

**Predictive Study and Classification of Diabetes Using
Machine Learning Techniques**

A DISSERTATION

SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE AWARD OF DEGREE
OF
MASTER OF TECHNOLOGY
IN
SOFTWARE ENGINEERING

Submitted by:

Krishan Kumar
2K20/SWE/12

Under the supervision of
Sanjay Patidar
(Assistant Professor)



DEPARTMENT OF SOFTWARE ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

MAY, 2022

DEPARTMENT OF SOFTWARE ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)


Bawana Road, Delhi – 110042

CANDIDATE'S DECLARATION

I, Krishan Kumar, Roll No. 2K20/SWE/12 student of M. Tech (Software Engineering), hereby declare that the project Dissertation titled “Predictive Study and Classification of Diabetes Using Machine Learning Techniques” which is submitted by me to the Department of Software Engineering, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of Master of Technology, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of and Degree, Diploma Associateship, Fellowship or other similar title or recognition.

Place: Delhi

Date: 30 May, 2022


Krishan Kumar

DEPARTMENT OF SOFTWARE ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi – 110042

CERTIFICATE

I hereby certify that the Project Dissertation titled “Predictive Study and Classification of Diabetes Using Machine Learning Techniques” which is submitted by Krishan Kumar, 2K20/SWE/12 Department of Software Engineering, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of Master of Technology, is a record of the project work carried out by the students under my supervision. To the best of my knowledge, this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi

Date: 31/05/2022



Mr. Sanjay Patidar

Assistant Professor

Department of

Software Engineering

ACKNOWLEDGEMENT

The success of this project requires the assistance and input of numerous people and the organization. I am grateful to everyone who helped in shaping the result of the project. I express my sincere thanks to **Mr. Sanjay Patidar**, my project guide, for providing me with the opportunity to undertake this project under his guidance. His constant support and encouragement have made me realize that it is the process of learning which weighs more than the end result. I am highly indebted to the panel faculties during all the progress evaluations for their guidance, constant supervision and for motivating me to complete my work. They helped me throughout with new ideas, provided information necessary and pushed me to complete the work. I also thank all my fellow students and my family for their continued support.



Krishan Kumar

2K20/SWE/12

ABSTRACT

Diabetes mellitus (DM) is a common but deadly disease in humans. It is caused by having excessive sugar levels existing for a long time. It causes around 30 to 40 lakh deaths worldwide each year. Technology plays a consequential role in the medical industry to assess diabetes prediction research studies. In this research, we trained four machine learning techniques so as to make predictions on whether a person is diabetic or not, on the basis of some health details of the individual. The Pima Indian Diabetes dataset is used, which consists of 768 samples, and each sample contains 8 attributes and one target class attribute.

Data pre-processing techniques are used to get the raw dataset cleaned, to remove the inconsistencies, anomalies and missing values present in the data which are not suitable for the machine learning models. K-Nearest Neighbour (KNN), Logistic Regression (LR), Support Vector Machine (SVM) and Artificial Neural Networks (ANN) are the techniques used for prediction of diabetes in this research. As a result, K-Nearest Neighbour which is a classification machine learning technique, performed the best, with an accuracy of 76.17% , while support vector machine, logistic regression and Artificial Neural Network gave 75.8%, 73.04% and 73.82% respectively.

CONTENTS

Candidate's Declaration	i
Certificate	ii
Acknowledgement	iii
Abstract	iv
Contents	v
List of Figures	viii
List of Tables	x
List of Abbreviations	xii
CHAPTER 1: INTRODUCTION	1
1.1 OVERVIEW	1
1.2 MOTIVATION	2
1.3 MACHINE LEARNING	2
1.4 TYPES OF DATASETS IN MACHINE LEARNING	3
1.4.1 LABELED DATA	3
1.4.2 UNLABELED DATA	4
1.5 TYPES OF MACHINE LEARNING	4

1.5.1 SUPERVISED MACHINE LEARNING	5
1.5.2 UNSUPERVISED MACHINE LEARNING	5
1.5.3 REINFORCEMENT MACHINE LEARNING	6
1.6 MODEL VALIDATION	6
1.6.1 CONFUSION MATRIX	7
CHAPTER 2: LITERATURE SURVEY	8
CHAPTER 3: PROPOSED MODEL	10
3.1 DATA EXTRACTION	11
3.1.1 DATA READING	11
3.1.2 DATA PRESENTATION	12
3.1.3 DATA CORRELATION	16
3.2 DATA PRE-PROCESSING	17
3.2.1 ELIMINATION OF NaN VALUES	17
3.2.2 EVALUATION OF CLASS DISTRIBUTION	20
3.2.3 DATA SCALING	20
3.3 DATA SPLITTING	20
3.4 MACHINE LEARNING METHODOLOGIES	21

3.4.1 K-NEAREST NEIGHBOR	21
3.4.2 SUPPPORT VECTOR MACHINE	23
3.4.3 LOGISTIC REGRESSION	24
3.4.4 ARTIFIFCIAL NEURAL NETWORKS	26
CHAPTER 4: RESULTS	27
4.1 RESULTS OF KNN	27
4.2 RESULTS OF SVM	28
4.3 RESULTS OF LR	29
4.4 RESULTS OF ANN	30
CHAPTER 5: CONCLUSION AND FUTURE SCOPE	32
REFERENCES	33

LIST OF FIGURES

Fig. 1.1. Artificial Intelligence, Machine Learning and Deep Learning	2
Fig.1.2. Conversion of unlabeled data into labelled data	3
Fig.1.3. Labelled Data	3
Fig.1.4 Unlabeled Data	4
Fig.1.5. Types Of Machine Learning	4
Fig.1.6. Supervised Learning	5
Fig.1.7 Unsupervised Learning	6
Fig.3.1. Proposed Architecture	10
Fig.3.2. Histogram of No of Pregnancies attribute	13
Fig.3.3. Histogram of Glucose attribute	13
Fig.3.4. Histogram of Blood Pressure attribute	13
Fig.3.5. Histogram of Skin Thickness attribute	14
Fig.3.6. Histogram of Insulin attribute	14
Fig.3.7. Histogram of BMI attribute	15
Fig.3.8. Histogram of Diabetes Pedigree Function attribute	15
Fig.3.9. Histogram of Age attribute	15
Fig.3.10. Histogram of Outcome attribute	16
Fig.3.11. Plot Graph for Correlation Matrix of attributes	17
Fig.3.12. Data values count before replacing NaN values	18
Fig.3.13. Data values count before replacing NaN values	19
Fig.3.14. Outcome dataset	20

