DIABETES PREDICTION USING MACHINE LEARNING

BTech Project Report

Submitted to The Dr Y.S.R ANU College of Engineering and Technology,
Acharya Nagarjuna University in partial fulfillment of
Requirement for the award of Degree

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

SUBMITTED
By

G. JAHNAVI	**(Y19CS3215)**
D. PRAVEEN	**(Y19CS3210)**
S. VAMSI KRISHNA	**(L20CS3275)**

Under the Guidance
of

P. CHARAN, M.Tech(Ph.D),MBA.
Assistant Professor, Dept of CSE, Dr Y.S.R ANUCET

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
Dr Y.S.R ANU COLLEGE OF ENGINEERING & TECHNOLOGY
ACHARYA NAGARJUNA UNIVERSITY
NAGARJUNA NAGAR -522510, GUNTUR, A.P., INDIA
2019-2023

Dr Y.S.R ANU COLLEGE OF ENGINEERING & TECHNOLOGY

ACHARYA NAGARJUNA UNIVERSITY

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the project entitled **"DIABETES PREDICTION USING MACHINE LEARNING"** is a Bonafide record of the project work done by **G. Jahnavi (Y19CS3215), D. Praveen(Y19CS3210),** and **S. Vamsi Krishna(L20CS3275),** under my supervision and guidance, in partial fulfillment of the requirements for the award of Degree in Department of Computer Science and Engineering from Dr Y.S.R ANU College of Engineering & Technology, Guntur for the academic year 2022-23.

…………………………..

P. Charan
Assistant Professor, Dept. of C.S.E

…………………………..

Dr. V. Balaji
Coordinator, Dept of C.S.E

External Examiner

DECLARATION

We hereby declare that the project entitled, **"DIABETES PREDICTION USING MACHINE LEARNING"** was carried out and written by me under the guidance of **P. CHARAN**, Assistant Professor, Department of Computer Science and Engineering, Dr Y.S.R ANU College of Engineering & Technology, Acharya Nagarjuna University. This work has not been previously formed the basis for the award of any degree or diploma or certificate nor has been submitted elsewhere for the award of any degree or diploma.

Place: **ANUCET** Student Signature

Date: Student 1:

Student 2:

Student 3:

ACKNOWLEDGEMENTS

In our opinion writing acknowledgement is a difficult but most beautiful part of the project report ending. It is difficult to express the heartfelt accomplishments and the gratitude for those who really or virtually supported for making the work successful and memorable.

We would like to convey our sincere gratitude and admiration to our Guide **P. Charan**, Assistant Professor, Dr Y.S.R ANU College of Engineering & Technology, Acharya Nagarjuna University (ANU), Guntur, for his excellent support and valuable guidance throughout our study. His marvelous advice motivated us to work on this, and make it successful. His supportive nature and thought-provoking ideas made this work possible.

We are very grateful to **Dr. V. Balaji,** Coordinator, Dept. of CSE, Dr Y.S.R ANU College of Engineering & Technology, Acharya Nagarjuna University, for his keen interest, counseling, constant encouragement, personal affection and valuable suggestions in completion of this research work.

We would like to express our thanks to **Prof. P. Siddaiah**, Principal, Dr Y.S.R ANU College of Engineering & Technology, Acharya Nagarjuna University, Guntur, for guidance, mentorship, and valuable insights that have been instrumental in the successful completion of this work

We would like to express our sincere appreciation and gratitude to **Prof E. Sreenivasa Reddy**, Dean Faculty of Engineering, Dr Y.S.R ANU College of Engineering & Technology, Acharya Nagarjuna University, Guntur, for his unwavering support and encouragement throughout the duration of this project.

We are deeply indebted to Dr. **U. SATISH KUMAR** Project Coordinator, for Conducting reviews and giving constructive criticism throughout the project duration.

We are grateful to the **Authorities** of Dr Y.S.R ANU College of Engineering & Technology, Acharya Nagarjuna University, enabling us to pursue our work.

We are very thankful to all **Teaching** and **Non-teaching staff** of Department of Electronics & Communication Engineering, Dr Y.S.R ANU College of Engineering & Technology, Acharya Nagarjuna University, Guntur, for their encouragement, co-operation and assistance during the work period.

We would like to express our grateful thanks to **our parents** our whole beloved family for their love, unwavering support and encouragement throughout the time of our study.

Finally, we thank one and all, who helped us directly or indirectly for completing this work.

With gratitude

G. Jahnavi	**(Y19CS3215)**
D. Praveen	**(Y19CS3210)**
S. Vamsi Krishna	**(L20CS3275)**

ABSTRACT

Diabetes is a chronic disease with the potential to cause a worldwide health care crisis. According to International Diabetes Federation 382 million people are living with diabetes across the whole world. Diabetes is an illness caused because of high glucose level in a human body. Diabetes should not be ignored if it is untreated then Diabetes may cause some major issues in a person like: heart related problems, kidney problem, blood pressure, eye damage and it can also affect other organs of human body. To achieve this goal this project work we will do early prediction of Diabetes in a human body or a patient for a higher accuracy through applying, Various Machine Learning Techniques. Machine learning techniques Provide better result for prediction by constructing models from datasets collected from patients. In this work we will use Machine Learning Classification and ensemble techniques on a dataset to predict diabetes. Which is K-Nearest Neighbor (KNN), Logistic Regression (LR), Decision Tree (DT), Support Vector Machine (SVM) and Random Forest (RF). The accuracy is different for every model when compared to other models. The Project work gives the accurate or higher accuracy model shows that the model is capable of predicting diabetes effectively. Our Result shows that Random Forest achieved higher accuracy compared to other machine learning techniques. The aim of this project is to develop a system which can perform early prediction of diabetes for a patient with a higher accuracy by combining the results of different machine learning techniques.

Keywords: Diabetes, Machine Learning, Prediction, Dataset, Logistic Regression (LR), Decision Tree, Support Vector Machine (SVM), Random Forest

G. Jahnavi (Y19CS3215)
D. Praveen (Y19CS3210)
S. Vamsi Krishna (L20CS3275)

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS AND ABBREVIATIONS

ML – Machine learning

DT - Decision Tree

KNN - K-nearest Neighbor

LR - Logistic Regression

RF- Random Forest

SVM - Support Vector Machine

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

Diabetes is chronic diseases in the world. Diabetes caused because of obesity or high blood glucose level, and so forth. It affects the hormone insulin, resulting in abnormal metabolism of carbohydrates and improves level of sugar in the blood. Diabetes occurs when body does not make enough insulin. According to (WHO) World Health Organization about 422 million people suffering from diabetes particularly from low- or idle-income countries. And this could be increased to 490 billion up to the year of 2030. However, prevalence of diabetes is found among various Countries like Canada, China, and India etc. Population of India is now more than 100 million so the actual number of diabetics in India is 40 million.

Diabetes is not only a disease but also a creator of different kinds of diseases like heart attack, blindness, kidney diseases, etc. The normal identifying process is that patients need to visit a diagnostic center, consult their doctor, and sit tight for a day or more to get their reports. Moreover, every time they want to get their diagnosis report, they have to waste their money in vain. Diabetes Mellitus (DM) is defined as a group of metabolic disorders mainly caused by abnormal insulin secretion and/or action. Insulin deficiency results in elevated blood glucose levels (hyperglycemia) and impaired metabolism of carbohydrates, fat and proteins. DM can be divided into several distinct types. However, there are two major clinical types, type 1 diabetes (T1D) and type 2 diabetes (T2D), according to the etiopathology of the disorder. T2D appears to be the most common form of diabetes (90% of all diabetic patients), mainly characterized by insulin resistance. The main causes of T2D include lifestyle, physical activity, dietary habits and heredity, whereas T1D is thought to be due to autoimmunological destruction of the Langerhans islets hosting pancreatic-β cells. T1D affects almost 10% of all diabetic patients worldwide, with 10% of them ultimately developing idiopathic diabetes. The symptoms of DM include polyuria, polydipsia, and significant weight loss among others. Diagnosis depends on blood glucose levels (fasting plasma glucose = 7.0 mmol/L.

Diabetes is major cause of death in the world. Early prediction of disease like diabetes can be controlled and save the human life. To accomplish this, this work explores prediction of diabetes by taking various attributes related to diabetes disease. For this purpose, we use the Kaggle Indian Diabetes Dataset, we apply various Machine Learning classification and ensemble Techniques to predict diabetes. Machine Learning Is a method that is used to train computers or machines explicitly. Various Machine Learning Techniques provide efficient result to collect Knowledge by building various classification and ensemble models from collected dataset. Such collected data can be useful to predict diabetes. Various techniques of Machine Learning can capable to do prediction, however it’s tough to choose best technique. Thus, for this purpose we apply popular classification and ensemble methods on dataset for prediction

1.2 AIM OF THE PROJECT

The main aim of this project is to develop a system which can perform early prediction of diabetes for a patient with a higher accuracy by combining the results of different machine learning techniques. we will explore how Random Forest classifier can be used for diabetes predictions.

1.3 SIGNS & SYPTOMS OF DIABETES

The classic symptoms of untreated diabetes are unintended weight loss, polyuria (increased urination). polydipsia (increased thirst), and polyphagia (increased hunger). Symptoms may develop rapidly (weeks or months) in type 1 diabetes, while they usually develop much more slowly and may be subtle or absent in type 2 diabetes.

Several other signs and symptoms can mark the onset of diabetes although they are not specific to the disease. In addition to the known symptoms listed above. they include blurred vision, headache, fatigue, slow healing of cuts, and itchy skin.

Prolonged high blood glucose can cause glucose absorption in the lens of the eye. which leads to changes in its shape, resulting in vision changes. Long-term vision loss can also be caused by diabetic retinopathy. A number of skin rashes that can occur in diabetes are collectively known as diabetic dermadromes.

1.4 COMPLICATIONS OF DIABETES

The major long-term complications relate to damage to blood vessels. Diabetes doubles the risk of cardiovascular disease and about 75% of deaths in people with diabetes are due to coronary artery disease. Other macrovascular diseases include stroke, and peripheral artery disease.

The primary complications of diabetes due to damage in small blood vessels include damage to the eyes, kidneys, and nerves. Damage to the eyes, known as diabetic retinopathy, is caused by damage to the blood vessels in the retina of the eye, and can result in gradual vision loss and eventual blindness. Diabetes also increases the risk of having glaucoma, cataracts, and other eye problems.

1.5 TYPES OF DIABETES

1.5.1 TYPE 1 DIABETES

Type 1 diabetes is characterized by loss of the insulin-producing beta cells OT the as pancreatic islets, leading to insulin deficiency. This type can be further classified immune-mediated or idiopathic. The majority of type 1 diabetes is of a mediated nature, in which at immune beta cells and cell-mediated autoimmune attack leads to the loss or thus insulin. It causes North America and approximately 10% of diabetes mellitus cases in healthy weight when Europe. Most affected people are otherwise healthy and of an onset occurs. Sensitivity and

usually responsiveness to insulin are diabetes" normal, especially in the early stages. Although it has been called "juvenile due to the frequent onset in children, the majority of individuals living with type 1 diabetes are now adults.

"Brittle" diabetes, also known as unstable diabetes or labile diabetes, is a term that was traditionally used to describe the dramatic and recurrent Wings in glucose levels, often occurring for no apparent reason in insulin-dependent diabetes. This term, however, has no biologic basis and should not be used. 44) Still, type 1 diabetes can be accompanied by irregular and unpredictable high blood sugar levels, and the potential for diabetic ketoacidosis or serious low blood sugar levels.

Type 1 diabetes is partly inherited, with multiple genes, including certain HLA genotypes, known to influence the risk of diabetes. In genetically susceptible people, the onset of diabetes can be triggered by one or more environmental factors, such as a viral infection or diet. Several viruses have been implicated, but to date there is no stringent evidence to support this hypothesis in humans. Among dietary factors, data suggest that gliadin (a protein present in gluten) may play a role in the development of type 1 diabetes, but the mechanism is not fully understood.

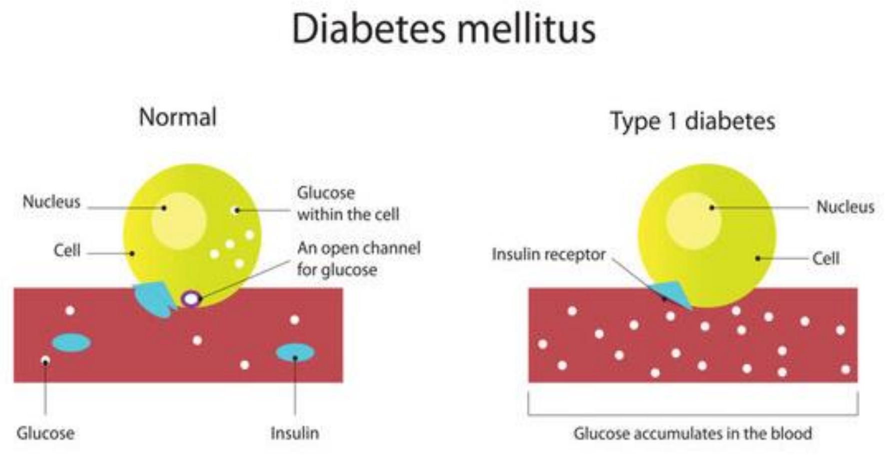


Figure 1.1 TYPE! DIABETES

Type 1 diabetes can occur at any age, and a significant proportion is diagnosed during adulthood. Latent autoimmune diabetes of adults (LADA) is the diagnostic term applied when type 1 diabetes develops in adults; it has a slower onset than the same Condition in children. Given this difference, some use the unofficial term "type 1.5 diabetes" for this condition. Adults with LADA are frequently initially misdiagnosed as having type 2 diabetes, based on age rather than a cause.

1.5.2 TYPE 2 DIABETES

Type 2 diabetes is characterized by insulin resistance, which may be combined with relatively reduced insulin secretion. The defective responsiveness of body tissues to insulin is believed to involve the insulin receptor. However, the specific defects are not known. Diabetes mellitus cases due to a known defect are classified separately. Type 2 diabetes is the most common type of diabetes mellitus. Many people with type 2 diabetes have evidence of prediabetes (impaired fasting glucose and/or impaired glucose tolerance) before meeting the criteria for type 2 diabetes. The progression of prediabetes to overt type 2 diabetes can be slowed or reversed by lifestyle changes or medications that improve insulin sensitivity or reduce the liver's glucose production. Type 2 diabetes is primarily due to lifestyle factors and genetics.

