neural network is formed by information arrangement, rules removing and governs appraisal three stages. Initial one is information planning, at that point Rules separating and last one is Rules appraisal.

Artificial neural network is entirely appropriate for taking care of the issues of data mining since its attributes of good power, self-arranging versatile, parallel preparing, disseminated capacity and high level of adaptation to non-critical failure. The utilization of neural network in data mining is a promising field of research particularly given the prepared accessibility of vast mass of informational collections and the detailed capacity of neural network to recognize and absorb connections between a substantial numbers of factors.

### 3.2.3 Decision Tree (DT)

Decision tree (DT) is one of the well-known directed learning-based classification algorithms. For tackling relapse and classification problems, DT can be utilized for both. Decision Tree is to make a preparation show which can use to predict class or estimation of target factors by taking in decision rules inferred from earlier information (training data). DT is a classification technique which breaks a dataset into littler subsets and last arrangement with related decision tree is gradually created **(Safavian, & Landgrebe, 1991)**.

**Decision Tree Algorithm Pseudocode**

```
1.    Place the best characteristic of the dataset at the foundation of the tree.

2.    Split the training set into subsets. Subsets ought to be made so that every
      subset contains information with a similar incentive for a trait.

3.    Repeat stage 1 and stage 2 on every subset until the point that you
      discover leaf hubs in every one of the parts of the tree.
```

There are many Decision Tree like Classification and Regression Trees (CART), Iterative Dichotomiser 3 (ID3), C4.5 etc. CART uses Gini Index (Classification) as metric. ID3 uses Entropy function and Information gain as metrics. C4.5 is the improvement version of ID3.

ID3 grows tree classifiers in three steps:

1. Determination of target characteristic and count of entropy of characteristics.
2. Select characteristic with most noteworthy data gain measure
3. Make hub containing that characteristic. Iteratively apply these steps to new tree branches and quit developing tree in the wake of checking of stop foundation.

The ID3 decision makes use of two concepts when creating a tree from top-down **(Fong, & Weber-Jahnke, 2012)**:

1. Entropy
2. Information Gain

Using these two concepts, the nodes to be created and the characteristic to split on can be determined.

C4.5 algorithm is enhancement to ID3. C4.5 can handle continuous input attribute. It follows three steps during tree growth **(Karaolis, Moutiris, Hadjipanayi, & Pattichis, 2010)**:

1. Part of all out categorical attribute is same to ID3 algorithm. Ceaseless attribute dependably produces paired parts
2. Characteristic with most elevated gain proportion is chosen.
3. Iteratively apply these steps to new tree branches and quit developing tree in the wake of checking of stop model.

### 3.2.4 Support Vector Machine (SVM)

Support vector machine (SVM) is a supervised learning calculation. It can utilize for both grouping or relapse issues however generally it is utilized in characterization issues. SVM function admirably for some, human services issues and can comprehend both linear and non-linear issues.

**(Vapnik, 1995)** and **(Chervonenkis, 2013)** presented the Support vector machine grouping strategy which is endeavor to pass a linearly separable hyperplane to order the dataset into two classes. At long last, the model can without a doubt gauge the objective groups (labels) for new cases.

Algorithm

1. Define an ideal hyperplane: maximize margin
2. Extend the above definition for non-directly distinguishable issues: have a punishment term for misclassifications.
3. Map information to high dimensional space where it is less demanding to arrange with linear decision surfaces: reformulate issue so information is mapped verifiably to this space.

For implementing SVM we need some libraries like NumPy, Pandas, Matpot.lib. besides this dataset also need to import. Then data preprocessing that involves:

1. Dividing the information into attributes and labels
2. Dividing the information into training and testing sets.

After that we have to prepare SVM on the training data. At that point expectation and assess the algorithm. Toward the end a last outcome will be found.

# CHAPTER 4
# RESULTS & DISCUSSION

## 4.1 Measurement of Classification Techniques

Performance of all the classification algorithms are assessed by different statistical measurement aspects such as accuracy, sensitivity, specificity etc. These classification measurement factors are calculated by the terms: True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN).

**True Positive (TP):** These are the correctly predicted positive values which means that the value of actual class is yes and the value of predicted class is also yes.

**True Negative (TN):** These are the correctly predicted negative values which means that the value of actual class is no and value of predicted class is also no.

**False Positive (FP):** When actual class is no and predicted class is yes is called False Positive.
It is a type 1 error.

**False Negative (FN):** When actual class is yes but predicted class in no. It is a type 2 error.

**Formula:**

**Accuracy:** Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations.

Accuracy = $(TP+TN)$ /$(TP+FP+FN+TN)$

**Precision:** Precision is the ratio of correctly predicted positive observations to the total predicted positive observations.

Precision = $TP$ $(TP+FP)$

**Recall** (Sensitivity): Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes.

Recall = $TP$ /$(TP+FN)$

**F1 score**: F1 Score is the weighted average of Precision and Recall. F1 is usually more useful than accuracy but is not as easy to understand as accuracy.

F1= $2*(Recall*Precision)$ $Recall+Precision$

**Specificity:** Specificity is a measure which defines the ratio of the patients that do not have diabetes, and also predicted by the model as non-diabetes. Specificity is the opposite of recall.

Specificity = $TN$ ($TN$+$FP$)

## 4.2 Analysis of the Results

Four data mining techniques were analyzed to evaluate the performance for diabetes prediction. We analyzed 768 data samples where there are training and test data. There are 70% training data and 30% test data. The dataset containing Original True cases are 268 (34.90%), Original False cases are 500 (65.10%, Training True cases are 188 (35.01%), Training False cases are 349 (64.99%), Test True cases are 80 (34.63%), Test False cases are 151 (65.37%). We also found that two columns are correlated with each other's. They are thickness and skin whereas 1 to 1. Hence, we dropped by the duplicate skin column by del function. Figure 1 shows the Heat map for checking correlated columns. Figure 2 shows dropped redundant column from datasets and figure 3 shows Statistical results from diabetes datasets. Figure 4 present classification performance of Data Mining Techniques (True Cases) and Figure 5 present classification performance of Data Mining Techniques (False Cases).