Fig.3.15. Data Splitting	21
Fig.3.16. Learning curve of KNN.	22
Fig. 3.17 Confusion matrix of KNN.	22
Fig. 3.18 Learning Curve of SVM.	23
Fig. 3.19 Confusion Matrix of SVM.	24
Fig. 3.20 Learning Curve of Logistic Regression.	25
Fig. 3.21 Confusion Matrix of Logistic Regression.	25

LIST OF TABLES

Table 1.1. Confusion Matrix	7
Table 3.1. List of attributes present in the dataset and their type	11
Table 3.2. Head of the PIMA Indian Diabetes Dataset	12
Table 3.3. Tail of the PIMA Indian Diabetes Dataset	12
Table 3.4. Correlation matrix of attributes	16
Table 3.5. Data statistics before replacing NaN values	18
Table 3.6. NaN values count	19
Table 3.7. Data statistics after replacing NaN values	19
Table 3.8. Verification of data class distribution	20
Table 3.9. Verification of data splitting	31
Table 4.1 Training score of KNN	27
Table 4.2 Testing score of KNN	27
Table 4.3 Training score SVM	28
Table 4.4 Testing score SVM	28
Table 4.5 Training score of LR	29
Table 4.6 Testing score of LR	29
Table 4.7. Training scores of Artificial Neural Networks	30
Table 4.8. Training scores of Artificial Neural Networks	30
Table 4.9. Performance measures of all the four algorithms	31

LIST OF ABBREVIATIONS

ML	Machine Learning
KNN	K-Nearest Neighbor
ANN	Artificial Neural Networks
SVM	Support Vector Machine
LR	Logistic Regression

CHAPTER 1

INTRODUCTION

1.1 OVERVIEW

In today's world, technology surrounds us in every possible way, usage of modern-day technologies to get some ease in our work, some of the examples we can see are driverless cars, maps GPS navigation, voice-controlled devices and many others. All these modern technologies use the data of the real world to train their models and finally test them with the help of data again and again to generate better accuracy. The degree of ease and accuracy of the technology is directly proportional to its usage. Machine learning is also a similar technique in which various algorithms are used to train the model using some part of data and then test the model with some other part of data to generate the outcome.

This modern-day technique can be used in the prediction of some disease with the help of associated symptoms. Algorithms can train the model in which symptoms values will be used as inputs and generation of outcome will be tested to get the best accuracy and some other different measures, which will help the patient and the medical field to deal with the patient's condition in the earlier stage [1]. We have gathered the data from PIMA Indian Diabetes dataset. Then some preprocessing techniques are applied on the dataset, to get the dataset in suitable format for the machine learning models. As the dataset that we selected has labeled data in it, we have used four supervised machine learning algorithms, to check which one gives the best accuracy of making predictions.

Our study is structured in the following order - Section 2 contains literature review. The next, section 3 explains the procedural approach, the machine learning techniques used and the model evaluation. Section 4 discusses about the final result obtained from the research. The last Section 5 contains the conclusion and future work.

1.2 MOTIVATION

Diabetes is a major cause of death, metabolic disorders in humans as well as leads to commercial and productivity loss throughout the world due to lower levels of efficiency of man power. It is a metabolic disorder, characterized by high blood sugar levels which is caused by low insulin production in the pancreas. It increases the risk of long-term complications. It increase the chances of heart disease and about 75% people having this disease die due to coronary artery disease.

In this research, four machine learning algorithms are compared in order to predict risk of someone having diabetes in future. Classification algorithms are used to classify the target outcomes independently.

1.3 MACHINE LEARNING

The modern world is booming with all the new artificial intelligence technologies. The aim of artificial intelligence is to make machines to think link human minds. And machine learning is a sub part of artificial intelligence, as shown in fig. 1.1[2].

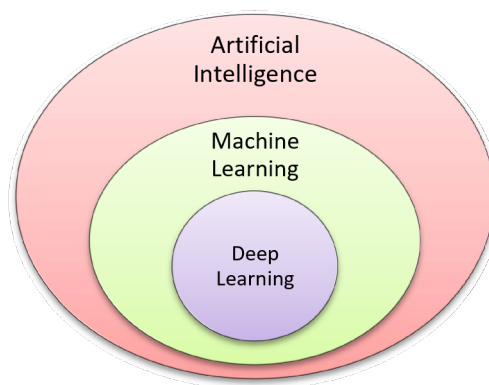


Fig. 1.1. Artificial Intelligence, Machine Learning and Deep Learning[2].

Machine learning is exactly what its name is saying, combination of “machine” and “learning”. Machine learning is the ability of the machine to self-learn about the data from a particular input of dataset that we give to the machine. And by the help of this learning feature ability, the machine can make predictions on the basis

of the learning it has already done by the dataset that we gave to it as input data in the very beginning.

1.4 TYPES OF DATASETS IN MACHINE LEARNING

Data extraction is the very first step in making a machine learning model. The data is what we give to our machine learning model as input to get the model learn about that dataset by training on it, to make predictions on the basis of the information present in that dataset.

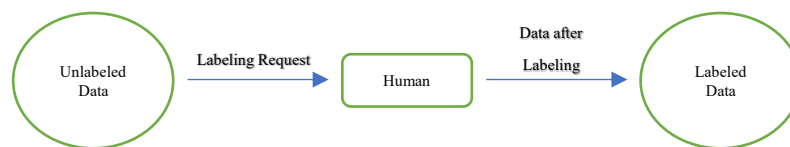


Fig.1.2. Conversion of unlabelled data into labelled data.

There are two types of data that we talk about in machine learning, the first is the labelled data, and the other one is the unlabelled data.