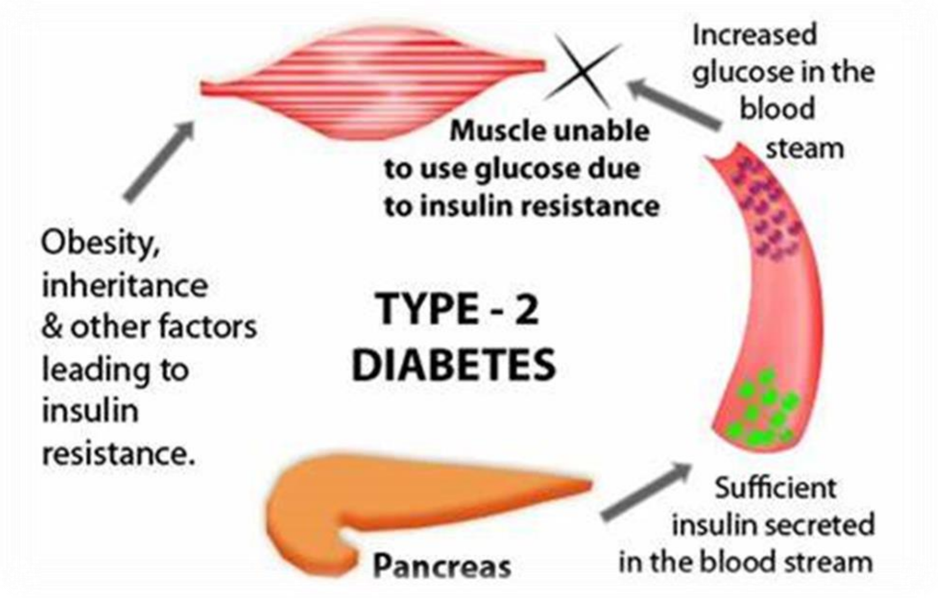

Increased glucose in the blood steam
Muscle unable to use glucose due to insulin resistance
Obesity, inheritance & other factors leading to insulin resistance.
TYPE - 2 DIABETES
Pancreas
Sufficient insulin secreted in the blood stream

Figure 1.2: TYPE2 DIABETES

1.5.3 GESTATIONAL DIABETES

Gestational diabetes resembles type 2 diabetes in several respects, involving a combination of relatively inadequate insulin secretion and responsiveness. It occurs in about 2-10% of all pregnancies and may improve or disappear after delivery. It is recommended that all pregnant women get tested starting around 24-28 weeks gestation. It is most often diagnosed in the second or third trimester because of the increase in insulin-antagonist hormone levels that occurs at this time. However, after pregnancy approximately 5-10% of women with gestational' diabetes are found to have another form of diabetes, most commonly type 2. Gestational diabetes is fully treatable, but requires careful medical supervision throughout the pregnancy. Management may include dietary changes, blood glucose monitoring, and in some cases, insulin may be required.

1.6 ARCHITECTURE

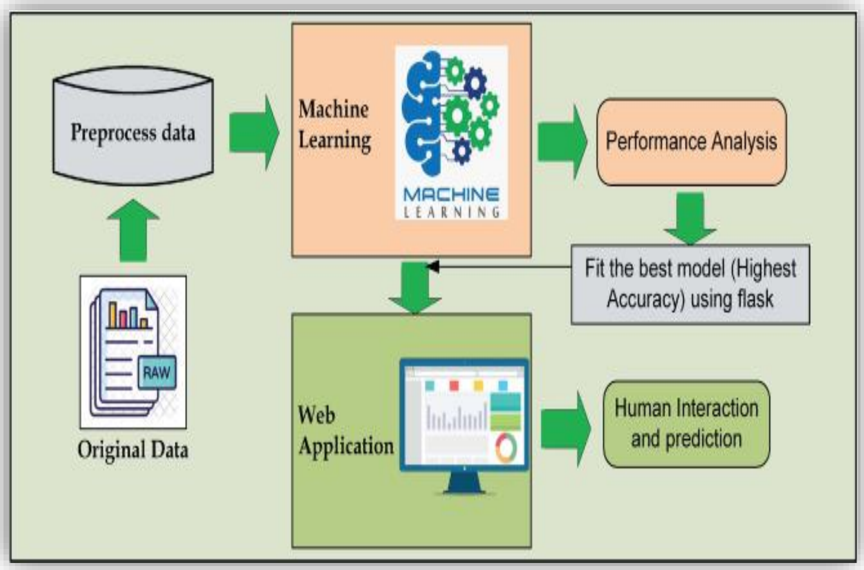


Figure 1.3: PROCEDURE

The system will detect whether the person has diabetes or not using the dataset. If diabetes is detected the classification value will be 1 and if not, the value will be 0. We will be using 4 machine learning model in order to detect the disease. The models used are logistic regression, KNN, Random Forest. has diabetes or not on the basis of using a trained dataset The system architecture will help in the future too diagnosis diabetes. It will also predict whether.

- Survey of diabetes death rates among different category of people:

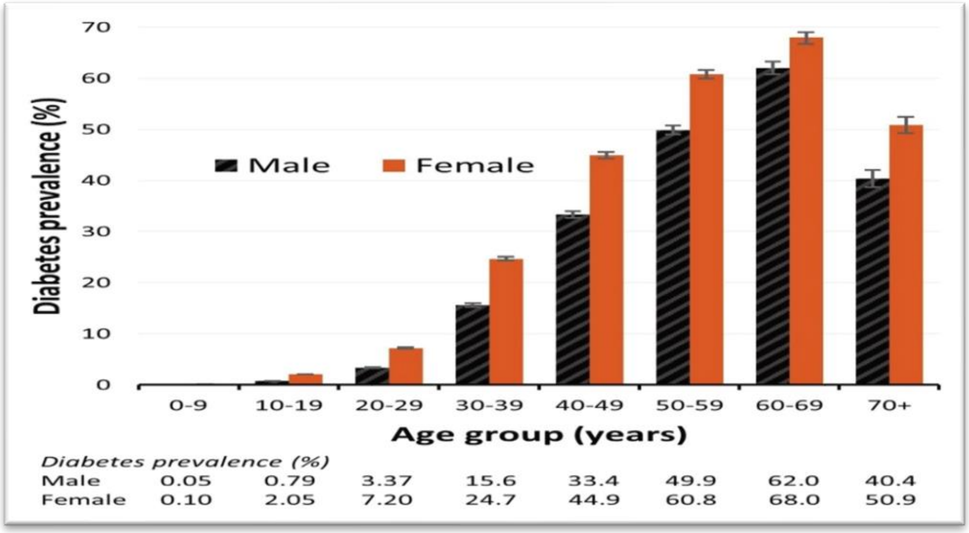


Figure 1.4: DIABETES SURVEY

1.7 CAUSES

- Exact cause of this condition is not known.
- Type 1 Diabetes IS a result of an cells that produce insulin auto-immune destruction of pancreatic islet cell that produce insulin.
- Genetics also plays a role in type 1 diabetes.
- Pancreatic the diseases can also be the cause of type 1 diabetes
- The risk factors include: Age - The children between the ages 4 and 7 years old and also in 10 years and 14 years are at high-risk .
- Genetics Family history

1.8 DIAGNOSIS

Diabetes mellitus by is diagnosed with a test for the glucose content in the blo0d, demonstrating any one of the following:

- Fasting plasma glucose level 7.0 mm/L (126 mg/dL), For this test, blood is taken had after a period of fasting, i.e., in the morning before breakfast, after the patient sufficient time to fast overnight.
- Plasma glucose > 11.1 mmol/L (200 mg/dL) two hours after a 75-gram oral glucose load as in a glucose tolerance test (OGTT).
- Symptoms of high blood sugar and plasma glucose > 11.1 mmol/L (200 mg/dL) either while fasting or not fasting.

WHO diabetes diagnostic criteria						
Condition	**2-hours glucose**		**Fasting Glucose**		**HbAt$_c$**	
Unit	Mmol/L	Mg/dL	Mmol/L	Mg/dl	Mmol/mol	DCCT %
Normal	<7.8	<140	<6.1	<110	<42	<6-0
Impaired glucose glycaemia	<7.8	<140	6.1-7.0	110-125	42-46	6.0-6.4
Diabetes mellitus	>=11.1	>=200	>=7.0	>=126	>=48	>=6.5

TABLE 1.1: WHO DIABETES DIAGNOSTIC CRITERIA

A positive result, in the absence of unequivocal high blood sugar, should be vomited by a repeat of any of the above methods on different day. It is preferable O measure a fasting allulose level because of the ease of measurement and the Considerable time commitment of formal glucose tolerance testing, which takes two us to complete and offers no prognostic advantage over the fasting test. According ne current definition two fasting glucose measures above 7.0 mmol/L.

- Glycated hemoglobin (HbA1c) 48 mmol/mol (2 6.5 DCCT %).

1.8.1 STEPS OF DETECTION

1.Data Collection

2.Defining Data

3.pre-processiong

4.Building model

5.Analysis

6.Results

1.9 PREVENTION

There is no known preventive measure for type 1 diabetes. Type 2 diabetes which accounts for 85-90% of all cases worldwide-can often be prevented or delayed by maintaining a normal body weight, engaging in physical activity, and eating a healthy diet. Higher levels of physical activity (more than 90 minutes per day) reduce the risk of diabetes by 28%o. Dietary changes known to be effective in helping to prevent diabetes include maintaining a diet rich in whole grains and fiber, and choosing good fats, such as the polyunsaturated fats found in nuts, vegetable oils, and fish. Limiting sugary beverages and eating less red meat and other sources of saturated fat can also help prevent diabetes. Tobacco smoking is also associated with an increased risk of diabetes and its complications, so smoking cessation can be an important preventive measure as well.

The relationship between type 2 diabetes and the main modifiable risk factors (excess weight, unhealthy diet, physical inactivity and tobacco use) is similar in all regions of the world. There is growing evidence that the underlying determinants of diabetes are a reflection of the major forces driving social, economic and cultural change: globalization, urbanization, population aging, and the general health policy environment.

1. Losing weight and keeping it off....
2. Following a healthy eating plan. .
3. Get regular exercise. .
4. Don't smoke. ...
5. Talk to your health care provider to see whether there is anything else you can do to delay or to prevent type 2 diabetes.

1.9.1 MANAGEMENT

Diabetes management concentrates on keeping blood sugar levels as close to normal, without causing low blood sugar. This can usually be accomplished with dietary changes, exercise. weight loss, and use of appropriate medications (insulin, oral medications).

Learning about the disease and actively participating in the treatment is portent, since complications are far less common and less severe in people who have well-managed blood sugar levels. Per the American College of Physicians, the anal of treatment is a Haik level of 7-8%o. Attention is also paid to other health include smoking, high blood pressure, metabolic syndrome obesity, and lack of These nullar exercise. Specialized footwear is widely used to reduce the risk of ulcers in at irk diabetic feet although evidence for the efficacy of this remains equivocal.

1.9.2 LIFESTYLE

People with diabetes can benefit from treatment, dietary education about the disease and changes, and exercise, with the goal of keeping both long-term blood glucose short-term and levels within associated higher acceptable bounds. In addition, given the risks of recommended to Control blood cardiovascular pressure.

Weight the loss can prevent progression from prediabetes to diabetes type 2, risk of cardiovascular disease, or result in a partial with diabetes. No remission in people dietary patterns, single dietary pattern is best for all people with diabetes. Healthy such as the Mediterranean diet, are often low-carbohydrate diet, or DASH diet, recommended, although evidence does not support one over the others.

1.9.3 MEDICATIONS

Glucose controlment-diabetic medication Most medications used to treat diabetes act by lowering blood sugar levels through different mechanisms. There is broad consensus that when people with diabetes maintain tight glucose control keeping the glucose levels in their blood within normal ranges they experience fewer complications, such as kidney problems or eye problems. There is however debate as to whether this is appropriate and cost effective for people later in life in whom the risk of hypoglycemia may be more significant.

There are a number of different cl333asses of anti-diabetic medications. Type 1 diabetes requires treatment with insulin, ideally using a "basal bolus" regimen that most closely matches normal insulin release: long-acting insulin for the basal rate and Short-acting insulin with meals. Type 2 diabetes is generally taken with medication that Is taken by mouth (e.g., metformin) although some eventually require injectable treatment with insulin or GLP-1 agonists.

1.9.4 SUPPORT

In countries take place using mainly a general practitioner system, such as the United Kingdom, outside hospitals, with hospital-based specialist care used only in case of complications, difficult blood her sugar control, or research projects. In circumstances, general practitioners and specialists share care in a team approach. Home telehealth support can be an effective management technique.

CHAPTER 2

LITERATURE REVIEW

K.VijiyaKumar et al. [11] proposed Random Forest algorithm for the Prediction of diabetes develop a system which can perform early prediction of diabetes for a patient with a higher accuracy by using Random Forest algorithm in machine learning technique. The proposed model gives the best results for diabetic prediction and the result showed that the prediction system is capable of predicting the diabetes disease effectively, efficiently and most importantly, instantly.

Noons Nnameka et al. [13] presented predicting diabetes onset: an ensemble supervised learning approach they used five widely used classifiers are employed for the ensembles and a meta-classifier is used to aggregate their outputs. The results are presented and compared with similar studies that used the same dataset within the literature. It is shown that by using the proposed method, diabetes onset prediction can be done with higher accuracy. Teja's N. Joshi et al. [12] presented Diabetes Prediction Using Machine Learning Techniques aims to predict diabetes via three different supervised machine learning methods including: SVM, Logistic regression, ANN. This project proposes an effective technique for earlier detection of the diabetes disease.

Dheeraj Shetty et al. [15] proposed diabetes disease prediction using data mining assemble Intelligent Diabetes Disease Prediction System that gives analysis of diabetes malady utilizing diabetes patient's database. In this system, they propose the use of algorithms like Bayesian and KNN (K-Nearest Neighbor) to apply on diabetes patient's database and analyze them by taking various attributes of diabetes for prediction of diabetes disease.

The profound physiological model for T1DM blood glucose prediction was given by Munoz-Organero et al. [25]. &e carbohydrate differential equations and the physiologic model for insulin absorption were designed based on the long-term memory (LSTM) cell-specific recurring neural network (RNN). &e findings revealed that the RMSE values for hypothetical patients were below 5 mg/dL and that, for actual patients, the RMSE values were below 10 mg/dL. In comparison to state-of-the-art approaches, the device has done higher. &e dataset had, nonetheless, the number of days for each participant to report the results. It was also impossible for the model to estimate potential BG values.

Mahbub [26] proposes a rigorous diabetes prediction voting strategy focused on conventional approaches. Here, eleven popular machine learning algorithms such as the Naive Bayes, the KNN, Support Vector Machine, Random Forest, Logistic Regression, Gradient Boosting, Ada Boosting, and several more have been used for early detection of diabetes. &e 11 algorithms were analyzed using different criteria, such as performance, precision, f-mean sacrament, and reminder. After cross-validation and hyper tuning, the best three machine learning algorithms were designed and employed in the Ensemble Voting Classifiers. &e experimental results confirmed the Pima Indians Diabetes Database's good results, which are around 86 percent correct. &e precision was comparatively good, but the estimation outcomes were not 100 percent, and the preparation period was high.

Zhou et al. [27] studied the diabetes prediction using an initial deep neural network. First, data has been split into preparation and evaluation data. Next, people were categorized depending on their medical problems. &e tests proved the efficacy and efficiency of the method. &e highest accuracy data collection for diabetes was 94.02174%, and the most reliable was that for Pima Indians at 99.4112%.