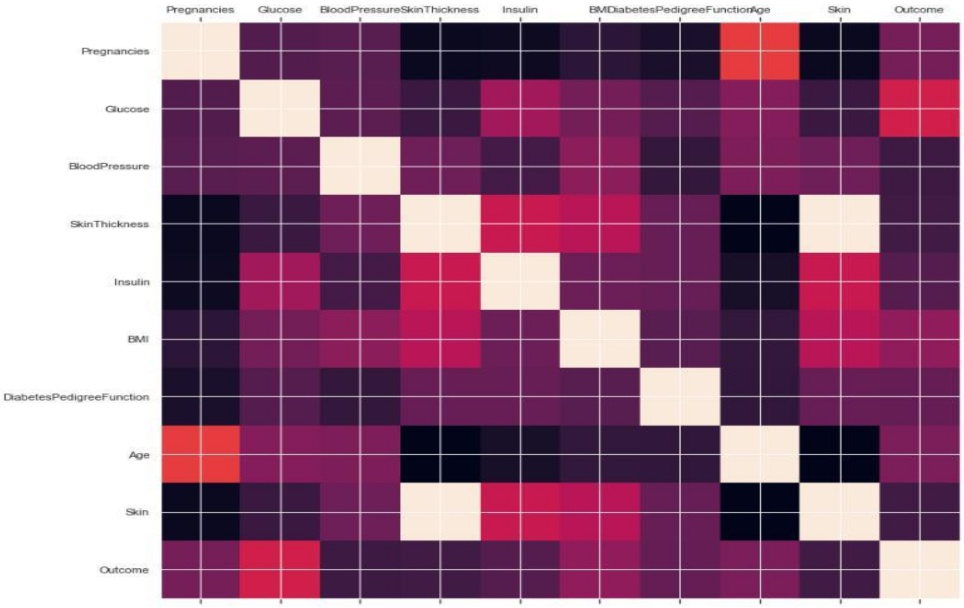
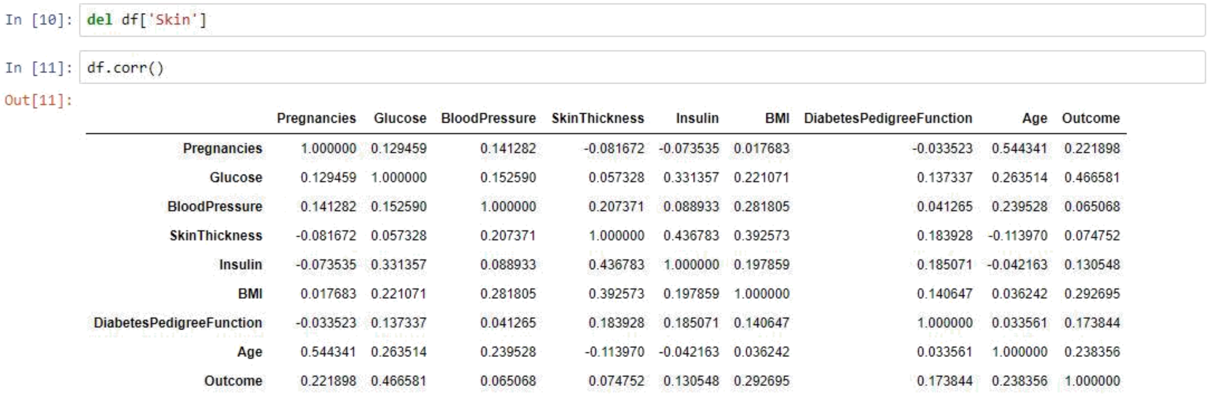


Figure 4.1: Heat map for checking correlated columns.



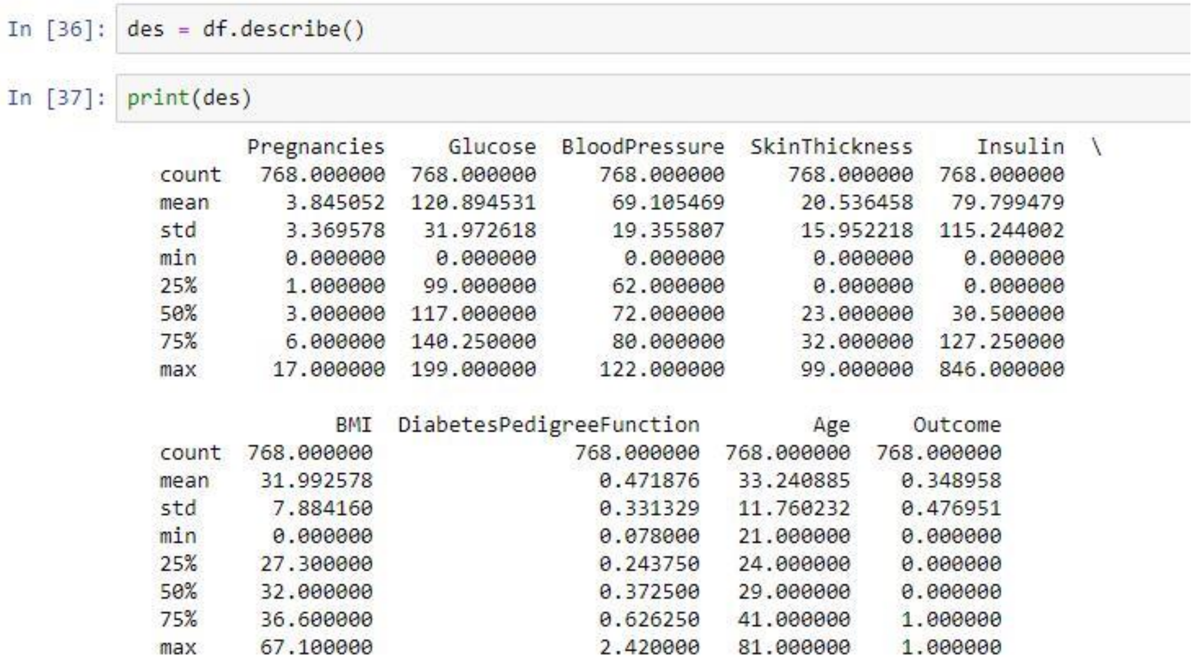
```
In [10]: del df['Skin']
```

```
In [11]: df.corr()
```