1.4.1 LABELLED DATA

Labelled data is the dataset type which consist of both input and output features in a form which is easily readable by the machine. Whole data is labelled. So, the machine can understand about the fixed relation between both input and output. As shown in fig 1.3[3].

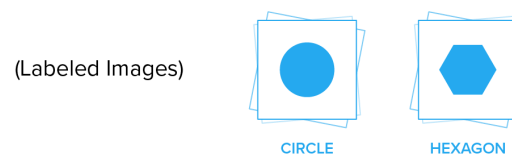


Fig.1.3. Labelled Data[3]

1.4.2 UNLABELLED DATA

Unlabelled data is the dataset type which contains data which is not easy to understand by the machine. The data is not easy to be read by the machine as it is not labelled as shown in fig 1.4.[3].

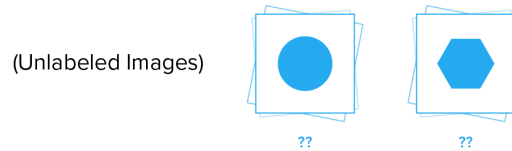


Fig.1.4 Unlabelled Data[3]

1.5 TYPES OF MACHINE LEARNING

Machine learning techniques can be divided into these following categories: also shown in fig1.5[4].

- SUPERVISED ML.
- UNSUPERVISED ML.
- REINFORCEMENT ML.

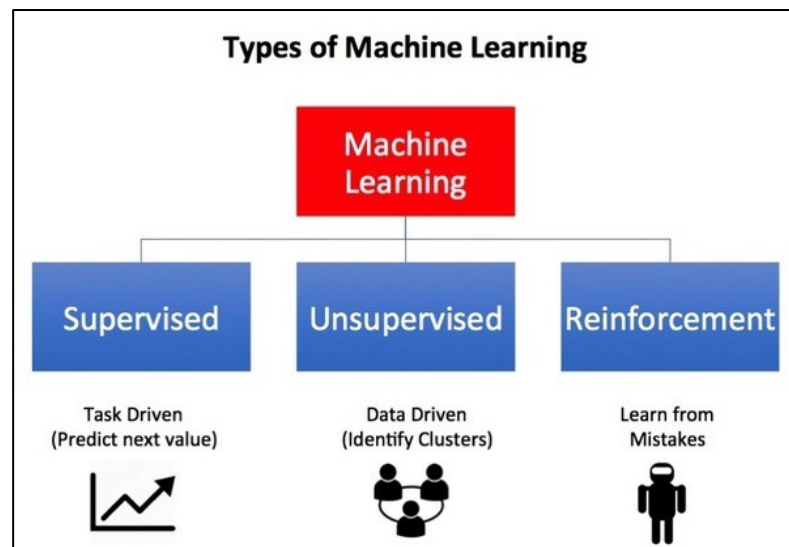


Fig.1.5. Types Of Machine Learning[4]

1.5.1 SUPERVISED MACHINE LEARNING

Supervised machine learning techniques work on the data which is easily readable by the machine i.e., labelled data. The labelled data is put into labelled form by the human labour in the starting, and then it is passed on to supervised machine learning techniques. In the supervised learning, the machine gets the well readable labelled data as input and then tries to get the idea of the exact relation between the input and the output attributes as shown in fig 1.6[5]

Divided into:

- Classification
- Regression

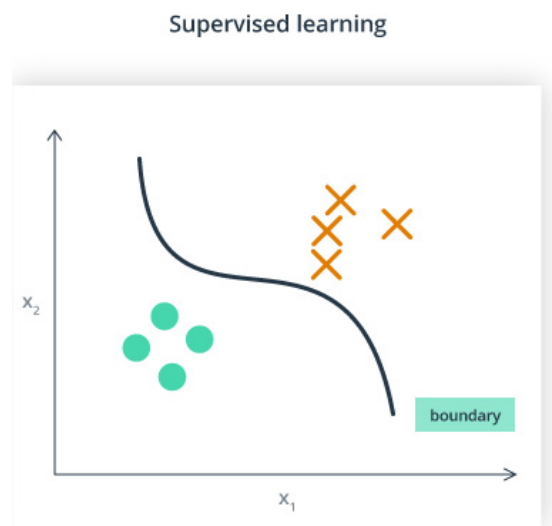


Fig.1.6. Supervised Learning[5]

1.5.2 UNSUPERVISED MACHINE LEARNING

Unsupervised machine learning techniques work with the data which is not easy to be read by the machine. It has the advantage over supervised machine learning techniques, that it can deal with unlabelled data as shown in fig 1.7[5]. Unsupervised machine learning techniques are versatile in nature, as they find hidden patterns in the unlabelled dataset.

Divided into:

- Clustering
- Association

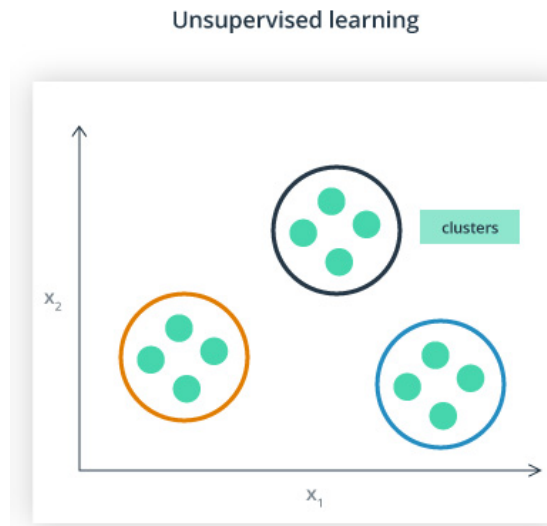


Fig.1.7 Unsupervised Learning[5]

1.5.3 REINFORCEMENT MACHINE LEARNING

Reinforcement machine learning directly acts as human, as how we see data and learn from it in our daily lives. It works on algorithm that self improves by using and hit-and-trial method and learns from new situations. Encouragement or reinforcement is given to the outputs which are correctly predicted, and discouragement or punishment is shown to the output which are not in favour.