Muhammad Azeem Sarwar et al. [10] proposed study on prediction of diabetes using machine learning algorithms in healthcare they applied six different machine learning algorithms Performance and accuracy of the applied algorithms is discussed and compared. Comparison of the different machine learning techniques used in this study reveals which algorithm is best suited for prediction of diabetes.

Diabetes Prediction is becoming the area of interest for researchers in order to train the program to identify the patient are diabetic or not by applying proper classifier on the dataset. Based on previous research work, it has been observed that the classification process is not much proved. Hence a system is required as Diabetes Prediction is important area in computers, to handle the issues identified based on previous research.

CHAPTER 3

MACHINE LEARNING CONCEPTS

Machine Learning is a system of computer algorithms that can learn from through self-improvement without being explicitly coded by a Machine programmer. learning is a part of artificial Intelligence which combines data with statistical tools to predict an output which can be used to make actionable insights. The breakthrough comes with the idea that a machine can singularly learn from the data (i.e., mining example) and Bayesian to produce accurate results. Machine learning is closely related to data Uses a predictive modeling. The machine receives data as input and algorithm to formulate answers. A typical machine learning tasks are to provide a recommendation, for those who have a Netflix account, all recommendations of movies or series are based on the user's historical data. Tech companies are using unsupervised learning to improve the user experience with personalizing recommendation. Machine learning is also used for a variety of tasks like fraud detection, predictive maintenance, portfolio optimization, automatize task and so on.

3.1 SUPERVISED LEARNING

In supervised learning, the system must "learn" inductively a function called target function, which is an expression of a model describing the data. The objective function is used to predict the value of a variable, called dependent variable or output variable, from a set of variables, called independent variables or input variables or characteristics or features. The set of possible input values of the function, i.e., its domain, are called instances. Each case is described by a set of characteristics (attributes or features). A subset of all cases, for which the output variable value is known, is called training data or examples In supervised learning, there are two kinds of learning. Tasks: classification and regression. Classification models try to predict distinct classes, such as e.g., blood groups, while regression models predict numerical values. Some of the most common techniques are Decision Trees (DT), Rule Learning, and Instance Based Learning (BL), such as k-Nearest Neighbors (k-NN), Genetic

Algorithms (GA**),** Artificial Neural Networks (ANN), and Support Vector Machines (SVM).

3.2 UNSUPERVISED LEARNING

In unsupervised learning, the system tries to discover the hidden structure of data or associations between variables. In that case, training data consists of instances without any corresponding labels. Association Rule Mining appeared much later than machine learning and is subject to greater influence from the research area of databases. Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning. pattern recognition, image analysis, information retrieval, bioinformatics', data compression. and computer graphics.

3.3 REINFORCEMENT LEARNING

The term Reinforcement Learning is a general term given to a family or techniques, in which the system attempts to learn through direct interaction with the environment so as to maximize some notion of cumulative reward. It is important to mention that the system has no prior knowledge about the behavior of the environment and the only way to find out is through trial and failure (trial and error). Reinforcement learning is mainly applied to autonomous independence in relation to its environment.

Classification is one of the most important decision-making techniques in many real-world problems. In this work, the main objective is to classify the data as diabetic or non-diabetic and improve the classification accuracy. For many classifications problem, the higher number of samples chosen but it doesn't lead to higher classification accuracy. In many cases, the performance of algorithm is high in the context of speed but the accuracy of data classification is low. The main objective of our model is to achieve high accuracy. Classification accuracy can be increase it we use much of the data set for training and few data sets for testing.

3.4 ARTIFICIAL NUERAL NETWORK

The artificial neural network is much similar as natural neural network of a brain. Artificial Neural networks (ANN) typically consist of multiple layers or a cube design, and the signal path traverses from front to back. Back propagation is the use of forward stimulation to reset weights on the "front" neural units and this is sometimes done in combination with training where the correct result is known. More modern networks are a bit freer flowing in terms of stimulation and inhibition with connections interacting in a much more chaotic and complex fashion. Dynamic neural networks are the most advanced, in that they dynamically can, based on rules, for new connections and even new neural units while disabling. Generally, the artificial neural network is consisting of the layers and network function, the layers of the network are including: input layer, hidden layer and output layer

EXAMPLE

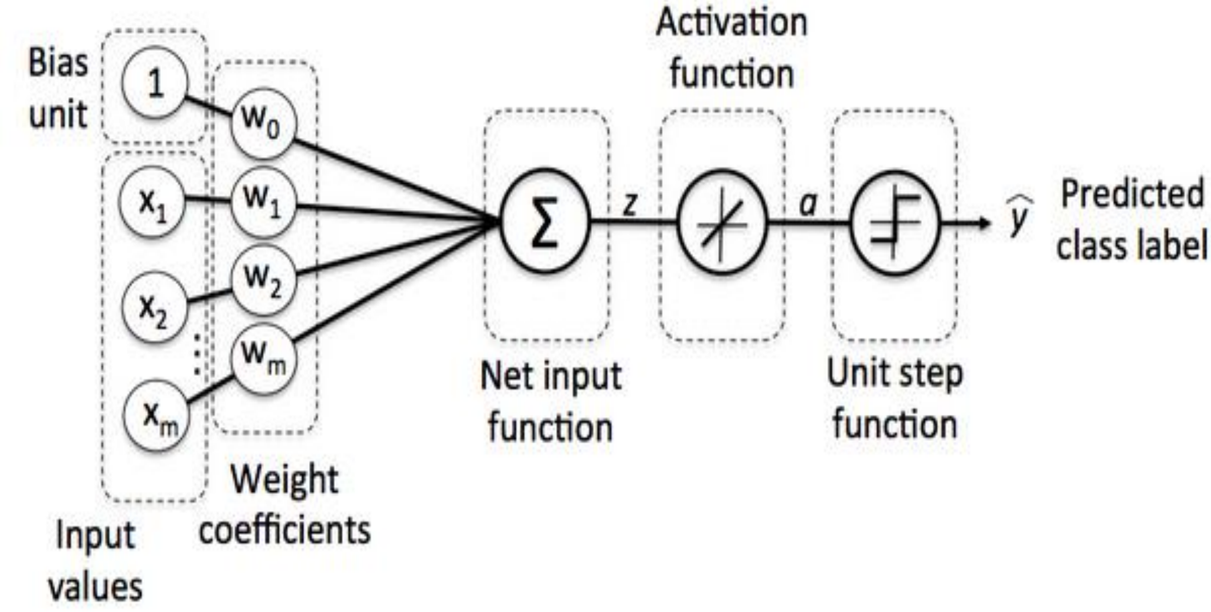


Figure 3.1: SINGLE LAYER NEURAL NETWORK

The input neurons define all the input attribute Generally, the artificial neural network is consisting of the layers and network function, the layers of the network are including: input layer, hidden layer and output layer Generally, the artificial neural network is consisting of the layers and network function, the layers of the network are including: input layer, hidden layer and output layer values for the data mining model. In our work, the number of neurons is 7.

CHAPTER 4

PYTHON LIBRARIES

4.1 PANDAS

Pandas is a Python package providing last, flexible, and expressive data secures designed to make working with "relational" or "labeled" data both easy and illusive, it aims to be the fundamental high-level bulling block for doing practical, real world data analysis in Python. Additionally, it has the broader goal of becoming the most powerful and flexible open-source data analysis/manipulation tool available in any language, l is already well on its way toward this (goal Pandas is well suited for many different kinds of data:

- Tabular data with heterogeneously-typed columns, as in an SQL table or Excel spreadsheet
- Ordered and unorder6d (not necessarily fixed-frequency) time series data.
- Arbitrary matrix data (hormogon60usly typed or heterogeneous) with row and column labels
- Any other form of observational / statistical data sots the data need not be labeled at all to be placed into a panda's data structure

The two primary data structures of pandas, Series (1dimensional) and Data frame (2 dimensional), handle the vast majority of typical use cases in finance, statistics, social sciences, and many areas of engineering. For R users, Data Frame provides everything that R's data frame provides and much more, pandas is built on top of NumPy and is intended to integrate Well within a scientific computing environment with many other3rd party libraries

Here are just a few of the things that pandas does well

- Easy handling of missing data (represented as NaN) in floating point as well as non-floating-point data

- Size mutability: columns can be inserted and deleted from Data Frame and higher dimensional objects
- Automatic and explicit data alignment: objects can be explicitly aligned to a set of labels, or the user can simply ignore the labels and let Series, Data Frame, etc. automatically align the data for you in computations
- Powerful, flexible group by functionality to perform split-apply-combine operations on data sets, for both aggregating and transforming data
- Make it easy to convert ragged, differently-indexed data in other Python and NumPy data structures into Data Frame objects
- Intelligent label-based slicing, fancy indexing, and sub setting of large data sets
- Intuitive merging and joining data sets
- Flexible reshaping and pivoting of data sets
- Hierarchical labeling of axes (possible to have multiple labels per tick)
- Robust lOS tools for loading data from flat files (CSV and delimited), Excel files databases, and saving / loading data from the ultrafast HDF5 format
- Time series-specific functionality: date range generation and frequency Conversion, moving window statistics, date shifting, and lagging.

many of these principles are here to address the shortcomings frequently experienced Using other languages / scientific research environments.

- pandas are fast. Many of the low-level algorithmic bits have been extensively tweaked in Python code. However, as with anything else sacrifices performance. So, if your generalization usually focus on one feature for you may application you be able to create a faster specialized tool.
- pandas are a dependency of stats models, making it a statistical computing important part of the ecosystem in Python.
- pandas has been used extensively in production in financial applications.

The best way to think applications. about the pandas data structures is as lower dimensional data. flexible containers for example, Data Frame is a container for Series, and Series is a container for scalars. We would like to be able to insert and these containers in a remove object from dictionary-like fashion. which take

Also, we would like sensible default behaviors for the common API functions into account the typical orientation of time series and sets. When cross-sectional data using the N-dimensional array (ND arrays) to store 2- and data, a burden is 3-dimensional writing placed on the user to consider the orientation of the data set when functions; axes are considered more or less equivalent (except when C- or Fortran-contiguousness matters for performance). In pandas, the axes are intended to lend more semantic meaning to the data; i.e., for a particular data set, there is likely to be a "right way to orient the data. The goal, then, is to reduce the amount of mental effort required to code up data transformations in downstream functions. For example, with tabular data (Data Frame) it is more semantically helpful to think of the index (the rows) and the columns rather than axis 0 and axis 1. Iterating through the columns of the Data Frame thus results in more readable code.

If you're interested in contributing, please visit the contributing guide. pandas is a NumFOCUS sponsored project. This will help ensure the success of the development of pandas as a world-class open-source project and makes it possible to donate to the project. The governance process that pandas project has used Informally since its inception in 2008 is formalized in Project Governance documents. The documents clarify how decisions are made and how the various elements of our community interact, including the relationship between open-source collaborative development and work that may be funded by for-profit or non-profit entities

4.2 SCIKIT-LEARN

Scikit-learn was initially developed by David Cournapeau as a Google summer of code project in 2007. Later Mathieu Burcher joined the project and started to use it as a part of his thesis work. In 2010 INRIA got involved and the first public release (v0.1 beta) was published in late January 2010.The project now has more than 30 active contributors and has had paid sponsorship from INRIA, Google, Tiny clues and the Python Software Foundation.

Scikit-learn provides a range of supervised and unsupervised learning algorithms via a consistent interface in Python. It is licensed under a permissive simplified BSD license and is distributed under many Linux distributions, encouraging academic and commercial use. The library is built upon the SciPy (Scientific Python) that must be installed before you can use scikit-learn.

This stack that includes:

- **NumPy**: Base n-dimensional array package
- **SciPy**: Fundamental library for scientific computing
- **Matplotlib**: Comprehensive 2D/3D plotting
- **IPython**: Enhanced interactive console
- **Sympy**: Symbolic mathematics
- **Pandas**: Data structures and analysis

Extensions or modules for SciPy care conventionally named SciKits. As such, the module provides learning algorithms and is named scikit-learn. The vision for the library is a level of robustness and support required for use in production systems. This means a deep focus on concerns such as ease of use, code quality, collaboration, documentation and performance.

Although the interface is Python, c-libraries are leverage for performance such as Numpy for arrays and matrix operations, LAPACK, LibsVM and the careful use of Python. The library is focused on modeling data. It is not focused on loading. manipulating and summarizing data. For these features, refer to NumPy and Pandas.

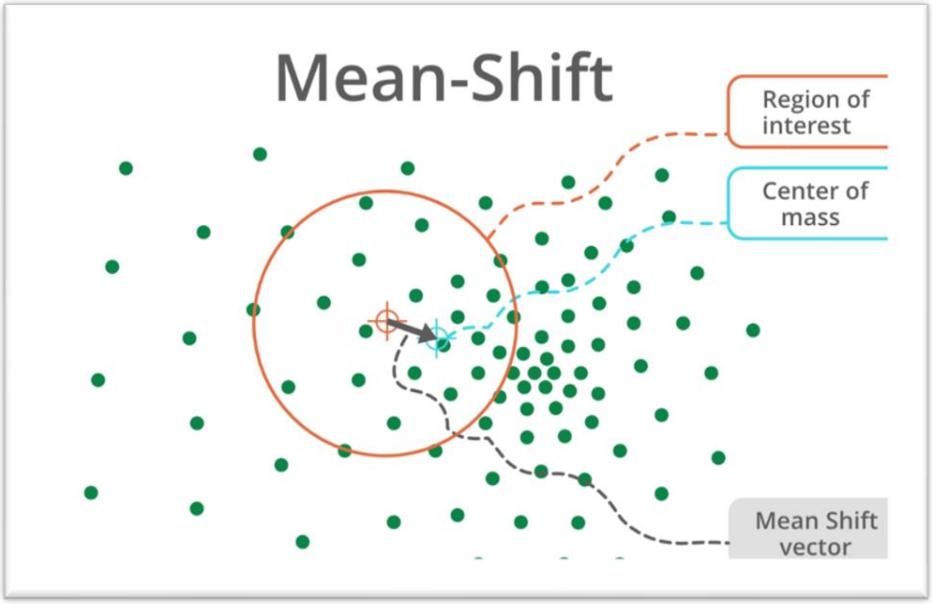


Figure 4.1: MEAN SHIFT CLUSTERING ALGORITHM

Some popular groups of models provided by scikit-learn include:

- **Clustering**: tor grouping unlabeled data such as KMeans.
- **Cross Validation**: for estimating the performance of supervised models on unseen data.
- **Datasets**: for test datasets and for generating datasets with specific properties for investigating model behavior.
- **Dimensionality Reduction**: for reducing the number of attributes in data for Summarization, visualization and feature selection such as Principal component analysis.
- **Ensemble methods**: for combining the predictions of multiple supervised models.
- **Feature extraction**: for defining attributes in image and text data.
- **Feature selection**: for identifying meaningful attributes from which to create Supervised models.
- **Parameter Tuning**: for getting the most out of supervised models.

- **Manifold Learning**: For summarizing and depicting complex multi-dimensional data.
- **Supervised Models:** a vast array not limited to generalized linear models, discriminate analysis, naive Bayes, lazy methods, neural networks, support vector machines and decision trees.