Out[11]:

| | Pregnancies | Glucose | BloodPressure | SkinThickness | Insulin | BMI | DiabetesPedigreeFunction | Age | Outcome |
|---|---|---|---|---|---|---|---|---|---|
| **Pregnancies** | 1.000000 | 0.129459 | 0.141282 | -0.081672 | -0.073535 | 0.017683 | -0.033523 | 0.544341 | 0.221898 |
| **Glucose** | 0.129459 | 1.000000 | 0.152590 | 0.057328 | 0.331357 | 0.221071 | 0.137337 | 0.263514 | 0.466581 |
| **BloodPressure** | 0.141282 | 0.152590 | 1.000000 | 0.207371 | 0.088933 | 0.281805 | 0.041265 | 0.239528 | 0.065068 |
| **SkinThickness** | -0.081672 | 0.057328 | 0.207371 | 1.000000 | 0.436783 | 0.392573 | 0.183928 | -0.113970 | 0.074752 |
| **Insulin** | -0.073535 | 0.331357 | 0.088933 | 0.436783 | 1.000000 | 0.197859 | 0.185071 | -0.042163 | 0.130548 |
| **BMI** | 0.017683 | 0.221071 | 0.281805 | 0.392573 | 0.197859 | 1.000000 | 0.140647 | 0.036242 | 0.292695 |
| **DiabetesPedigreeFunction** | -0.033523 | 0.137337 | 0.041265 | 0.183928 | 0.185071 | 0.140647 | 1.000000 | 0.033561 | 0.173844 |
| **Age** | 0.544341 | 0.263514 | 0.239528 | -0.113970 | -0.042163 | 0.036242 | 0.033561 | 1.000000 | 0.238356 |
| **Outcome** | 0.221898 | 0.466581 | 0.065068 | 0.074752 | 0.130548 | 0.292695 | 0.173844 | 0.238356 | 1.000000 |

Figure 4.2: Dropped redundant column from datasets



```
In [36]: des = df.describe()
```

```
In [37]: print(des)
```

```
       Pregnancies     Glucose  BloodPressure  SkinThickness     Insulin  \
count   768.000000  768.000000     768.000000     768.000000  768.000000
mean      3.845052  120.894531      69.105469      20.536458   79.799479
std       3.369578   31.972618      19.355807      15.952218  115.244002
min       0.000000    0.000000       0.000000       0.000000    0.000000
25%       1.000000   99.000000      62.000000       0.000000    0.000000
50%       3.000000  117.000000      72.000000      23.000000   30.500000
75%       6.000000  140.250000      80.000000      32.000000  127.250000
max      17.000000  199.000000     122.000000      99.000000  846.000000

              BMI  DiabetesPedigreeFunction         Age     Outcome
count  768.000000                768.000000  768.000000  768.000000
mean    31.992578                  0.471876   33.240885    0.348958
std      7.884160                  0.331329   11.760232    0.476951
min      0.000000                  0.078000   21.000000    0.000000
25%     27.300000                  0.243750   24.000000    0.000000
50%     32.000000                  0.372500   29.000000    0.000000
75%     36.600000                  0.626250   41.000000    1.000000
max     67.100000                  2.420000   81.000000    1.000000
```

Figure 4.3: Statistical results from diabetes datasets

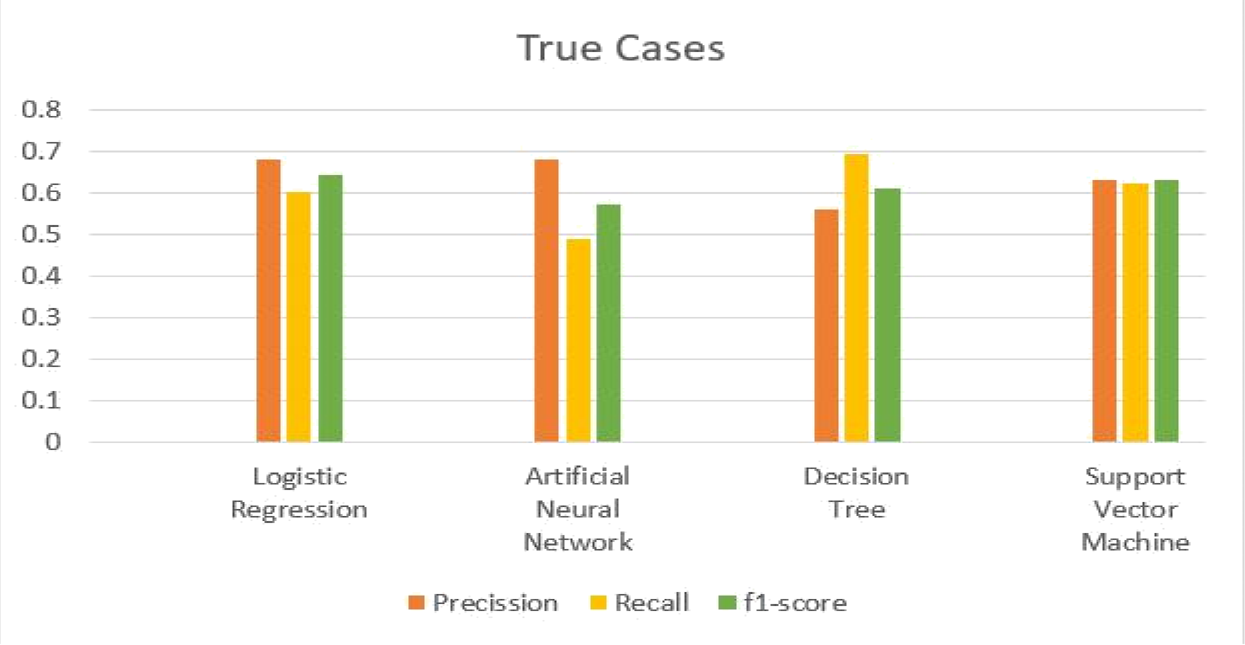


Figure 4.4: Classification Performance of Data Mining Techniques (True Cases)