1.6 MODEL VALIDATION

The methodologies were performed on the jupyter notebook. The data was analyzed using data visualization techniques and conformed via performance evaluation metrics such as accuracy, precision, recall, f-score, etc. Cross-validation method was used for evaluation. In k-fold cross validation, we broke the data into k distinct sets which are exclusive in nature and have equal size, with one set used for training purpose and other for testing.

1.6.1 CONFUSION METRICS

The study is evaluated/validated via confusion matrix using metrics such as accuracy, precision, recall and f-score. Table 1.1 shows the confusion matrix.

Table 1.1. Confusion Matrix

Output		Predicted Values	
		Diabetic	Non-Diabetic
Actual Values	Diabetic	TP	FN
	Non-Diabetic	FP	TN

True Positives (TP) : Total predicted diabetic cases, validated as diabetic.

True Negative (TN) : Total predicted non-diabetic cases, validated as non-diabetic.

False Positives (FP) : Total predicted diabetic cases, validated as non-diabetic.

False Positives (FN) : Total predicted non-diabetic cases, validated as diabetic.

$$\text{Precision} = \frac{TP}{TP+FP}, \text{ Recall} = \frac{TP}{TP+FN},$$

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}, \text{ F1 - score} = \frac{2TP}{2TP+FP+FN}$$

CHAPTER 2

LITERATURE REVIEW

Diabetes diagnosis and treatment has been a crucial topic in medical research from a very long period of time. With the help of Machine learning, a really good progress has been made in the process of predicting diabetes in people. This prediction is made by the help of machine learning models, which are trained on the dataset consisting of medical information of patients. After the training phase the model is evaluated by passing the testing data to the model, to check how efficiently the model is working.

Humar Kahramanli et al [6] used amalgamation of Artificial Neural Networks and fuzzy logics to make a model with good accuracy to predict diabetes. A.Kumar Dwivedi et al. [7] compared five machine learning algorithms to predict diabetes. The algorithms used were artificial neural networks, classification tree, KNN, SVM and logistic regression. The author in [8] used two classification algorithms, deep neural networks and artificial neural networks. And also used principal component analysis. Using deep neural networks they achieved better accuracy of 82.67%.

Nawaz Mahamudally et al. [9] used k-means clustering, neural networks and C4.5 decision tree algorithm to predict diabetes in patients. Bayesian network, Artificial neural network, SVM, Decision tree and KNN were used to predict diabetes, by M. Heydari[10]. Hasans Temurtas el at. [11] made a model which was trained by Levenberg–Marquardt (LM) algorithm, and the model was combined with multilayer neural network structure. Vijayashree et al. [12] proposed a system that uses recursive feature elimination and principal component analysis for prediction of diabetes. Mukesh Kumari et al. [13] used data mining techniques to predict diabetes mellitus.

K. Rajesh and V. Sangeetha (2012) used classification technique. They used C4.5 decision tree algorithm to find hidden patterns from the dataset for classifying efficiently[14]. Mani Butwall and Shraddha Kumar (2015) proposed a model using Random Forest Classifier to forecast diabetes behaviour[15].

Goncalves et al. [16] introduced a system to predict diabetes using hierarchical Neuro-Fuzzy BSP method. Ashiquzzaman et al. [17] proposed a prediction framework for the diabetes mellitus using deep learning approach where the overfitting is diminished by using the dropout method. Zhu et al. [18] proposed a system using multiple classifiers and improved the accuracy of complex disease prediction like diabetes. B.M. Patil proposed Hybrid Prediction Model which includes Simple K-means clustering algorithm, followed by application of classification algorithm to the result obtained from clustering algorithm. In order to build classifiers C4.5 decision tree algorithm is used[19].

Patients can have several symptoms and some of the symptoms and factors are included in the data set like Age, Insulin level, Glucose level, Diabetes Pedigree Function, Blood Pressure level, Skin Thickness and BMI. Prediction of the outcome from data has been done using various traditional machine learning techniques and artificial neural networks. In order to apply these algorithms, we need to preprocess the data which includes standardizing, normalizing and cleaning of the data. Then proposed algorithms are applied and their performances are validated. These prerequisite actions are necessary so that optimal levels of accuracy, precision and recall can be obtained.

In this paper, classification algorithms are used on the diabetic patients data set to predict the outcome of diabetes presence in patients and we achieved a success rate on the test set of 76%. Moreover, we were able to obtain this much accuracy with traditional machine learning approaches, by adding some data pre-processing techniques.

CHAPTER 3

PROPOSED MODEL

So the architecture that we have proposed in this thesis is shown in fig.3.1. The first step starts with data extraction, in data extraction we have gathered the data from the PIMA Indian diabetes dataset. After extraction, the data is sent for the data pre-processing step. The raw data that we get in the data extraction process is not in the condition to be directly used by the machine learning models. So to get the data in the form that it can be used by the machine learning models, we have applied some data pre-processing techniques.