This dataset is provided as an example dataset with the library and is loaded. The classifier is fit on the data and then predictions are made on the training data. The scikit-learn testimonials page lists India, Medley, wise.io , Evernote, Telecom Aristech and Weber as users of the library. If this is a small indication of companies that have presented on their use, then there are very likely tens to hundreds of larger organizations using the library. It has good test coverage and managed releases and is suitable for prototype and production projects alike. If you are interested in learning more, checkout the Scikit-Learn homepage that includes documentation and related resources.

This post, you will discover the challenge of model selection for machine learning After reading this post, you will know:

- Model selection IS the process of choosing one among many candidate models for a predictive modeling problem.
- There beyond may model be many competing concerns when performing model selection resources. Performance, such as complexity, maintainability, and available
- The and two resampling main classes methods. of model selection techniques are probabilistic measures
- Model selection is the process of selecting one final machine learning model from among a collection of candidate machine learning models for a training dataset.

Model selection is a process that can be applied both across different types of models (e.g., logistic regression, SVM, KNN, etc.) and across models of the same type configured with different model hyper parameters (e.g., different kernels in an SVM).

When we have a variety of models of different complexity (e.g., linear or logistic regression models with different degree polynomials, or KNN classifiers with different values of K), how should we pick the right one?

Machine Learning: A Probabilistic Perspective, 2012. For example, we may have a dataset for which we are interested in developing a classification or regression predictive model. We do not know beforehand as to which model will perform best on this problem, as it is unknowable. Therefore, we fit and evaluate a suite of different models on the problem

Model selection is the process of choosing one of the models as the final model that addresses the problem. Model selection is different from model assessment.

Firstly, we need to get over the idea of a "best" model. All models have some predictive error given the statistical noise in the data, the incompleteness of the data Sample, and the limitations of each different model type. Therefore, the notion of a Pallet or best model is not useful. Instead, we must seek a model that is "good enough"., The project stakeholders may have specific requirements, such as ability and limited model complexity. As such, a model that has lower skill but is simpler and easier to understand may be preferred. Alternately, if model skill is need to get over the idea of a "best" model. All models have some predictive error, given the statistical noise in the data, the incompleteness of the data sample, and the limitations of each different model type. Therefore, the notion of a Pllect or best model is not useful. Instead, we must seek a model that is "good enough". The project stakeholders may have specific requirements, such as ability and limited model complexity. As such, a model that has lower skill but is simpler and easier to understand may be preferred. Alternately, if model skill is prized above all other concerns, then the ability of the model to perform well on out of-sample data will be preferred regardless of the computational complexity involved. Therefore, a "good enough" model may refer to many things and is specific to your project, such as:

- A model that meets the requirements and constraints of project stakeholders.
- A model that is sufficiently skillful given the time and resources available.
- A model that is skillful as compared to naive models.
- A model that is skillful relative to other tested models.
- A model that is skillful relative to the state-of-the-art.
- Next, we must consider what is being selected.

For example, we are not selecting a fit model, as all models will be discarded. This is because once we choose a model, we will fit a new final model on all available data and start using it to make predictions. Therefore, are we choosing among algorithms used to fit the models on the training dataset. Some algorithms require specialized data preparation in order to best expose the structure of the problem to the learning algorithm. Therefore, we must go one step further and consider model selection as the process of selecting among model development pipelines.

Each pipeline may take in the same raw training dataset and outputs a model that can be evaluated in the same manner but may require different or overlapping computational steps,

such as:

- Data filtering.
- Data transformation.
- Feature selection.
- Feature engineering.
- And more...
- The closer you look at the challenge of model selection, the more nuance you will discover.

Now that we are familiar with some considerations involved in model selection. let's review some common methods for selecting a model.

The best approach to model selection requires "'sufficient" data, which may be nearly infinite depending on the complexity of the problem. In this ideal situation, we Would split the data into training, validation, and test sets, then fit candidate models on the training set, evaluate and select them on the validation set, and report the performance of the final model on the test set..

Historically various 'information criteria' have been proposed that attempt to correct for the bias of maximum likelihood by the addition of a penalty term to compensate for the over-fitting of more complex models. Pattern Recognition and Machine Learning, 2006. A model with fewer parameters is less complex, and because of this, is preferred because it is likely to generalize better on average. Four commonly used probabilistic model selection measures include:

- Akaike Information Criterion (AIC).
- Bayesian Information Criterion (BIC).
- Minimum Description Length (MDL).
- Structural Risk Minimization (SRM).

4.3 RESAMPLING METHODS

Resampling methods seek to estimate the performance of a model (or more precisely, the model development process) on out-of-sample data. This is achieved by splitting the training dataset into sub train and test sets, fitting a model on the sub train set, and evaluating it on the test set.

This process may then be repeated multiple times and the mean performance across each trial is reported. It is a type of Monte Carlo estimate of model performance on out-of-sample data, although each trial is not strictly independent as depending on

The resampling method chosen, the same data may appear multiple times in different training datasets, or test datasets. Three common resampling methods selection methods include:

- Random train/test splits.
- Cross-Validation
- Bootstrap.

Most of the time probabilistic measures (described in the previous section) are not available, therefore resampling methods are used. By far the most popular is the cross-validation family of methods that includes many subtypes. Probably the simplest and most widely used method for estimating prediction error is cross-validation.

The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2017. An example is the widely used k-fold cross-validation that splits the training dataset into k folds where each example appears in a test set only once.

4.4 CROSS VALIDATION

Cross-validation (CV) is used to estimate the test error associated with a model to evaluate its performance or to select the appropriate level of flexibility. Evaluating a model's performance is usually defined as model assessment, and model selection is used for selecting the level of flexibility. The terminology is widely used in the field of data science.

CHAPTER 5

METHODOLOGIES

5.1 SUPPORT VECTOR MACHINE (SVM)

The Support Vector Machine (SVM) was first proposed by Vapnik, and SVM is a set of related supervised learning method always used in medical diagnosis for classification and empirical regression. SVM simultaneously minimize the classification error and maximize the geometric margin. So SVM is called Maximum Margin Classifiers. SVM is a general algorithm based on guaranteed risk bounds of statistical learning theory, so called structural risk minimization principle. SVMs can efficiently perform nonlinear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

The kernel trick allows constructing the classifier without explicitly knowing the feature space. Recently, SVM has attracted a high degree of interest in the machine learning research community. Several recent studies have reported that the SVM (support vector machines) generally are capable of delivering higher performance in terms of classification accuracy than the other data classification algorithms. SVM is a technique suitable for binary classification tasks, so we choose SVM to predict the is SVM is well known for its discriminative power for diabetes. The reason classification, especially in the cases where a large number of features are involved, and in our cases where the dimension of the feature of 7

5.1.1 HYPER PLANES AND SUPPORT VECTORS IN SVM

I. HYPER PLANE

There can be multiple lines/decision boundaries to segregate the classes in n dimensional space, but we need to find out the best decision boundary that helps to classify the data points. This best boundary is known as the hyperplane of SVM.

The dimensions of the hyperplane depend on the features present in the dataset. which means if there are 2 features (as shown in image), then hyperplane will be a straight line. And if there are 3 features, then hyperplane will be a 2-dimension plane.

The working of the SVM algorithm can be understood by using an example. Suppose we have a dataset that has two tags (green and blue), and the dataset has two features x1 and X2. We want a classifier that can classify the pair(x1, x2) of coordinates in either green or blue. Consider the below image

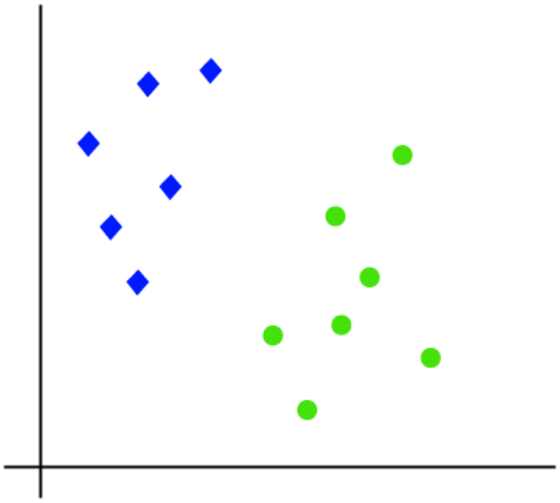


Figure 5.1: DATA SET

So, as it is 2-d space so by just using a straight line, we can easily separate these two classes. But there can be multiple lines that can separate these classes.

Consider the below image:

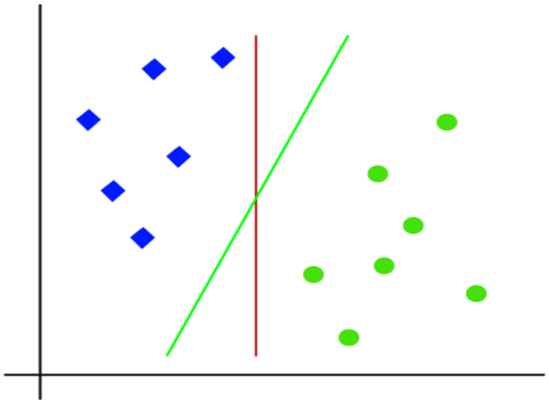


Figure 5.2: CLASSIFICATION

Hence, the SVM algorithm helps to find the best line or decision boundary; this best boundary or region is called as a **hyperplane**. SVM algorithm finds the closest point of the lines from both the classes. These points are called support vectors. The distance between the vectors and the hyperplane is called as **margin**. And the goal of SVM is to maximize this margin. The hyperplane with maximum margin is called the **optimal hyperplane**.

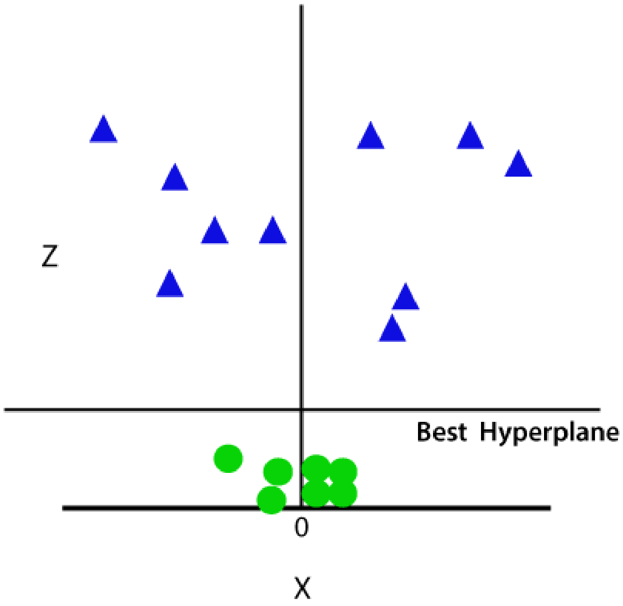


Figure 5.3: SVM CLASSIFIER

Since we are in 3-d Space, hence it is looking like a plane parallel to the x-axis. If we convert it in 2d space with z=1, then it will become as:

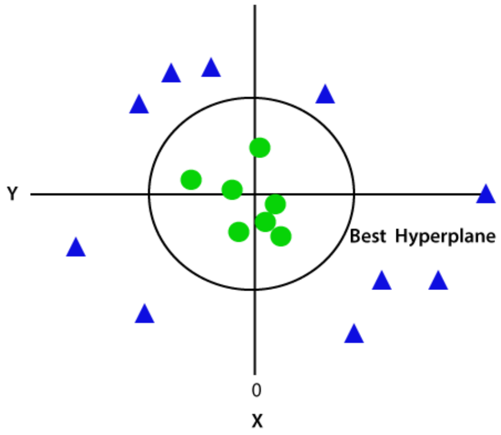


Figure 5.4: BEST HYPERPLANE

Hence, we get a circumference of radius 1 in case of non-linear data

II. SUPPORT VECTORS

The data points or vectors that are the closest to the hyperplane and which vectors support position of the hyperplane are termed as Support Vector.

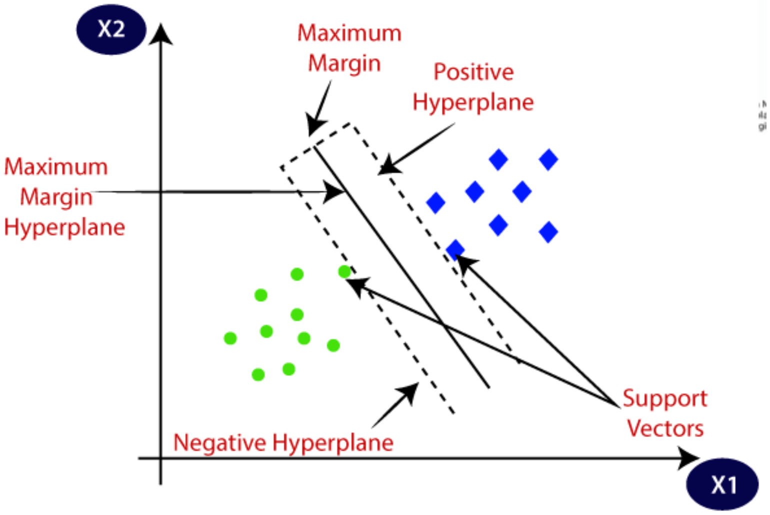


Figure 5.5: SUPPORT VECTOR MACHINE

Since these the hyperplane, hence called a Support vector. Consider the below image

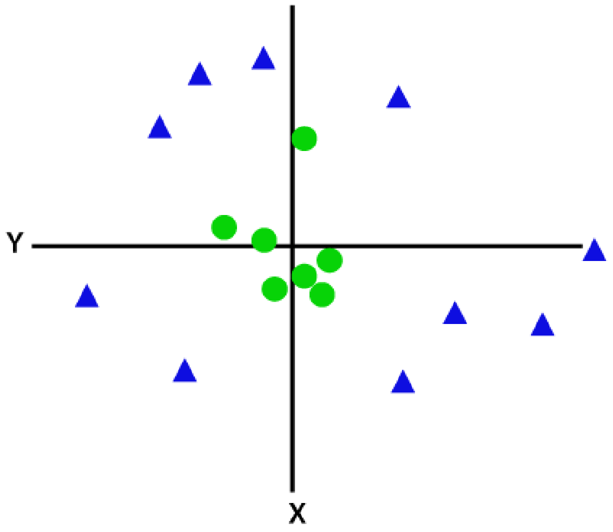


Figure 5.6: AFTER SVM CLASSIFY

So, to separate these data points, we need to add one more dimension. For linear data, we have used two dimensions' x and y, so for non-linear data, we will add a third-dimension z.