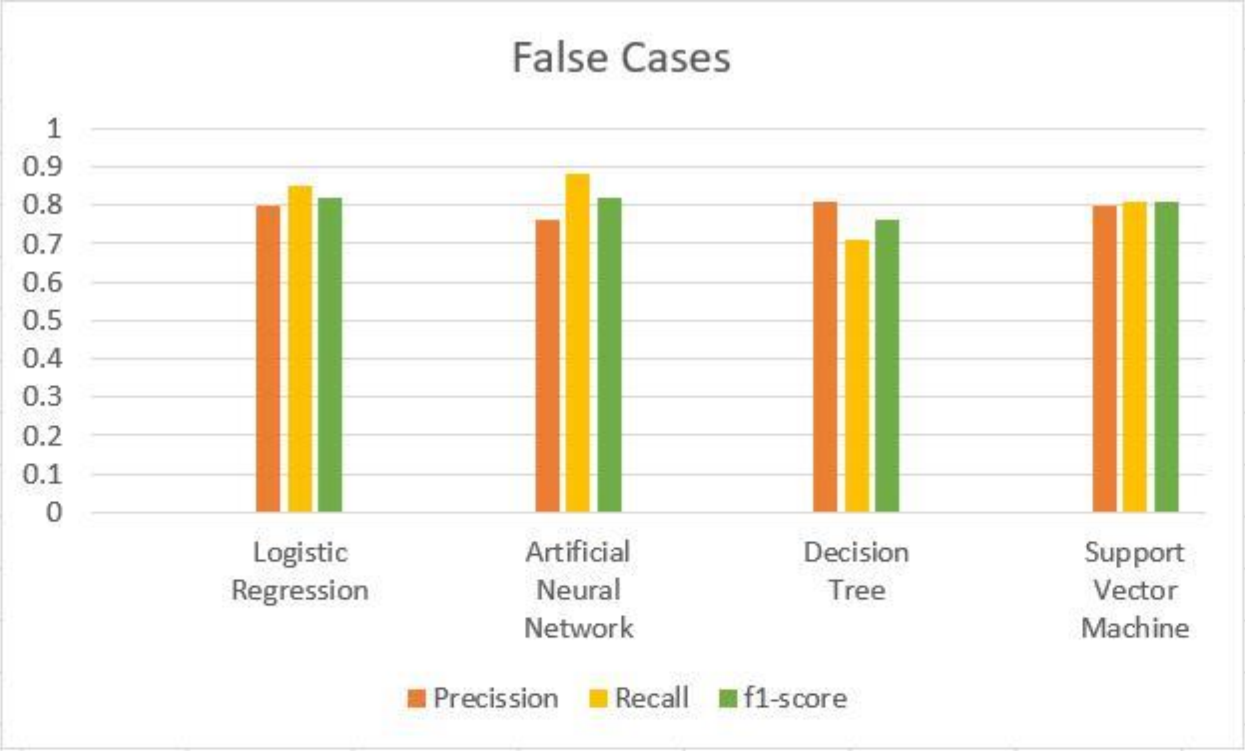


Figure 4.5: Classification Performance of Data Mining Techniques (False Cases)

In this experiment, we consider different analysis to investigate the four-machine learning algorithms for the classifications of Diabetes Datasets. Moreover, from the diabetes datasets all of the samples are evaluated. Figure 6 shows the confusion matrix (Classification results of the computational Intelligence for prediction of diabetes. Here, TP true positive, FP false positive, TN true negative, FP false positive) of the four classification algorithms. Figure 7 present the accuracy of four supervised based classifications techniques. Hence, LR achieved the best accuracy (i.e. 76%) and DT performed worst (i.e. 70%). Moreover, ANN and SVM are comparatively achieved same accuracy (i.e. 74%).