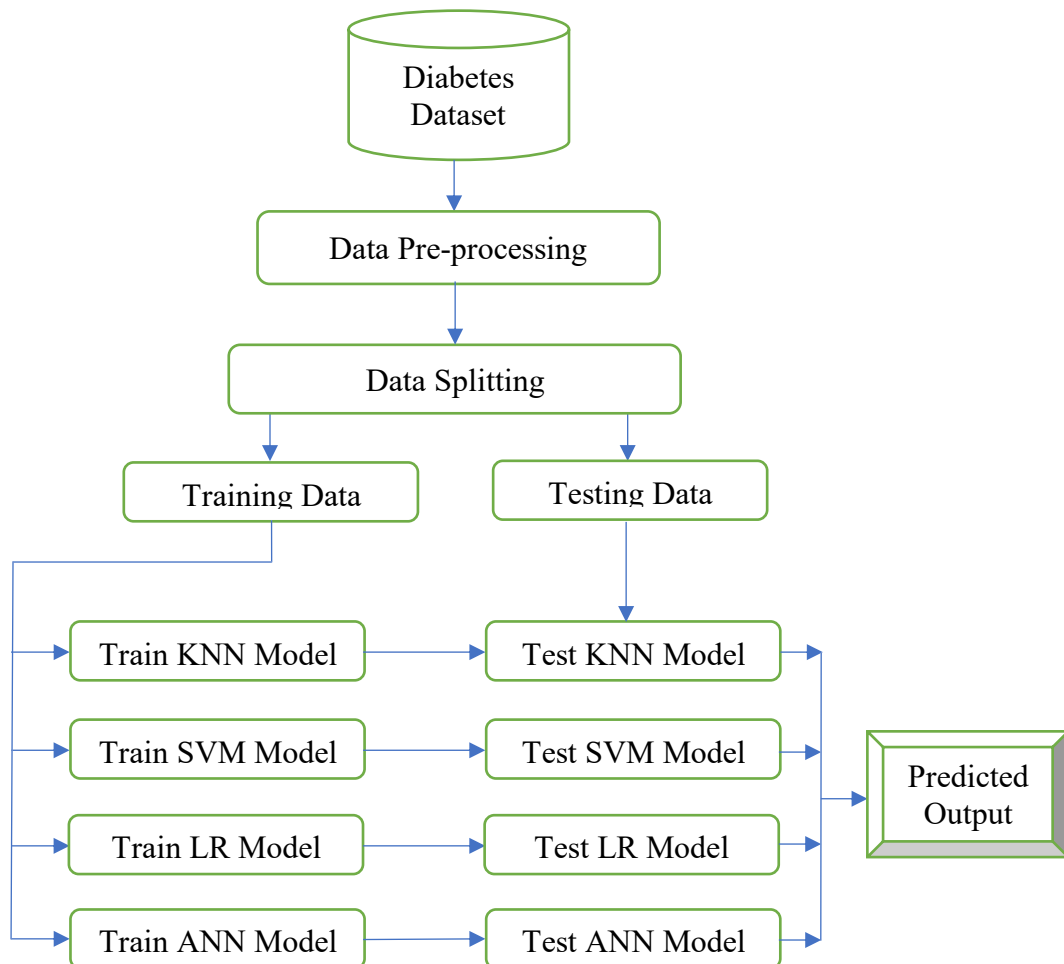


Fig.3.1. Proposed Architecture

Data pre-processing techniques deals with the inconsistencies, anomalies and missing values present in the data. After data pre-processing, data splitting is done.

Data splitting is performed to split the data into two parts, training data and testing data. The first one is the training data, which is used to train the machine learning models. And the second one is the testing data, which is used to check how well the model is performing after being trained, which is also known as model evaluation.

3.1 Data Extraction

The dataset which we have extracted is the PIMA Indian diabetes dataset. Aim of which is to predict whether or not a patient is diabetic, on the basis of several attributes present in the dataset. Different criteria were used on the selection of these values from the database. The list of attributes present in the dataset and their data type is shown in table 3.1.

Table 3.1. List of attributes present in the dataset and their data type.

ATTRIBUTE NAME	DATA TYPE
<i>Pregnancies</i>	<i>Integer</i>
<i>Glucose</i>	<i>Integer</i>
<i>BloodPressure</i>	<i>Integer</i>
<i>SkinThickness</i>	<i>Integer</i>
<i>Insulin</i>	<i>Integer</i>
<i>DiabetesPedigreeFunction</i>	<i>Float</i>
<i>BMI</i>	<i>Float</i>
<i>Age</i>	<i>Integer</i>
<i>Outcome</i>	<i>Integer</i>

3.1.1 Data Reading

The model is constructed using python language. So, to read the data from the dataset, we have used “pd.read_csv” command.

As shown in table 3.2 and table 3.3, the PIMA Indian diabetes dataset contains the medical details of 768 different patients, and these medical details were used for classifications. These medical details were stored in 768 rows and 9 columns. 9 columns consisting of 8 independent attributes and one class column ‘Outcome’ (diabetic or non-diabetic).

So, the head of the dataset looks like this, as shown in table. 3.2 showing details of first five patients containing 9 attributes along with it, out of which 8 attributes are the

medical details of the patient and one is the outcome class. Similarly, the tail of the dataset is as shown in table. 3.3.

Table. 3.2. Head of the PIMA Indian Diabetes Dataset.

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

Table. 3.3. Tail of the PIMA Indian Diabetes Dataset.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
762	9	89	62	0	0	22.5	0.142	33	0
763	10	101	76	48	180	32.9	0.171	63	0
764	2	122	70	27	0	36.8	0.340	27	0
765	5	121	72	23	112	26.2	0.245	30	0
766	1	126	60	0	0	30.1	0.349	47	1
767	1	93	70	31	0	30.4	0.315	23	0

3.1.2 Data Presentation

In data representation we will see how the attributes are distributed in their own respective ranges. For data presentation we used the histograms. Histogram is a type of bar plot, in histograms the x-axis show the range of the attribute, and the y-axis will show the frequency of the attribute.

Fig. 3.2 shows the histogram for the ‘number of pregnancies’ attribute. So, according to this histogram representation of the ‘number of pregnancies’ of the patient, the histogram shows that patients having no of pregnancies from 0 to 2, are present most in the dataset with a frequency of value from around 240 to 250.

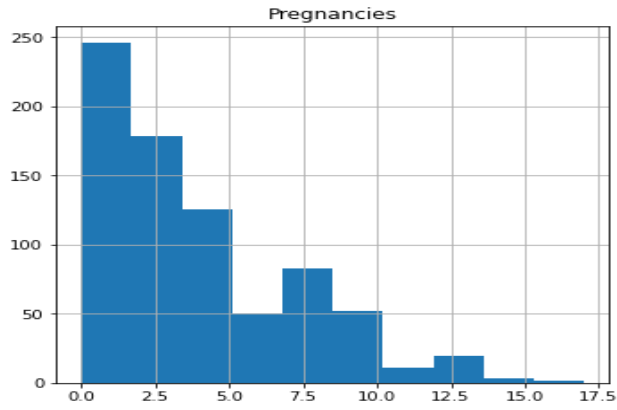


Fig.3.2. Histogram of no of pregnancies attribute.