By adding the third dimension, the sample space will become as below image:

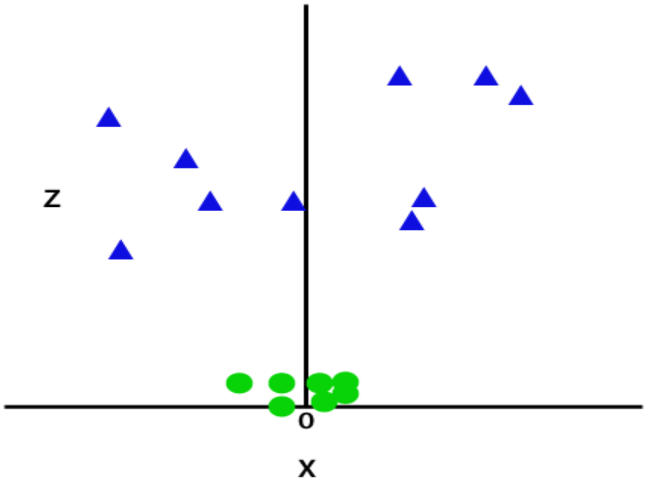


Figure 5.7: PERFECT CLASSIFICATION

It can be calculated as:

z=x? +y? X

5.1.2 TYPES OF SVM

I. LINEAR SVM

Linear SVM is used for data that are linearly separable i.e., for a dataset that can be categorized into two categories by utilizing a single straight line. Such data points are termed as linearly separable data and the classifier is used described As a Linear SVM classifier.

II. NON-LINEAR SVM

Non-Linear SVM is used for data that are non-linearly separable data i.e., a straight line cannot be used to classify the dataset. For this, we use something known as a kernel trick that sets data points in a higher dimension where they Can be separated using planes or other mathematical functions. Such data points are termed as non-linear data, and the classifier used is termed as a Non-linear SVM classifier.

III. APPLICATIONS OF SVM

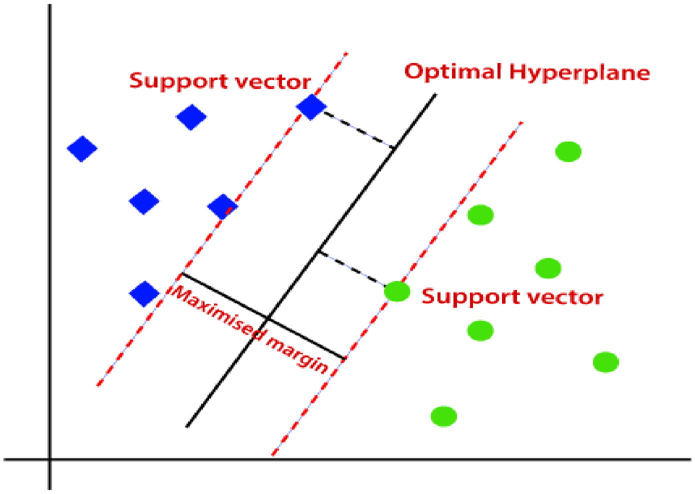


Figure 5.8: SVM

- Sentiment analysis.
- Spam Detection.
- Handwritten digit recognition.
- Image recognition challenge

5.2 LOGISTIC REGRESSION

In statistics Logistic regression is a regression model where the dependent variable is categorical, namely binary dependent variable-that is, where it can take only two values, "O" and "1", which represent outcomes such as pass/fail, win/lose, machine learning, most medical fields and social sciences. For example, the Trauma and Injury Severity Score (TRISS). which is widely used to predict mortality in injured patients, was originally developed using logistic regression. Many other medical Scales Used to assess severity of a patient have been developed using logistic regression.

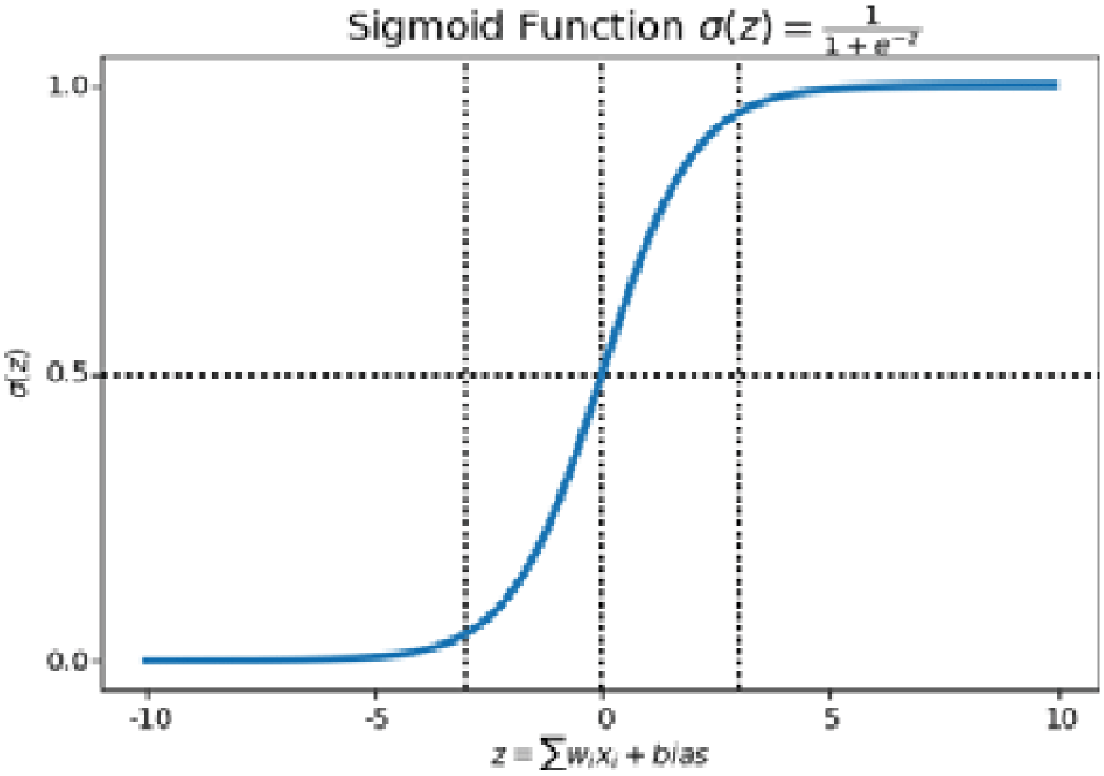


Figure5.9: LOGISTIC REGRESSION

Conditional random fields, an extension of logistic regression to sequential data, are used in natural language processing. In this paper, Logistic regression was used to predict whether a patient suffer from diabetes, based on seven observed characteristics of the patient.

In logistic regression, the target variable has two possible values like yes/no. Imagine if we represent the target variable y taking the value of "yes" as 1 and no" as 0. Then, according to the logistic model, the log-odds of y being 1 is a linear combination of one or more predictor variables. So, let's say that we have two predictor or independent variables namely x1 and x2, and let P be the probability of y being equal to 1. Then according to the logistic model

$$\mathbf{ln(p / 1 – p) = a + bx1 + cx2}$$

By exponentiation the equation, we can recover the odds:

$$\mathbf{p / 1 – p = e(a + bx1 + cx2)}$$

$$\mathbf{p = e(a + bx1 + cx2) / 1 + e(a + bx1 + cx2)}$$

$$\mathbf{p = 1 / 1 + e(a + bx1 + cx2)}$$

Which gives us the probability of y being 1. If p is closer to 0, then y=0 and when p is closer to 1 then y=1. Thus, the equation for logistic regression becomes:

$$\mathbf{y = 1 / 1 + e(a + bx1 + cx2)}$$

We can generalize this equation for n number of parameters and independent variables as follows:

$$\mathbf{y = 1 / 1 + e-(a + bx1 + cx2)}$$

5.2.1 TYPES OF LOGISTIC REGRESSION

On the basis of the categories, Logistic Regression can be classified into three types:

- **Binomial**: In binomial Logistic regression, there can be only two possible types of the dependent variables, such as 0 or 1, Pass or Fail, etc.
- **Multinomial:** In multinomial Logistic regression, there can be 3 or more possible unordered types of the dependent variable, such as "cat", "dogs", or "sheep"
- **Ordinal:** In ordinal Logistic regression, there can be 3 or more possible ordered types of dependent variables, such as "low", "Medium", or "High".

5.2.2 Sigmoid Function

- It is the logistic expression especially used in Logistic Regression.
- The sigmoid function converts any line into a curve which has discrete values like binary 0 and.
- In this session let's see how a continuous linear regression can be manipulated and converted into Classifies Logistic

$$\mathbf{P = 1 / 1 + e\text{-}(y)}$$

Were,

P represents Probability of Output class Y represents predicted output.

5.3 RANDOM FOREST CLASSIFIER

Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

The below diagram explains the working of the Random Forest algorithm:

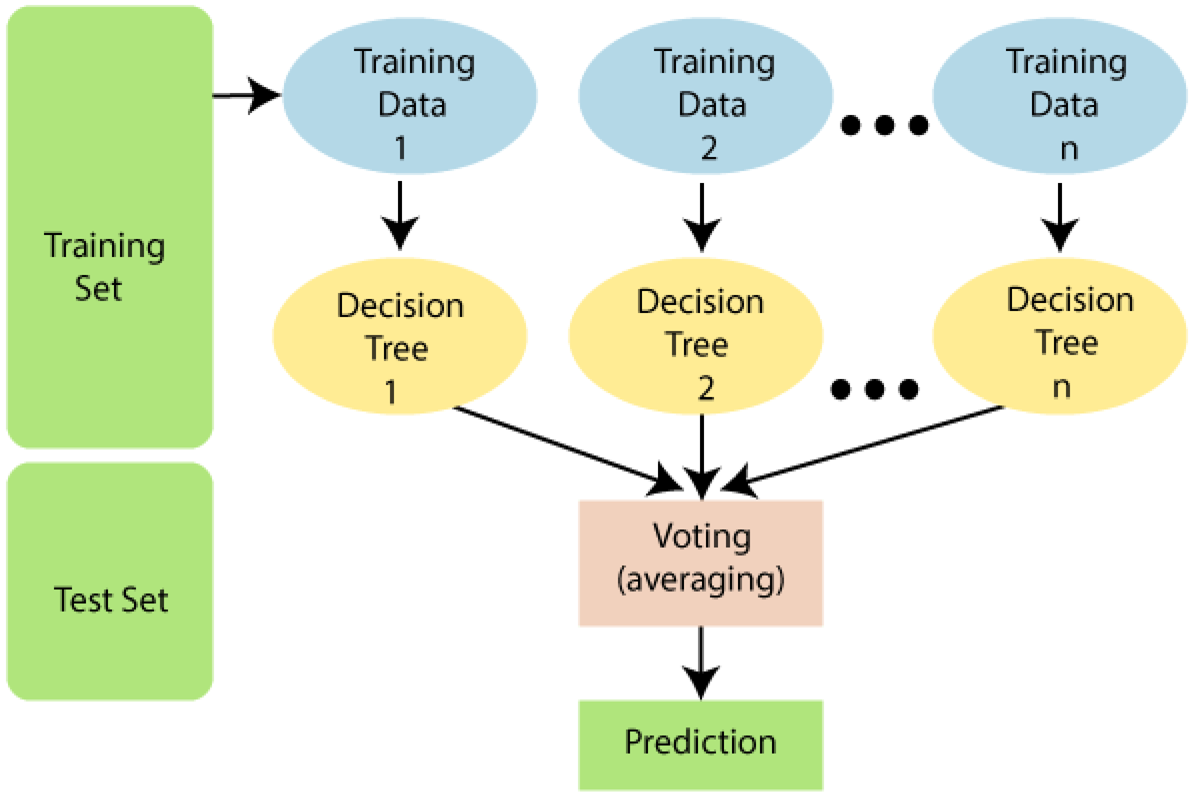


Figure 5.10: RANDOM FOREST

Below are some points that explain why we should use the Random Forest algorithm:

- It takes less training time as compared to other algorithms.
- It predicts output with high accuracy, even for the large dataset it runs efficiently.
- It can also maintain accuracy when a large proportion of data is missing.

Below The Working process can be explained in the below steps:

- **Step-1:** Select random K data points from the training set.
- **Step-2:** Build the decision trees associated with the selected data points (Subsets).
- **Step-3:** Choose the number N for decision trees that you want to build.
- **Step-4:** Repeat Step 1 & 2.
- **Step-5**: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

5.4 DECISION TREE

Decision Tree is a supervised learning method used in data mining for classification and regression methods. It is a tree that helps us in decision-making purposes. The decision tree creates classification or regression models as a tree structure It separates a data Set into Smaller subsets, and at the same time.

The Decision tree is steadily developed. The final tree is a tree with the decision nodes and A decision node has at least two branches. The leaf nodes show classification or decision. We can't accomplish more split on leaf nodes-The uppermost decision node in a tree that relates to the best predictor called the root node. Decision trees can deal with both categorical and numerical data.

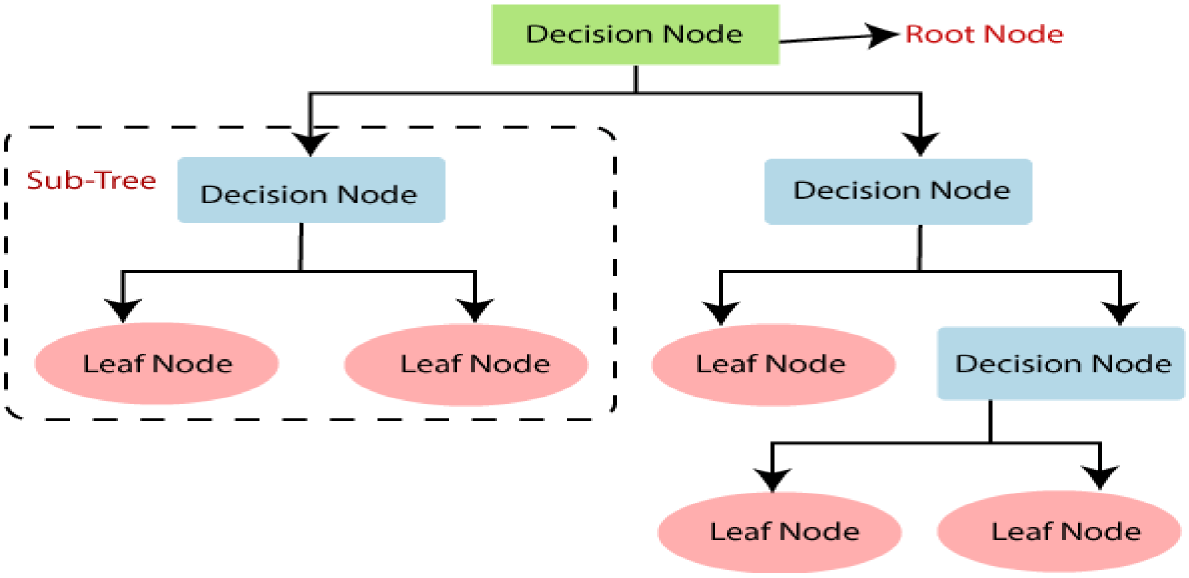


Figure 5.11: DECISION TREE

CHAPTER 6

MODULE DESCRIPTION & DATA SET

6.1 MODULE DESCRIPTION

In predictive analysis, the methods used, stands a very crucial role in the model. To get maximum accuracy, we need to select algorithms that would eventually benefit us from the beginning to the end. So before selecting the methods we need to have safe and clean data that would run easily on every model which we use and for that we need to pass our data through several possible stages to get it purely and cleanly.