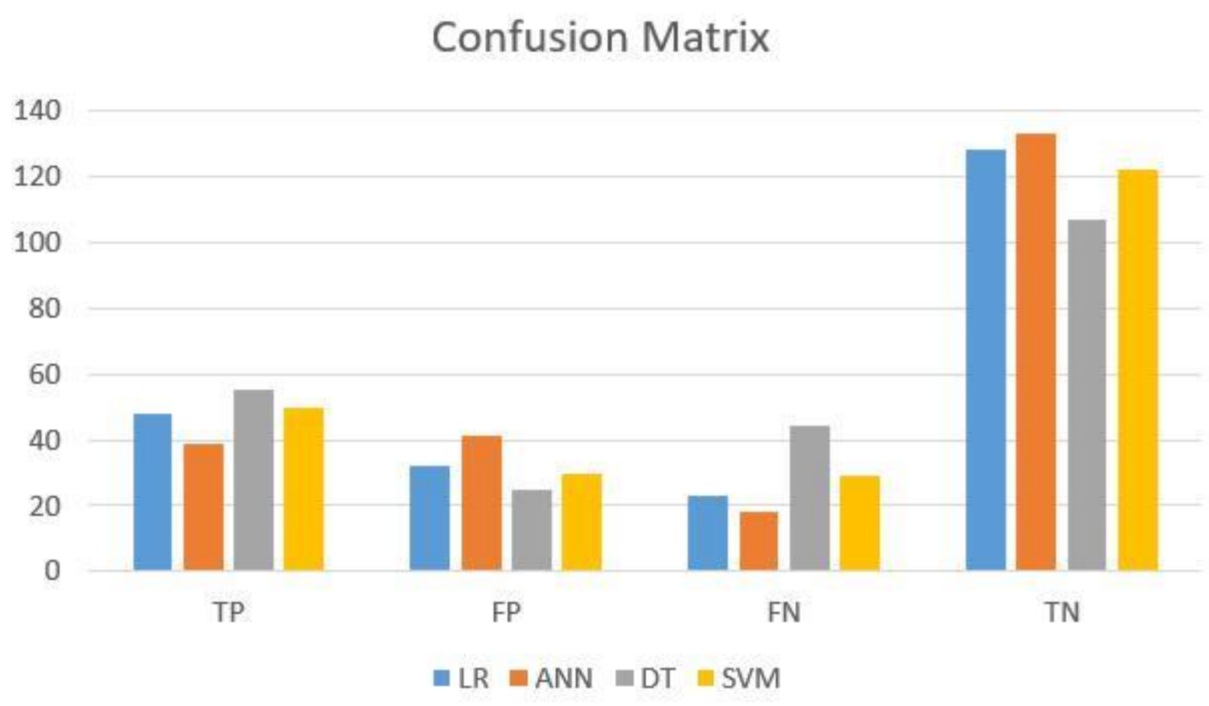


Figure 4.6: Confusion matrix of classification algorithm

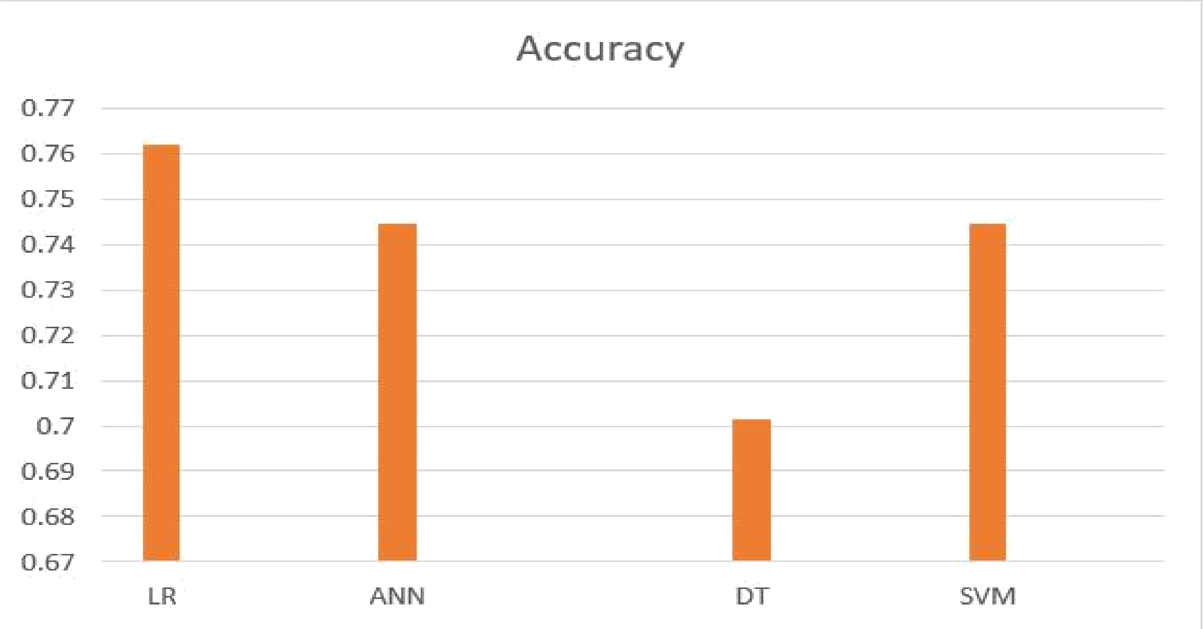


Figure 4.7: Classification accuracy of four classifiers.

### 4.3 Performance Evaluation

Figure 8 and table 1 show the performance of precision, recall, f-1 and specificity. In precision, LR and ANN achieved the highest performance 0.68% and DT is the lowest 0.56%. DT achieved the highest performance 0.695% in recall and ANN is the lowest 0.49%. In f-1, LR and in specificity, DT achieved the highest performance 0.64 and 0.81 respectively. Figure 8 represents the Receiver Operating Curve (ROC) for the selected four classification algorithms. It shows the true positive rate and false positive rate of the classifiers from the diabetes data analysis. For the best classification techniques, the area of the Receiver Operating Curve (ROC) must be close to one (1).