Fig.3.3 shows the histogram for the 'glucose' attribute. So, according to this histogram representation of the 'glucose' of the patient, the histogram shows that patients having glucose level between 100 to 125, are present most in the dataset with number of entries ranging from around 210 to 220.

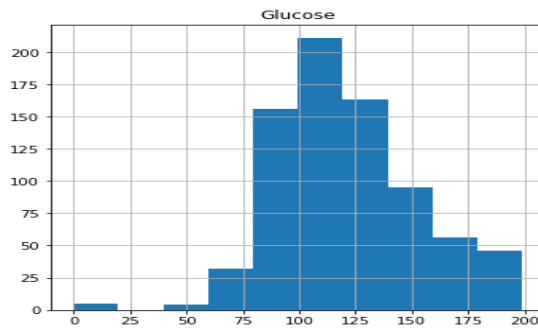


Fig.3.3. Histogram of glucose attribute.

Fig.3.4 shows the histogram for the 'blood pressure' attribute. So, according to this histogram representation of the 'blood pressure' of the patient, the histogram shows that patients having blood pressure level in the range 60 to 80, are present most in the dataset with number of entries ranging from around 260 to 280.

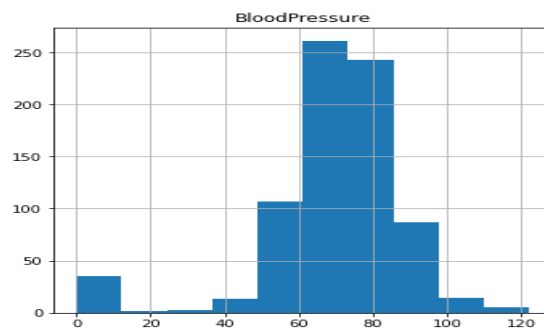


Fig.3.4. Histogram of Blood Pressure attribute.

Fig.3.5 shows the histogram for the ‘skin thickness’ attribute. So, according to this histogram representation of the ‘skin thickness’ of the patient, the histogram shows that patients having skin thickness in the range from 0 to 10, are present most in the dataset with number of entries ranging from around 220 to 240.

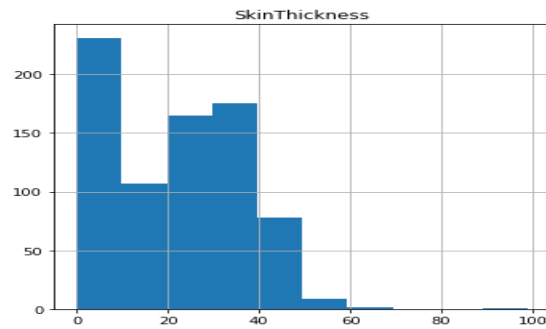


Fig.3.5. Histogram of Skin Thickness attribute.

Fig.3.6 shows the histogram for the ‘insulin’ attribute. So, according to this histogram representation of the ‘insulin’ of the patient, the histogram shows that patients having insulin in the range from 0 to 100, are present most in the dataset with number of entries ranging from around 460 to 500.

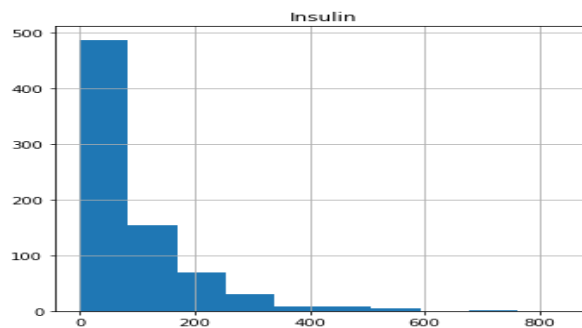


Fig.3.6. Histogram of Insulin attribute.

Fig.3.7 shows the histogram for the ‘BMI’ attribute. So, according to this histogram representation of the ‘BMI’ of the patient, the histogram shows that patients having BMI in the range from 25 to 35, are present most in the dataset with number of entries ranging from around 260 to 280.

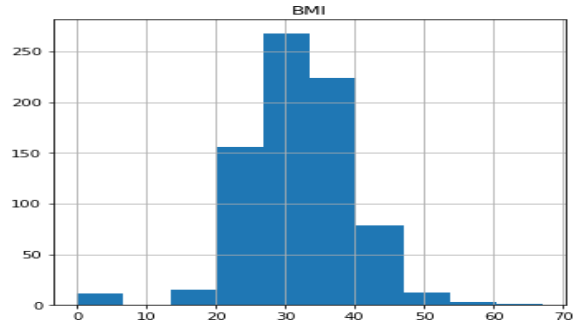


Fig.3.7. Histogram of BMI attribute.

Fig.3.8 shows the histogram for the ‘Diabetes Pedigree Function’ attribute. So, according to this histogram representation of the ‘Diabetes Pedigree Function’ of the patient, the histogram shows that patients having Diabetes Pedigree Function in the range from 0 to .25, are present most in the dataset with number of entries ranging from around 320 to 330.

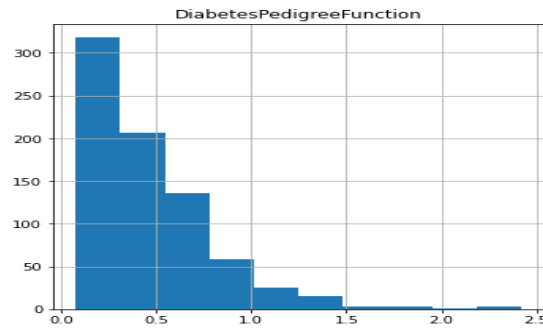


Fig.3.8. Histogram of Diabetes Pedigree Function attribute.

Fig.3.9 shows the histogram for the ‘Age’ attribute. So, according to this histogram representation of the ‘Age’ of the patient, the histogram shows that patients having Age in the range from 20 to 25, are present most in the dataset with 300 entries.