Some of them are mentioned below:

1.Dataset selection stage

2.Pre-processing and transformation stage

3.Feature selection stage

4.Implementation of classification algorithms

5.Performance evaluation stage

6.1.1 DATA SELECTION STAGE

Selecting an appropriate dataset always stands a prominent role at the beginning stage. We have to collect a dataset which would be suitable to predict Diabetes disease for data analysis and along with sufficient knowledge. To implement a machine learning model, we need to have a huge amount of data – in the form of a dataset. We have obtained our dataset from the Kaggle which is a platform that provides various

6.1.2 PREPROCESSING AND TRANSFORMATION STAGE

In this stage, we need to prepare the dataset into an attribute relation file. So that in future we would use the attribute to identify the required features. After that, we need to convert the file into a binomial pattern so that we could implement various associative rules and techniques. To be a standard data format, we must remove missing values, duplicate values, and outlier records.

6.1.3 FEATURE SELECTION STAGE

For the predictive model to work efficiently and to produce accurate results we need to make sure that we have selected good features. For better results, we have to use various tools and applications to get most of the promising features in a diabetes dataset.

So, for that, we need to use several search methods and feature evaluators for this purpose. Selecting best features is important process when we prepare a large dataset for training. So, we are using the Select Kbest method which selects the features according to the k highest score. Some of the features which we have selected are: Pregnancies, BMI, AGE for better results

6.1.4 IMPLEMENTATION OF CLASSIFICATION ALGORITHMS

In this study, we have included four different classifiers that use the selected dataset including a Decision Tree, Logistics Regression, Random Forest, SVM.

6.1.5 PERFORMANCE EVALUATION STAGE

Decision tree, Logistic regression, SVM, and Random Forest are trained and tested using the identified Diabetes dataset, and the performance and evaluation of each classifier are counter measured for correct classified instances of the approved identified dataset.

Disease prediction system: The best classifier among the four is going to form an intelligent diabetes prediction system for accurate prediction

6.2 GENERAL ARCHITECTURE

The overall schema of the predictive model will be illustrated in the figure below. At the beginning of this study, we only have the diabetes dataset. After that, we sent the diabetes dataset into feature selection stage where we receive the prominent and crucial features of our data which we can in the future purposes. So, we are going use the Select Best method, which helps us to eliminate the less important part of data and reduce a training time. After getting feature we now move on to the next stage preparing the training dataset. Now we are going to run the data against classification techniques, which Logistic Regression, Random Forest, Decision Tree, SVM and are going to evaluate the performance to detect the diabetes disease.

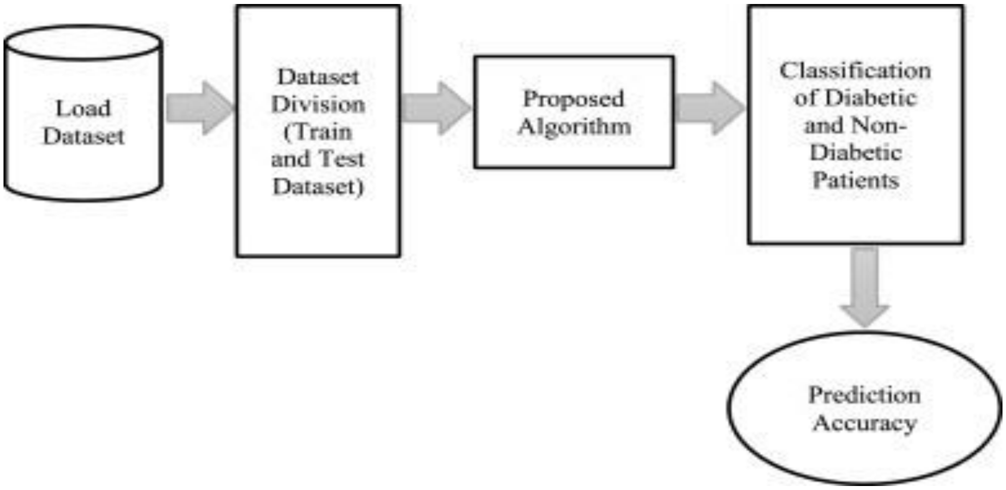

Load Dataset
Dataset Division (Train and Test Dataset)
Proposed Algorithm
Classification of Diabetic and Non-Diabetic Patients
Prediction Accuracy

Figure 6.1: ARCHITECTURE

6.3 DATA SET

6.3.1 KAGGLE DATA SET

The dataset collected is originally from Kaggle. It consists of several medical analyst variables and one target variable. The objective of the dataset is to predict whether the patient has diabetes or not. The dataset consists of several independent variables and one dependent variable, i.e., the outcome. Independent variables include the number of pregnancies the patient had their BMI, insulin level, age and so on as shown in following table:

SERIAL NO	ATTRIBUTE NAMES	DESCRIPTION
1	PREGNANCIES	NO. OF TIMES PREGNANT
2	GLUCOSE	PLASMA GLUCOSE CONCENTRATION
3	BLOOD PRESSURE	DIASTOLIC BLOOD PRESSURE
4	SKIN THICKNESS	TRICEPS SKIN FOLD THICKNESS
5	INSULIN	2-H SERUM INSULIN
6	BMI	BODY MASS INDEX
7	DPF	DIABETES PEDIGREE FUNCTION
8	AGE	AGE OF PATINET
9	OUTCOME	CLASS VARIABLE(0 OR 1)

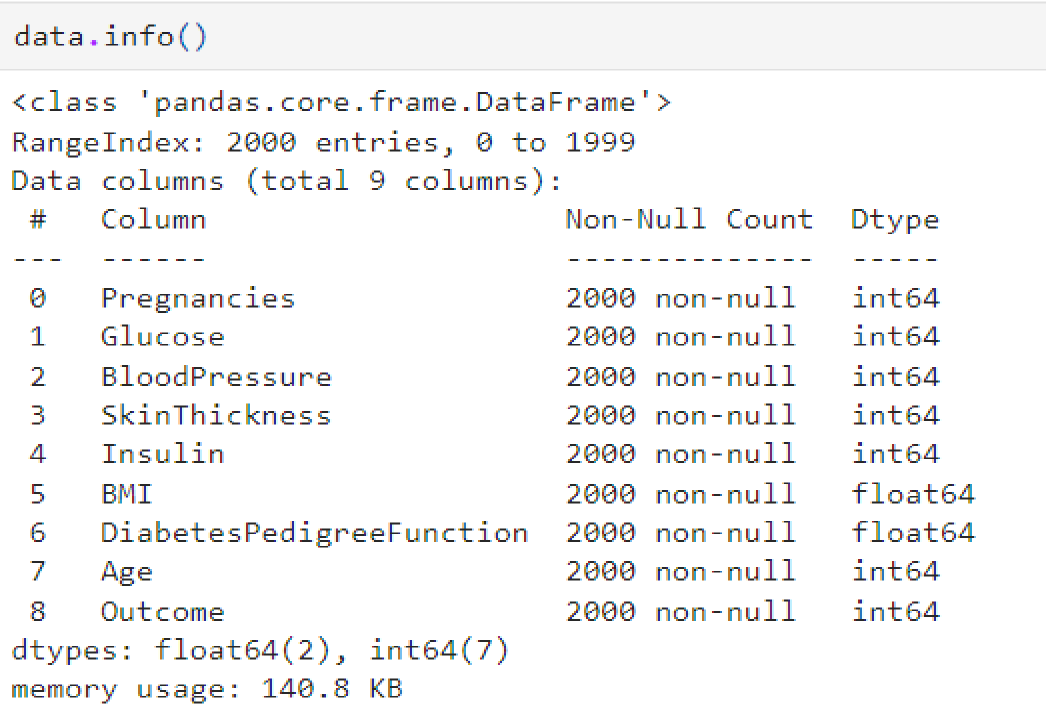
TABLE 6.1: DATASET DESCRIPTION

- The diabetes dataset consists of 2000 data points, with 9 features each.
- “outcome” is the feature we are going to predict, 0 means no diabetes and 1 means diabetes.
- There are no null values in the data.

6.3.2 CHECKING FOR MISSISNG VALUES

There are several reasons why data may be missing from a dataset in Python. Some common causes of missing data include user error, such as forgetting to fill in a field, data being lost while transferring manually from a legacy database, or programming errors. Missing data can also occur due to issues with data collection methods, such as surveys where respondents may not provide all the information requested.

Missing data can affect your algorithm and research in several ways. For example, missing values can provide a wrong idea about the data itself, causing ambiguity. Calculating an average for a column with half of the information unavailable or set to zero gives the wrong metric. Additionally, when data is unavailable, some algorithms may not work.



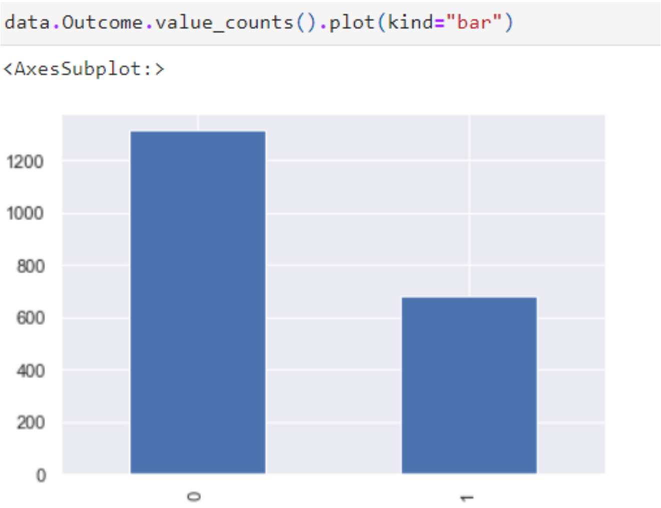
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   Pregnancies               2000 non-null   int64  
 1   Glucose                   2000 non-null   int64  
 2   BloodPressure             2000 non-null   int64  
 3   SkinThickness             2000 non-null   int64  
 4   Insulin                   2000 non-null   int64  
 5   BMI                       2000 non-null   float64
 6   DiabetesPedigreeFunction  2000 non-null   float64
 7   Age                       2000 non-null   int64  
 8   Outcome                   2000 non-null   int64  
dtypes: float64(2), int64(7)
memory usage: 140.8 KB
```

Figure 6.2: CODE FOR MISSING DATA CHECKING

6.3.3 Bar plot for outcome class

The above graph shows that the data is biased towards data points having outcome value is 0 where it means that diabetes was not present actually. The no. of non-diabetes is almost twice the number of diabetes prediction.

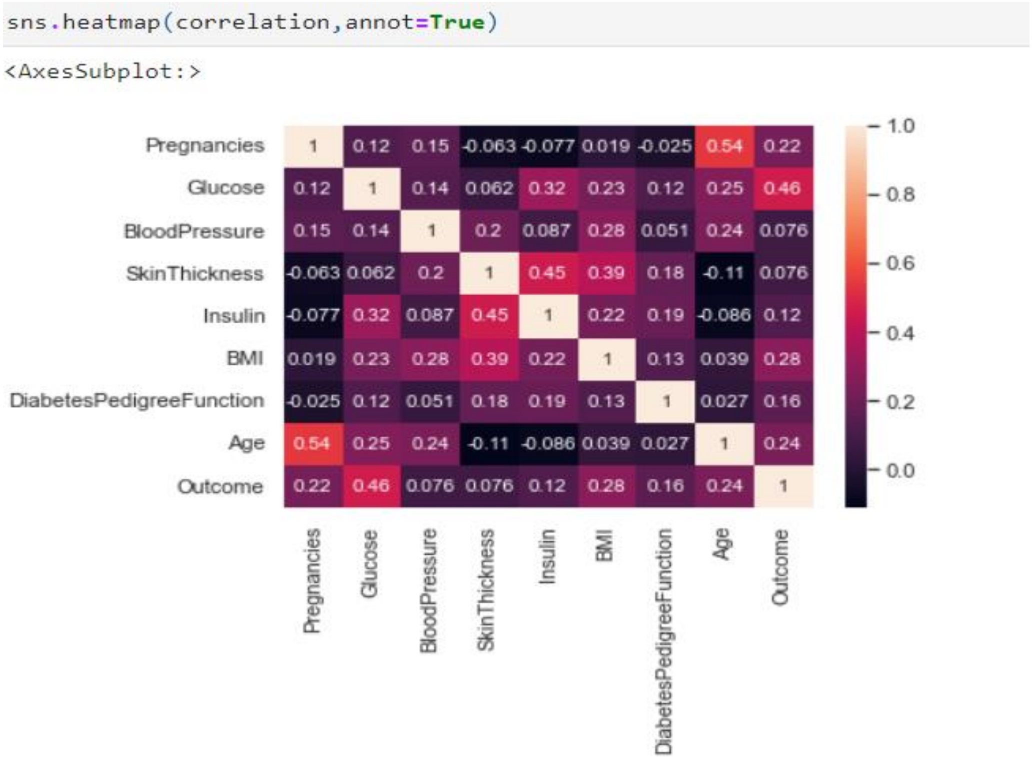