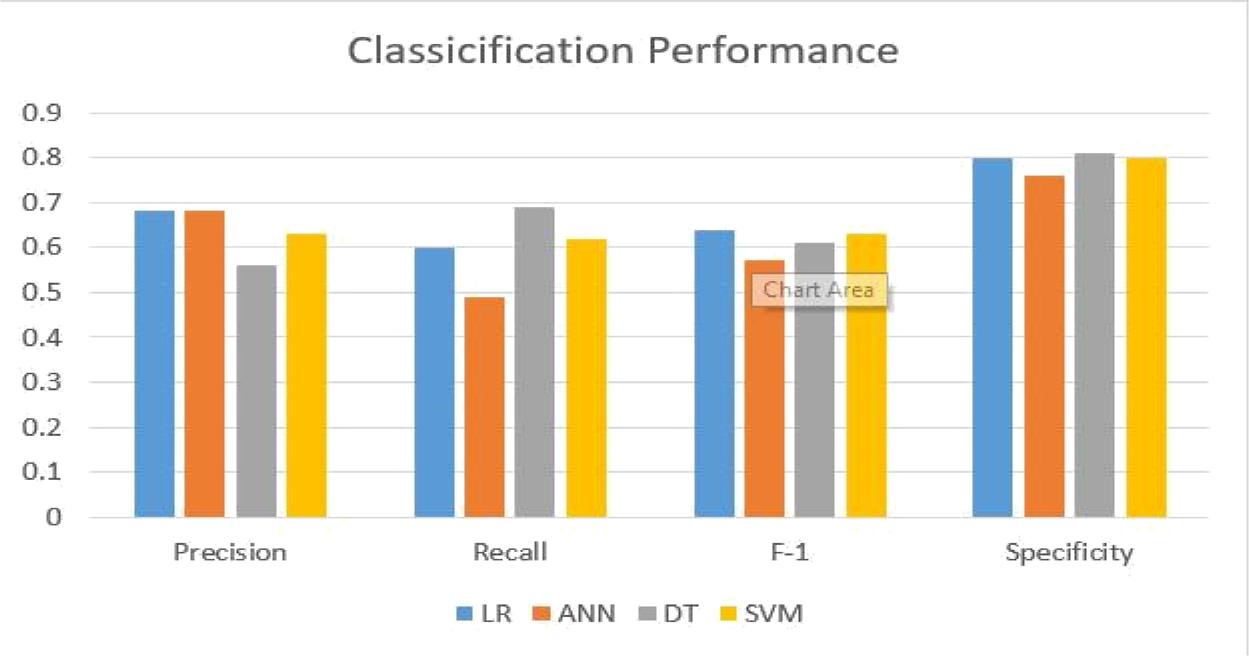


Figure 4.8: Classification Performance of Data Mining Techniques

| | LR | ANN | DT | SVM |
|---|---|---|---|---|
| Accuracy | 0.7619 | 0.7446 | 0.7013 | 0.7446 |
| Precision | 0.68 | 0.68 | 0.56 | 0.63 |
| Recall/Sensitivity/TPR | 0.60 | 0.49 | 0.69 | 0.62 |
| F-1 | 0.64 | 0.57 | 0.61 | 0.63 |
| Specificity/TNR | 0.80 | 0.76 | 0.81 | 0.80 |

Table 4.1: Classification Performance Measurements.

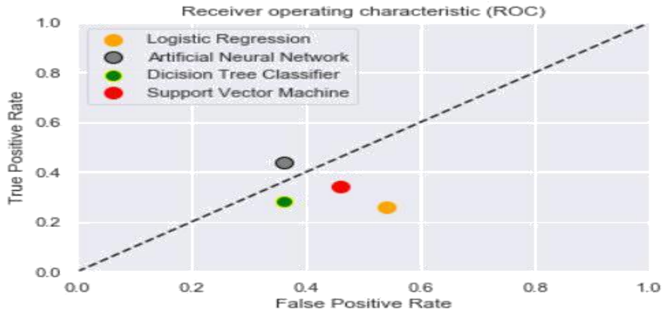


Figure 4.9: Receiver Operating Curve for four classification algorithms

In the above all figures show the Data Mining techniques performance results. All of the four Data Mining techniques Logistic Regression (LR) show the highest performance of accuracy is 0.76% and precision is 0.68%.

In our work logistic Regression show the highest performance. But there are no such things that LR always gives the best performance. SVM, Naïve Bayes, DT also can give best result in different situation or perspective.

**(I. Kavakiotis et al)** experimented between different supervised learning and using diabetes datasets. They showed SVM arise as the most successful and best performer algorithm.

**(Ramezankhani A, 2014)** described into their article, Decision Tree outperforms with the highest accuracy of 90.5% comparatively other algorithms and **(Xue-HuiMeng,** 2013) Decision tree model had the best accuracy is 77.87% than others, followed by the logistic regression model.

# CHAPTER 5

# CONCLUSION AND RECOMMENDATION

## 5.1 Conclusion

The following chapter concludes the findings in this study. This study analyzed different data mining techniques to find out the best data mining techniques for diabetes prediction. For the above study, first we search with key words and find out some articles, journals and papers that related to our analysis. We reviewed those articles individually and identified relevant articles from there. Then we collect a data set according to some diabetes features (Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes pedigree function, Age (years)) and performed four data mining techniques - Logistic Regression (LR), Artificial Neural Network (ANN), Decision Tee (DT) and Support Vector Machine (SVM). And then perform co-relation algorithm for omit internal dependency and redundant values. Among them Logistic Regression (LR) performed the best accuracy for our dataset. This work will help furthermore to predict diabetes accurately.

## 5.2 Recommendation for future work

There are many recommendations for future work. We used only four data mining techniques for this work but there are more data mining techniques. To predict more accurate, more algorithms can be used. By using more techniques performance can be improved.

# References

"How Many People Have Diabetes?" [Online]. [Accessed: 08-Jun-2018].

"Diabetes," 2017. [Online]. [Accessed: 08-Jun-2018].

Meng, X. H., Huang, Y. X., Rao, D. P., Zhang, Q., & Liu, Q. (2013). Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. *The Kaohsiung journal of medical sciences*, *29*(2), 93-99.

Concaro, S., Sacchi, L., Cerra, C., & Bellazzi, R. (2009, August). Mining administrative and clinical diabetes data with temporal association rules. In *MIE* (pp. 574-578).

Aljumah, A. A., Ahamad, M. G., & Siddiqui, M. K. (2013). Application of data mining: Diabetes health care in young and old patients. *Journal of King Saud University-Computer and Information Sciences*, *25*(2), 127-136.

Perveen, S., Shahbaz, M., Guergachi, A., & Keshavjee, K. (2016). Performance analysis of data mining classification techniques to predict diabetes. *Procedia Computer Science*, *82*, 115-121.

Concaro, S., Sacchi, L., Cerra, C., Stefanelli, M., Fratino, P., & Bellazzi, R. (2009). Temporal data mining for the assessment of the costs related to diabetes mellitus pharmacological treatment. In *AMIA Annual Symposium Proceedings* (Vol. 2009, p. 119). American Medical Informatics Association.

Sa-ngasoongsong, A., & Chongwatpol, J. (2012). An analysis of diabetes risk factors using data mining approach. *Oklahoma state university, USA.*

Joshi, S., & Borse, M. (2016, September). Detection and Prediction of Diabetes Mellitus Using Back-Propagation Neural Network. In *Micro-Electronics and Telecommunication Engineering (ICMETE), 2016 International Conference on* (pp. 110-113). IEEE.

Jayalskshmi, T., & Santhakumaran, A. (2010, February). Impact of preprocessing for diagnosis of diabetes mellitus using artificial neural networks. In *Machine Learning and Computing (ICMLC), 2010 Second International Conference on* (pp. 109-112). IEEE.

Georga, E. I., Protopappas, V. C., Mougiakakou, S. G., & Fotiadis, D. I. (2013, November). Short-term vs. long-term analysis of diabetes data: Application of machine learning and data mining techniques. In *Bioinformatics and Bioengineering (BIBE), 2013 IEEE 13th International Conference on* (pp. 1-4). IEEE.