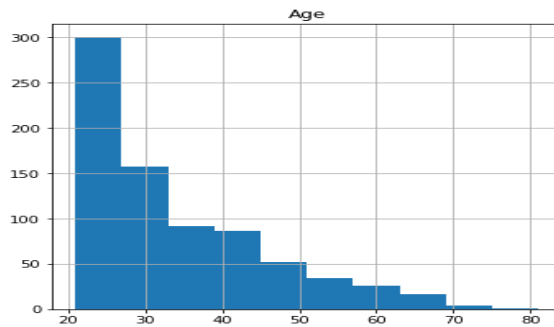


Fig.3.9. Histogram of Age attribute.

Fig.3.10 shows the histogram for the ‘outcome’ attribute. Outcome attribute contains only two outputs, which are 0 and 1. If the output is 0, it means the patient is not diabetic, and if the output is 1, it means the patient has diabetes.

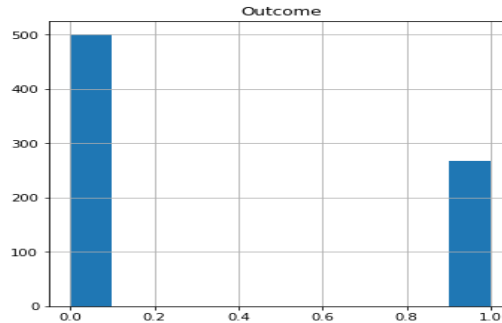


Fig.3.10. Histogram of Outcome attribute.

3.1.3 Data Correlation

Data correlation tells us about the co-dependency between the attributes. This co-dependency means if some changes are made to one attribute, will they affect the other attribute or not. If we make some changes in values of one attribute and those changes affects the values of the other attribute also, that means the attributes are correlated.

So, we have checked the data correlation between all the attributes of the dataset. After checking the data correlation, as shown in table 3.4. the outcome class has the highest correlation score of 0.46 with the ‘glucose’ attribute of the patient, followed by BMI, age and pregnancies attributes, with a correlation score of 0.29, 0.23 and 0.22, respectively. This means outcome class, which is whether a patient is diabetic or not, depends the most on the glucose level of the patient, and then the BMI, age and no. of pregnancies attributes of the patients.

Table 3.4. Correlation matrix of attributes.

	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

Just for a better presentation, along with tabular correlation table, we have added a colour correlation matrix of all the attributes present in the PIMA Indian Diabetes dataset, as shown in fig3.11.

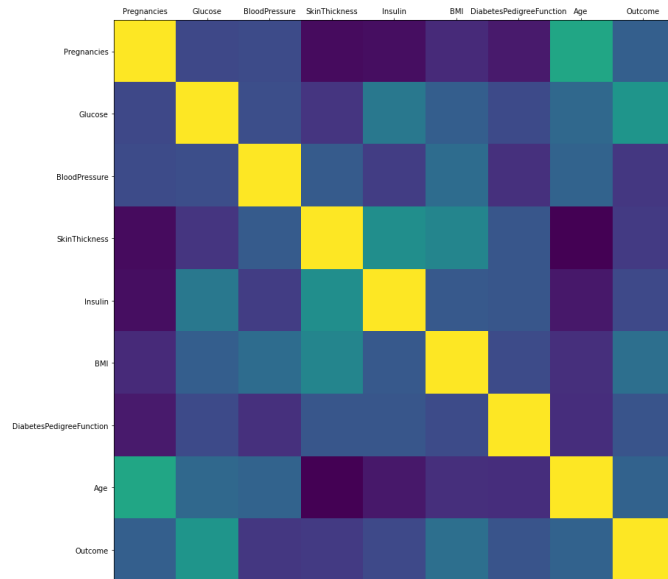


Fig.3.11. Plot Graph for Correlation Matrix of Attributes

3.2 Data Pre-processing

After the data is collected, it can't be directly used for the study, therefore it needs to be processed, cleaned and reduced to gather suitable information from the raw data useful for the study[20]. The raw data is expected to have many inconsistencies, anomalies, out of bound values, missing values or a format not suitable for our model. Hence, the data needs to be processed in order to use it for our study. Moreover, vast data in present day business, science, industry and academia scenarios needs complex mechanisms to analyze it. It includes data cleaning, transformation and normalization of data; and irregular data reduction tasks, used to reduce the convolution of data, determine and eliminate irrelevant and boisterous elements from the data through feature selection or discretization processes.

3.2.1 Elimination of NaN values.

The data was checked for any null values across all features and secondly in individual feature columns, Elimination of NaN(Not a number) values: We replaced these values

by the mean and median of the respected attributes. Table 3.6. shows the count of number of NaN values present in the data set before elimination of NaN values, and also after the elimination of NaN values.

Table 3.5. Data statistics before replacing NaN values.

	count	mean	std	min	25%	50%	75%	max
Pregnancies	768.0	3.845052	3.369578	0.000	1.00000	3.0000	6.00000	17.00
Glucose	768.0	120.894531	31.972618	0.000	99.00000	117.0000	140.25000	199.00
BloodPressure	768.0	69.105469	19.355807	0.000	62.00000	72.0000	80.00000	122.00
SkinThickness	768.0	20.536458	15.952218	0.000	0.00000	23.0000	32.00000	99.00
Insulin	768.0	79.799479	115.244002	0.000	0.00000	30.5000	127.25000	846.00
BMI	768.0	31.992578	7.884160	0.000	27.30000	32.0000	36.60000	67.10
DiabetesPedigreeFunction	768.0	0.471876	0.331329	0.078	0.24375	0.3725	0.62625	2.42
Age	768.0	33.240885	11.760232	21.000	24.00000	29.0000	41.00000	81.00
Outcome	768.0	0.348958	0.476951	0.000	0.00000	0.0000	1.00000	1.00

Table 3.5 shows the data statistic of the dataset before pre-processing. In fig 3.12 the data without NaN values are counted before pre-processing.