```
data.Outcome.value_counts().plot(kind="bar")
```

```
<AxesSubplot:>
```

Figure 6.3: BAR GRAPH

6.4 CORRELATION MATRIX



```
sns.heatmap(correlation,annot=True)
```

```
<AxesSubplot:>
```

Figure 6.4: CORRELATION MATRIX

6.5 Measurements

To find the efficient classifier for diabetes prediction we have applied performance matrices are confusion matrix and accuracy are discussed as follows: Confusion matrix: -

which provides output matrix with complete description performance of the model.

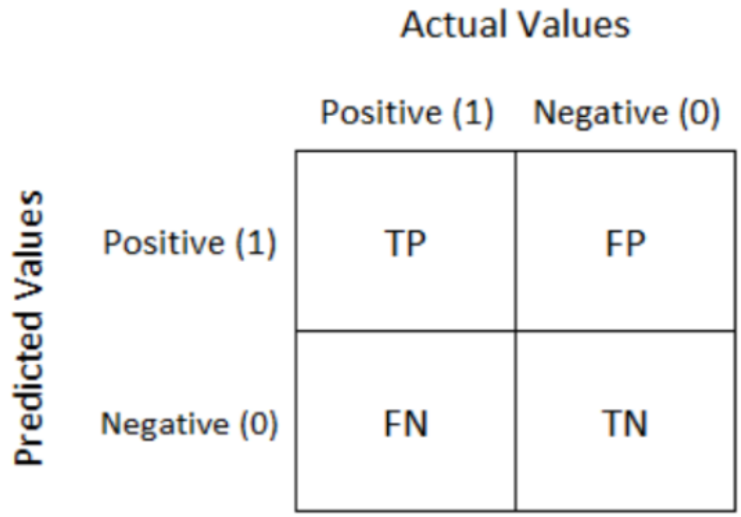
Here,

Tp: True positive

FP: False positive

TN: True negative

FN: False negative



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Figure 6.5: MEASUREMENTS

The following performance metrics are used to calculate the presentation of various algorithms.

- True positive (TP) – person has disease, and the prediction also has a positive
- True negative (TN) – person not having disease and the prediction also has a negative
- False positive (FP) – person not having disease but the prediction has a positive
- False negative (FN) – person having disease and the prediction also has a positive
- TP and TN can be used to calculate accuracy rate and the error rates can be computed using FP and FN values.
- True positive rate can be calculated as TP by a total number of persons have disease in reality.
- False positive rate can be calculated as FP by a total number of persons do not have disease in reality.
- Precision is TP/ total number of persons have prediction result is yes.
- Accuracy is the total number of correctly classified records

CHAPTER 7

IMPLEMENTATION AND TESTING

7.1 DATA COLLECTION

It includes data collection and understanding the data to study the hidden patterns and trends which helps to predict and evaluating the results. Dataset carries1405 rows i.e., total number of data and 10 columns i.e., total number of features. Features include Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, DiabetesPedigreeFunction, Age

7.2 DATA PREPROCESSING

This phase of model handles inconsistent data in order to get more accurate and precise results like in this dataset Id is inconsistent so we dropped the feature. This dataset doesn’t contain missing values. So, we imputed missing values for few selected attributes like Glucose level, Blood Pressure, Skin Thickness, BMI and Age because these attributes cannot have values zero. Then data was scaled using Standard Scaler.

7.3 FEATURE SELECTION

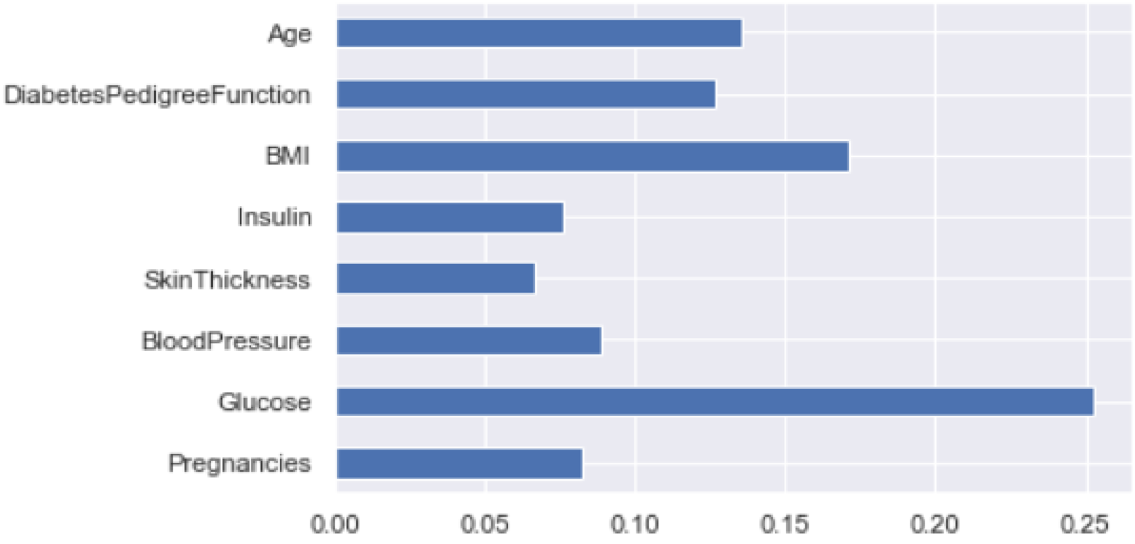


Figure 7.1: FEATURE IMPORTANCE

Pearson’s correlation method is a popular method to find the most relevant attributes/features. The correlation coefficient is calculated in this method, which correlates with the output and input attributes. The coefficient value remains in the range by between −1 and 1. The value above 0.5 and below −0.5 indicates a notable correlation, and the zero value means no correlation.

7.4 SCALING AND NORMALIZATION

We performed feature scaling by normalizing the data from 0 to 1 range, which boosted the algorithm’s calculation speed. scaling means that you're transforming your data so that it fits within a specific scale, like 0-100 or 0-1. You want to scale data when you're using methods based on measures of how far apart data points are, like support vector machines (SVM) or k-nearest neighbors (KNN)With these algorithms, a change of "1" in any numeric feature is given the same importance

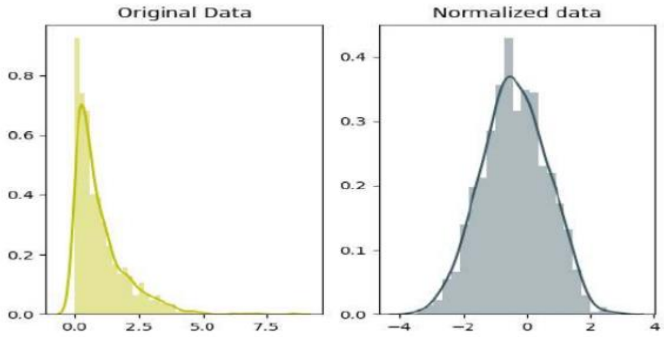


Figure 7.2: NORMALIZATION

7.5 SPLITTING THE DATA

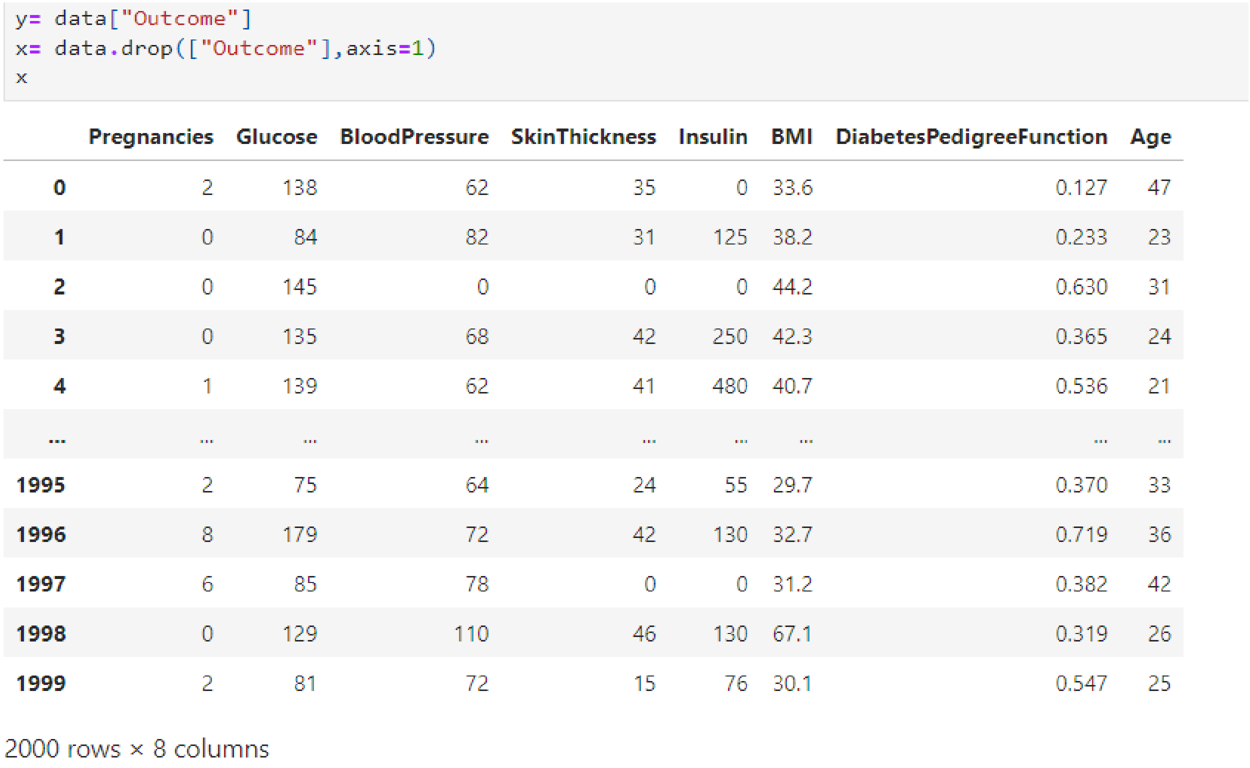
After data cleaning and pre-processing, the dataset becomes ready to train and test. In the train/split method, we split the dataset randomly into the training and testing set.

1. Independent dataset represented as x which is the feature dataset, and
2. The dependent dataset represented as y that contains the target dataset.

3. Here feature dataset and target dataset are divided into training and testing data.

4. We divide the entire dataset into 80% into training data and 20% into testing data that why we have used test size to 0.2

5. Once we do that, we have our training data and testing data ready to be deployed.



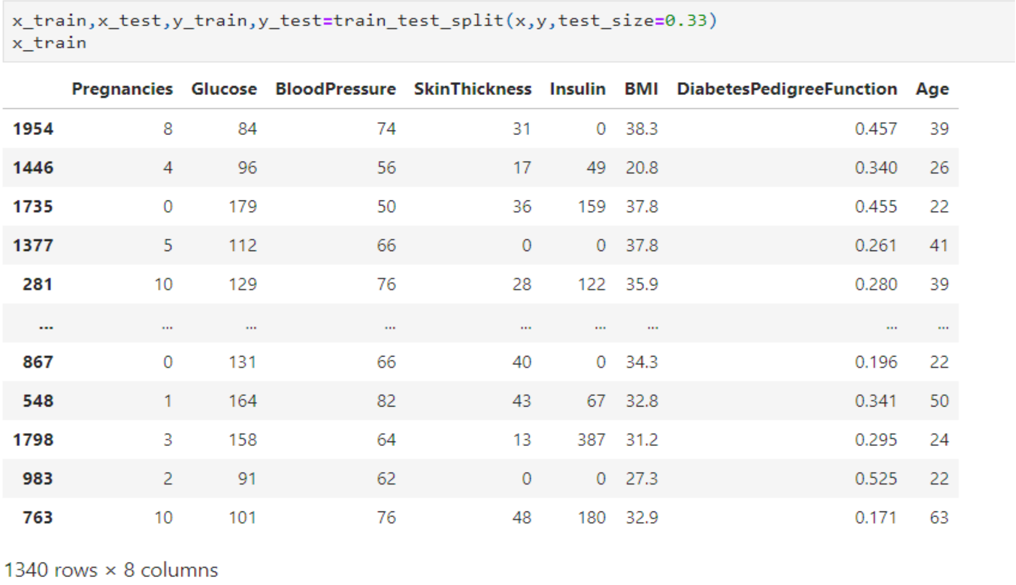
```
y= data["Outcome"]
x= data.drop(["Outcome"],axis=1)
x
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
0	2	138	62	35	0	33.6	0.127	47
1	0	84	82	31	125	38.2	0.233	23
2	0	145	0	0	0	44.2	0.630	31
3	0	135	68	42	250	42.3	0.365	24
4	1	139	62	41	480	40.7	0.536	21
...	...	...	...	...	...	...	...	...
1995	2	75	64	24	55	29.7	0.370	33
1996	8	179	72	42	130	32.7	0.719	36
1997	6	85	78	0	0	31.2	0.382	42
1998	0	129	110	46	130	67.1	0.319	26
1999	2	81	72	15	76	30.1	0.547	25

2000 rows × 8 columns

Figure 7.3: SPLITTING DATA

After dividing the data into independent and dependent data sets as X, Y. We can split the data accordingly as shown in the figure.



```
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.33)
x_train
```

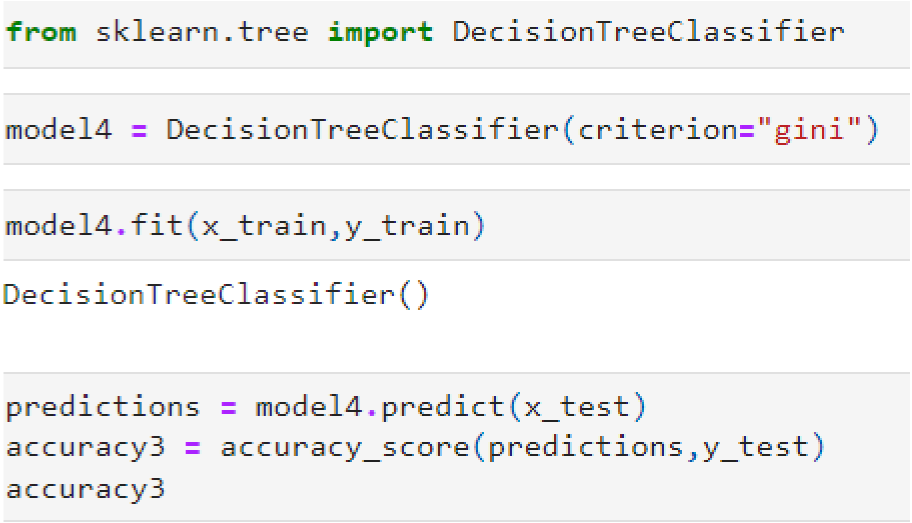
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
1954	8	84	74	31	0	38.3	0.457	39
1446	4	96	56	17	49	20.8	0.340	26
1735	0	179	50	36	159	37.8	0.455	22
1377	5	112	66	0	0	37.8	0.261	41
281	10	129	76	28	122	35.9	0.280	39
...	...	...	...	...	...	...	...	...
867	0	131	66	40	0	34.3	0.196	22
548	1	164	82	43	67	32.8	0.341	50
1798	3	158	64	13	387	31.2	0.295	24
983	2	91	62	0	0	27.3	0.525	22
763	10	101	76	48	180	32.9	0.171	63

1340 rows × 8 columns

Figure 7.4: TRAINING DATA

Next, we train the data set with each and every ML model and find the accuracy.