Sisodia, D., & Sisodia, D. S. (2018). Prediction of Diabetes using Classification Algorithms. *Procedia Computer Science*, *132*, 1578-1585.

Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. *Computational and structural biotechnology journal*, *15*, 104-116.

Wu, H., Yang, S., Huang, Z., He, J., & Wang, X. (2018). Type 2 diabetes mellitus prediction model based on data mining. *Informatics in Medicine Unlocked*, *10*, 100-107.

Research Summary | National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK).

[Online]. [Accessed: 08-Jun-2018].

Berkson, J. (1944). Application of the logistic function to bio-assay. *Journal of the American Statistical Association*, *39*(227), 357-365.

Gaur, P. (2012). Neural networks in data mining. *International Journal of Electronics and Computer Science Engineering (IJECSE, ISSN: 2277-1956)*, *1*(03), 1449-1453.

Hecht-Nielsen, R. (1992). Theory of the backpropagation neural network. In *Neural networks for perception* (pp. 65-93).

van Gerven, M., & Bohte, S. (2018). Editorial: Artificial Neural Networks as Models of Neural Information Processing. *Artificial Neural Networks as Models of Neural Information Processing*, 5.

Safavian, S. R., & Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics*, *21*(3), 660-674.

Fong, P. K., & Weber-Jahnke, J. H. (2012). Privacy preserving decision tree learning using unrealized data sets. *IEEE Transactions on knowledge and Data Engineering*, *24*(2), 353-364.

Karaolis, M. A., Moutiris, J. A., Hadjipanayi, D., & Pattichis, C. S. (2010). Assessment of the risk factors of coronary heart events based on data mining with decision trees. *IEEE Transactions on information technology in biomedicine*, *14*(3), 559-566.

Vapnik, V., Guyon, I. and T. H.-M. Learn, and undefined 1995. Support vector machines.

statweb.stanford.edu.

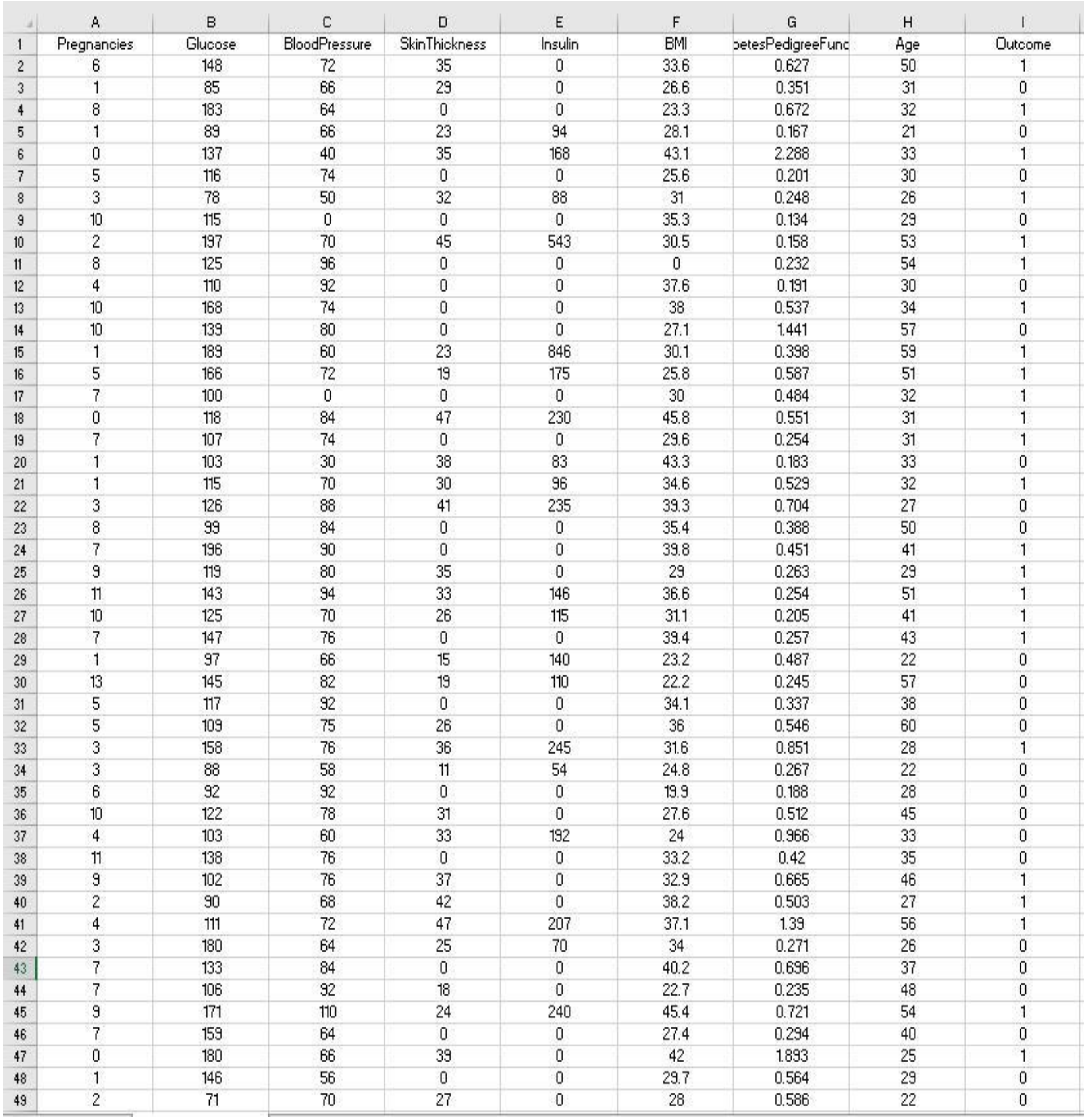
Chervonenkis, A. Y. (2013). Early history of support vector machines. In *Empirical Inference* (pp. 13-20). Springer, Berlin, Heidelberg.

Ramezankhani, A., Pournik, O., Shahrabi, J., Khalili, D., Azizi, F., & Hadaegh, F. (2014). Applying decision tree for identification of a low risk population for type 2 diabetes. Tehran Lipid and Glucose Study. *Diabetes research and clinical practice*, *105*(3), 391-398.