7.5.1 DECISION TREE



```
from sklearn.tree import DecisionTreeClassifier
```

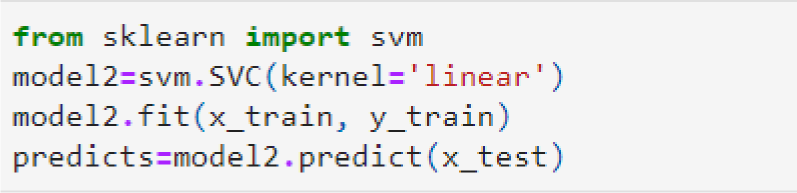
```
model4 = DecisionTreeClassifier(criterion="gini")
```

```
model4.fit(x_train,y_train)
```

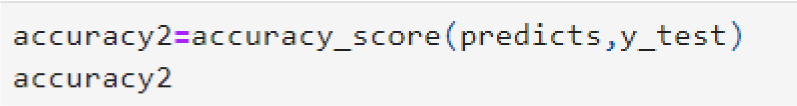
```
DecisionTreeClassifier()
```

```
predictions = model4.predict(x_test)
accuracy3 = accuracy_score(predictions,y_test)
accuracy3
```

7.5.2 SVM

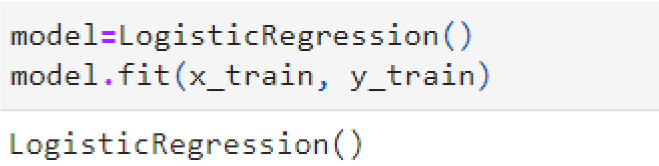


```
from sklearn import svm
model2=svm.SVC(kernel='linear')
model2.fit(x_train, y_train)
predicts=model2.predict(x_test)
```



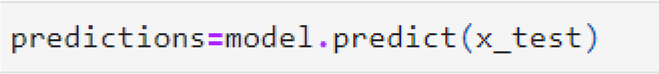
```
accuracy2=accuracy_score(predicts,y_test)
accuracy2
```

7.5.3 LOGOSTIC REGRESSION

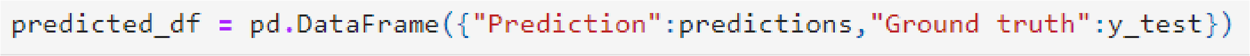


```
model=LogisticRegression()
model.fit(x_train, y_train)
```

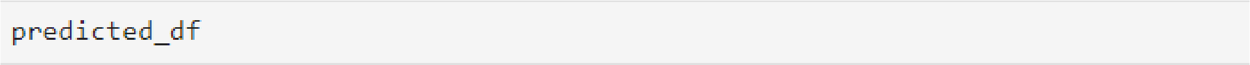
```
LogisticRegression()
```



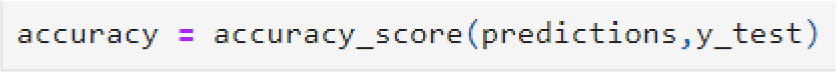
```
predictions=model.predict(x_test)
```



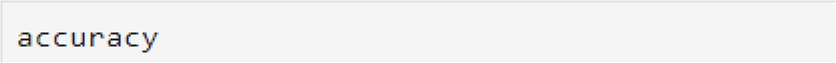
```
predicted_df = pd.DataFrame({"Prediction":predictions,"Ground truth":y_test})
```



```
predicted_df
```

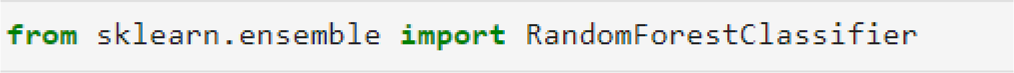


```
accuracy = accuracy_score(predictions,y_test)
```

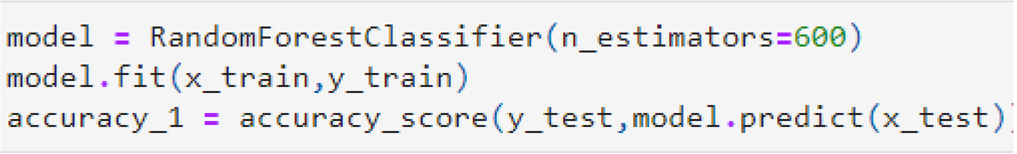


```
accuracy
```

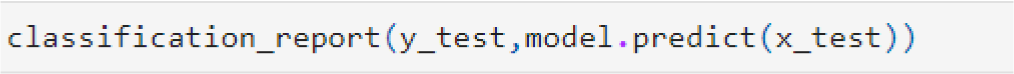
7.5.4 RANDOM FOREST



```
from sklearn.ensemble import RandomForestClassifier
```



```
model = RandomForestClassifier(n_estimators=600)
model.fit(x_train,y_train)
accuracy_1 = accuracy_score(y_test,model.predict(x_test))
```



```
classification_report(y_test,model.predict(x_test))
```

CHAPTER 8

RESULT AND DISCUSSION

8.1 Result

The result of each classifier has been evaluated using different evaluation metrics The experiments are conducted using Python 3.3 programming language through the Jupyter Notebook application. Several libraries from KAGGLE have been used, which is a free software for the machine learning library in Python. Each model generates different outputs depending on the different values of its parameter:

LOGISTIC REGRESSION – 76%

DECISION TREE – 94.6%

RANDOM FOREST CLASSIFIER– 96.5%

SUPPORT VECTOR MACHINE – 77%

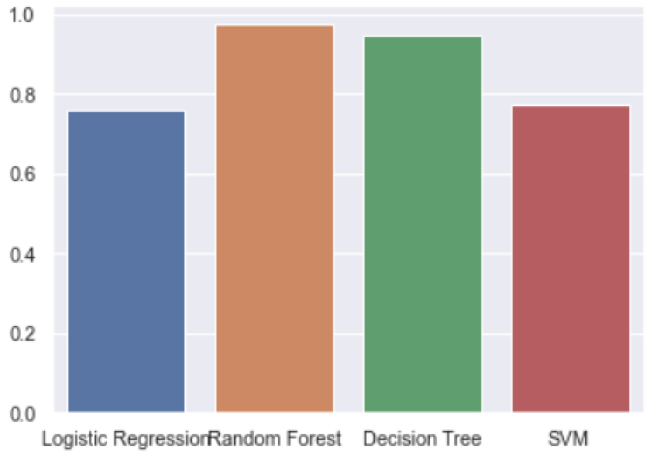


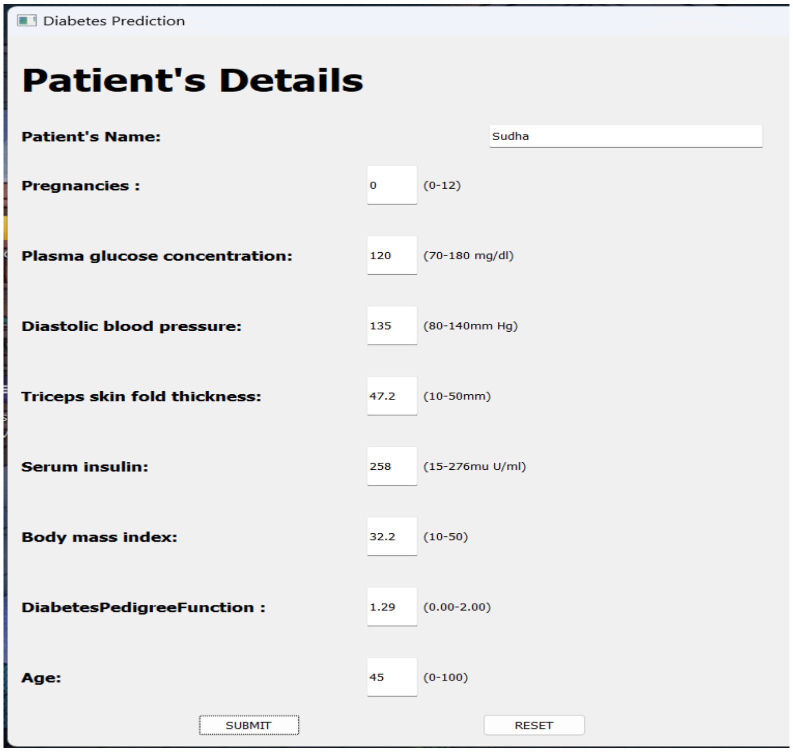
Figure 8.1: ACCURACY BARGRAPH

Among all classifiers we got near 100 % accuracy for Random Forest and Decision tree, in this two we are using random forest because decision tree has overfitting issue and it can be resolved using random forest and Decision Tree computational complexity is more when compared to Random Forest. In this we have only 600 instances but if we have more data decision tree makes it complex since it has lot of layers. By considering all these aspects, we are using Random Forest. When you don't bother much about interpreting the model but want better accuracy. Random forest will reduce variance part of error rather than bias part, so on a given training data set decision tree may be more accurate than a random forest. But on an unexpected inference data set, Random forest always wins in terms of accuracy.

8.2 Creating a User Interface for Accessibility:

The last part of the project is the Creation of a user interface for the model. This user interface is used to enter unseen data for the model to read and then make a prediction. The user interface is created using "PyQt5". PyQt5 is a cross-platform GUI toolkit and a set of Python bindings for the Qt v5 application development framework. It enables Python to be used as an alternative application development language to C++ on all supported platforms including iOS and Android. A GUI application consists of Front-end and Back-end. PyQt5 has provided a tool called 'QtDesigner' to design the front-end by drag and drop method so that development can become faster and one can give more time on back-end stuff.

The project predicts the onset of diabetes in a person based on the relevant medical details collected. When the person enters all the relevant medical data required in the online Web portal, this data is then passed on to the trained model for it to make predictions whether the person is diabetic or non-diabetic the model then makes the prediction with an accuracy of 97%, which is fairly good and reliable. Following figure shows the basic UI form which requires the user to enter the specific medical data fields. These parameters help determine if the person is prone to develop diabetes. Our research has the added benefit of an associated Web Site, which makes the model more user friendly and easily understandable for a novice



Patient's Details

Patient's Name: Sudha

Pregnancies : 0 (0-12)

Plasma glucose concentration: 120 (70-180 mg/dl)

Diastolic blood pressure: 135 (80-140mm Hg)

Triceps skin fold thickness: 47.2 (10-50mm)

Serum insulin: 258 (15-276mu U/ml)

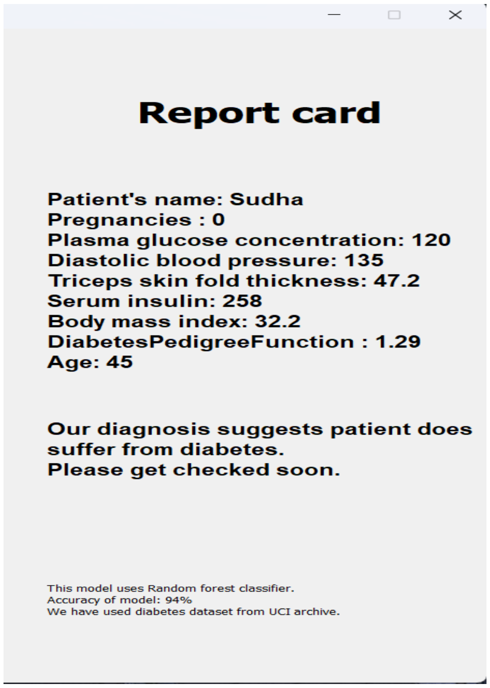
Body mass index: 32.2 (10-50)

DiabetesPedigreeFunction : 1.29 (0.00-2.00)

Age: 45 (0-100)

Figure 8.2: DATA ENTRY PAGE

On submission of this form the result will be showed in the form of test like below figure:



Report card

Patient's name: Sudha
Pregnancies : 0
Plasma glucose concentration: 120
Diastolic blood pressure: 135
Triceps skin fold thickness: 47.2
Serum insulin: 258
Body mass index: 32.2
DiabetesPedigreeFunction : 1.29
Age: 45

Our diagnosis suggests patient does suffer from diabetes.
Please get checked soon.

This model uses Random forest classifier.
Accuracy of model: 94%
We have used diabetes dataset from UCI archive.

Figure 8.3: RESULT PAGE

CHAPTER 9

CONCLUSION

The main aim of this project was to design and implement Diabetes Prediction Using Machine Learning Methods and Performance Analysis of that methods and it has been achieved successfully. The proposed approach uses various classification and ensemble learning method in which SVM, Decision Tree, Logistic Regression classifiers are used. And we get more than 80% classification accuracy has been achieved.

The Experimental results can be forwarded to health care for taking necessary precautions and make early decision to cure diabetes and save humans life.

Machine learning has the great ability to revolutionize the diabetes risk prediction with the help of advanced computational methods and availability of large amount of epidemiological and genetic diabetes risk dataset. Detection of diabetes in its early stages is the key for treatment. This work has described a machine learning approach to predicting diabetes levels. The technique may also help researchers to develop an accurate and effective tool that will reach at the table of Dillons to help them make better decision about the disease status. Logistic regression gives 76% accuracy and Decision tree accuracy 94% and Random Forest gives 96% accuracy.

Conclusion In this research study, we have used different types of machine learning algorithm for detection of diabetes. the dataset and We implemented machine learning algorithms on performed classification to signify the best machine learning algorithm for diabetes prediction on the bases of old data available. The higher accuracy the better prediction rate we will achieve. The random forest algorithm obtained the best accuracy and Roc. While on the other hand logistic regression had the lowest score. The overall experimentation displayed that random forest is better than other algorithms in diabetes prediction. In our future work we will work on the diagnosis of diabetes.

REFERENCES

1. Debadri Dutta, Debpriyo Paul, Parthajeet Ghosh, "Analyzing Feature Importance's for Diabetes Prediction using Machine Learning". IEEE, pp 942 928, 2018.

2. K. VijiyaKumar, B.Lavanya, 1.Nirmala, S.Sofia Caroline, "Random Forest Algorithm for the Prediction of Diabetes". Proceeding of International Conference on Systems Computation Automation and Networking, 2019. a Md.

3. Faisal Faruque, Asaduzzaman, lqbal H. Sarker. "Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus". International Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 February, 2019.

4. 4 Tejas N. Joshi, Prof. Pramila M.Chawan, "Diabetes Prediction Using Machine Learning Techniques". Int. Journal of Engineering Research and Application, Vol. 8, Issue 1, (Part -ll) January 2018, pp.-09-13

5. 5 Nonso Nhamoko, Abir Hussain, David England, "Predicting Diabetes Onset: an Ensemble Supervised Learning Approach ". IEEE Congress on Evolutionary Computation (CEC), 2018.

6. Han YE, Sun TJ, Han YQ, et al. Clinical perspectives on mesenchymal stem cells promoting wound healing in diabetes mellitus patients by inducing autophagy.Eur.Rev. Med.Pharmacol. Sci. 19(14), 2666-2670 (2015).

7. Dave SD, Vanikar AV, Trivedi HL,et al. Novel therapy for insulin-dependent diabetes mellitus: infusion of in vitro-generated insulin-secreting cells. Clin. Exp. Med. 15(1),41-45 (2015).

8. Chena L, Hammond H, Ye Z, Zhan X, Dravid G. Human adult marrow cells Support prolonged expansion of human embryonic stem cells in culture. Stem Cells. 21, 131142(2003).

9. Wendelin Schramm, Fabian Sailer, Monika Pobiruchin, Christian VWeiss: PROSIT Open Source Disease Models for Diabetes Mellitus.JCIMTH 2016: 115-118.

10. Lucas Felipe Klein, Sandro José Rigo, Sílvio César Cazella. Ângela .Jornada Ben: An Application for Mobile Devices Focused on Clinical Decision Support: Diabetes Mellitus Case. ISAml 2016: 57-65

11. Vaclay Burda, Daniel Novák, Jakub Schneider: Evaluation of diabetes mellitus Compensation after one year of using Mobiab system. EMBC 2016

12. Komi, Zhai. 2017. Application of Data Mining Methods in Diabetes Prediction.

13. Analysis of Various Data Mining Techniques to Predict Diabetes Mellitus, Omar Kassem Khalil Aissa Boudjella, 2016 Sixth International Conference on Developments in eSystems Engineering.

14. Nahla B., Andrew et al,"Intelligible support vector machines for diagnosis of diabetes mellitus. Information Technology in Biomedicine", IEEE Transactions. 14, (July. 2010), 1114-20.