Meng, X. H., Huang, Y. X., Rao, D. P., Zhang, Q., & Liu, Q. (2013). Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. *The Kaohsiung journal of medical sciences*, *29*(2), 93-99.

# APENDIX A

## Diabetes Data set



| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Pregnancies | Glucose | BloodPressure | SkinThickness | Insulin | BMI | betesPedigreeFunc | Age | Outcome |
| 2 | 6 | 148 | 72 | 35 | 0 | 33.6 | 0.627 | 50 | 1 |
| 3 | 1 | 85 | 66 | 29 | 0 | 26.6 | 0.351 | 31 | 0 |
| 4 | 8 | 183 | 64 | 0 | 0 | 23.3 | 0.672 | 32 | 1 |
| 5 | 1 | 89 | 66 | 23 | 94 | 28.1 | 0.167 | 21 | 0 |
| 6 | 0 | 137 | 40 | 35 | 168 | 43.1 | 2.288 | 33 | 1 |
| 7 | 5 | 116 | 74 | 0 | 0 | 25.6 | 0.201 | 30 | 0 |
| 8 | 3 | 78 | 50 | 32 | 88 | 31 | 0.248 | 26 | 1 |
| 9 | 10 | 115 | 0 | 0 | 0 | 35.3 | 0.134 | 29 | 0 |
| 10 | 2 | 197 | 70 | 45 | 543 | 30.5 | 0.158 | 53 | 1 |
| 11 | 8 | 125 | 96 | 0 | 0 | 0 | 0.232 | 54 | 1 |
| 12 | 4 | 110 | 92 | 0 | 0 | 37.6 | 0.191 | 30 | 0 |
| 13 | 10 | 168 | 74 | 0 | 0 | 38 | 0.537 | 34 | 1 |
| 14 | 10 | 139 | 80 | 0 | 0 | 27.1 | 1.441 | 57 | 0 |
| 15 | 1 | 189 | 60 | 23 | 846 | 30.1 | 0.398 | 59 | 1 |
| 16 | 5 | 166 | 72 | 19 | 175 | 25.8 | 0.587 | 51 | 1 |
| 17 | 7 | 100 | 0 | 0 | 0 | 30 | 0.484 | 32 | 1 |
| 18 | 0 | 118 | 84 | 47 | 230 | 45.8 | 0.551 | 31 | 1 |
| 19 | 7 | 107 | 74 | 0 | 0 | 29.6 | 0.254 | 31 | 1 |
| 20 | 1 | 103 | 30 | 38 | 83 | 43.3 | 0.183 | 33 | 0 |
| 21 | 1 | 115 | 70 | 30 | 96 | 34.6 | 0.529 | 32 | 1 |
| 22 | 3 | 126 | 88 | 41 | 235 | 39.3 | 0.704 | 27 | 0 |
| 23 | 8 | 99 | 84 | 0 | 0 | 35.4 | 0.388 | 50 | 0 |
| 24 | 7 | 196 | 90 | 0 | 0 | 39.8 | 0.451 | 41 | 1 |
| 25 | 9 | 119 | 80 | 35 | 0 | 29 | 0.263 | 29 | 1 |
| 26 | 11 | 143 | 94 | 33 | 146 | 36.6 | 0.254 | 51 | 1 |
| 27 | 10 | 125 | 70 | 26 | 115 | 31.1 | 0.205 | 41 | 1 |
| 28 | 7 | 147 | 76 | 0 | 0 | 39.4 | 0.257 | 43 | 1 |
| 29 | 1 | 97 | 66 | 15 | 140 | 23.2 | 0.487 | 22 | 0 |
| 30 | 13 | 145 | 82 | 19 | 110 | 22.2 | 0.245 | 57 | 0 |
| 31 | 5 | 117 | 92 | 0 | 0 | 34.1 | 0.337 | 38 | 0 |
| 32 | 5 | 109 | 75 | 26 | 0 | 36 | 0.546 | 60 | 0 |
| 33 | 3 | 158 | 76 | 36 | 245 | 31.6 | 0.851 | 28 | 1 |
| 34 | 3 | 88 | 58 | 11 | 54 | 24.8 | 0.267 | 22 | 0 |
| 35 | 6 | 92 | 92 | 0 | 0 | 19.9 | 0.188 | 28 | 0 |
| 36 | 10 | 122 | 78 | 31 | 0 | 27.6 | 0.512 | 45 | 0 |
| 37 | 4 | 103 | 60 | 33 | 192 | 24 | 0.966 | 33 | 0 |
| 38 | 11 | 138 | 76 | 0 | 0 | 33.2 | 0.42 | 35 | 0 |
| 39 | 9 | 102 | 76 | 37 | 0 | 32.9 | 0.665 | 46 | 1 |
| 40 | 2 | 90 | 68 | 42 | 0 | 38.2 | 0.503 | 27 | 1 |
| 41 | 4 | 111 | 72 | 47 | 207 | 37.1 | 1.39 | 56 | 1 |
| 42 | 3 | 180 | 64 | 25 | 70 | 34 | 0.271 | 26 | 0 |
| 43 | 7 | 133 | 84 | 0 | 0 | 40.2 | 0.696 | 37 | 0 |
| 44 | 7 | 106 | 92 | 18 | 0 | 22.7 | 0.235 | 48 | 0 |
| 45 | 9 | 171 | 110 | 24 | 240 | 45.4 | 0.721 | 54 | 1 |
| 46 | 7 | 159 | 64 | 0 | 0 | 27.4 | 0.294 | 40 | 0 |
| 47 | 0 | 180 | 66 | 39 | 0 | 42 | 1.893 | 25 | 1 |
| 48 | 1 | 146 | 56 | 0 | 0 | 29.7 | 0.564 | 29 | 0 |
| 49 | 2 | 71 | 70 | 27 | 0 | 28 | 0.586 | 22 | 0 |

# APENDIX B

## Research Questions

Is there any issue on various data mining techniques for diabetes prediction?