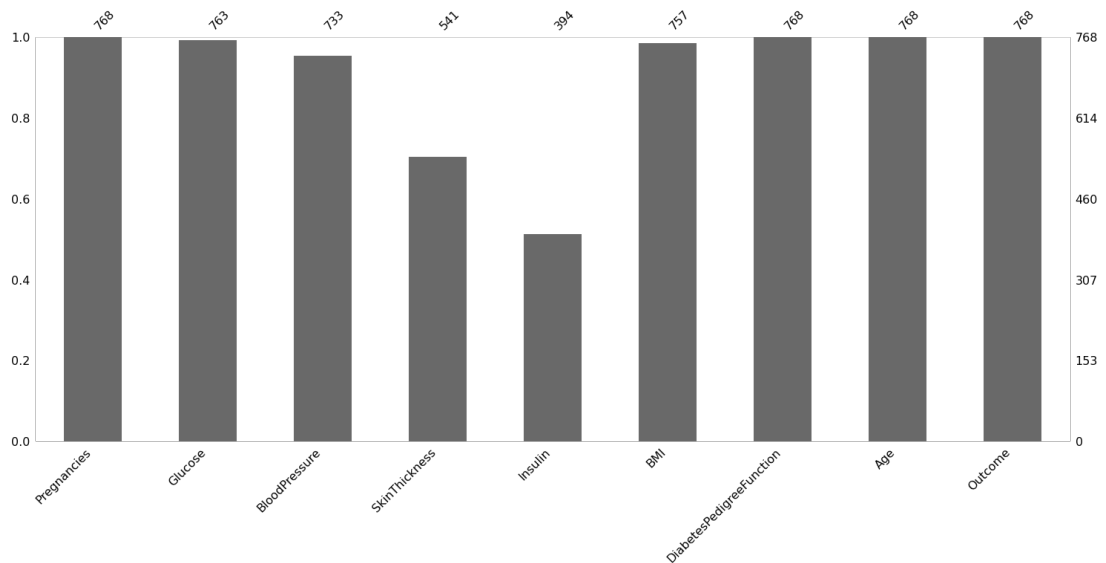


Fig.3.12. Data values count before replacing NaN values.

Table 3.6. NaN values count.

Attribute Name	NaN Values	
	Before Elimination	After Elimination
<i>Pregnancies</i>	0	0
<i>Glucose</i>	5	0
<i>BloodPressure</i>	35	0
<i>SkinThickness</i>	227	0
<i>Insulin</i>	374	0
<i>DiabetesPedigreFunction</i>	0	0
<i>BMI</i>	11	0
<i>Age</i>	0	0
<i>Outcome</i>	0	0

Table 3.7. Data statistics after replacing NaN values.

	count	mean	std	min	25%	50%	75%	max
Pregnancies	768.0	3.845052	3.369578	0.000	1.00000	3.000000	6.00000	17.00
Glucose	768.0	121.686763	30.435949	44.000	99.75000	117.000000	140.25000	199.00
BloodPressure	768.0	72.405184	12.096346	24.000	64.00000	72.202592	80.00000	122.00
SkinThickness	768.0	29.108073	8.791221	7.000	25.00000	29.000000	32.00000	99.00
Insulin	768.0	140.671875	86.383060	14.000	121.50000	125.000000	127.25000	846.00
BMI	768.0	32.455208	6.875177	18.200	27.50000	32.300000	36.60000	67.10
DiabetesPedigreeFunction	768.0	0.471876	0.331329	0.078	0.24375	0.372500	0.62625	2.42
Age	768.0	33.240885	11.760232	21.000	24.00000	29.000000	41.00000	81.00
Outcome	768.0	0.348958	0.476951	0.000	0.00000	0.000000	1.00000	1.00

Table 3.7 shows the data statistic of the dataset before pre-processing. In fig 3.13 the data without NaN values are counted before pre-processing.

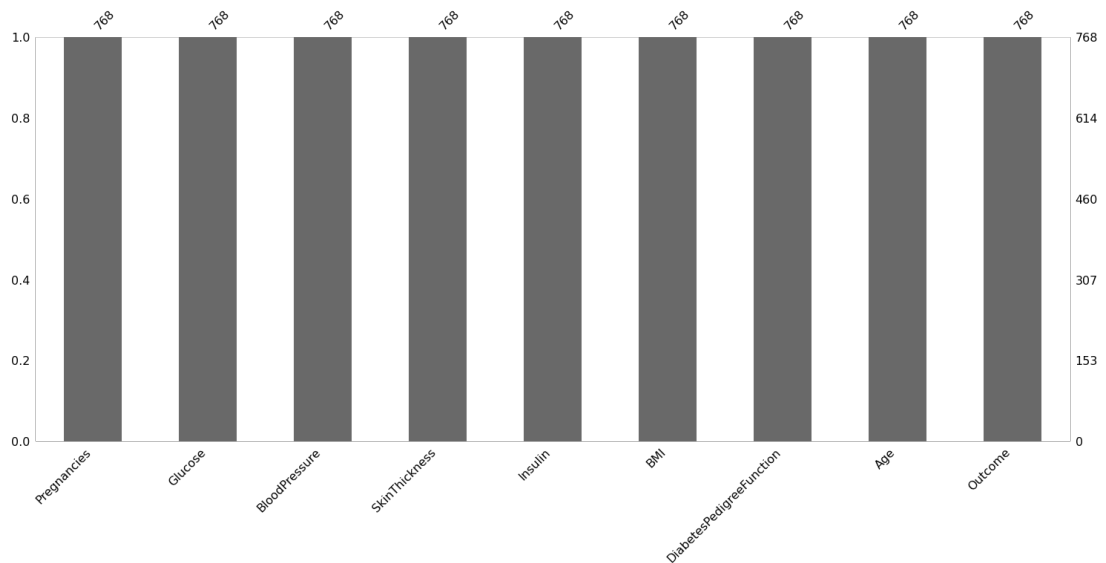


Fig.3.13. Data values count before replacing NaN values.

3.2.2 Evaluation of class distribution.

The data was checked to be distributed evenly between the target variable outcome as shown in table 3.8.

Table 3.8. Verification of data class diistribution.

Dataset Type	Outcome Ditrbutio
Original True	268 (34.90%)
Original False	500 (65.10%)
Training True	179 (34.96%)
Training False	333 (65.04%)
Test True	89 (34.77%)
Test False	167 (65.23%)

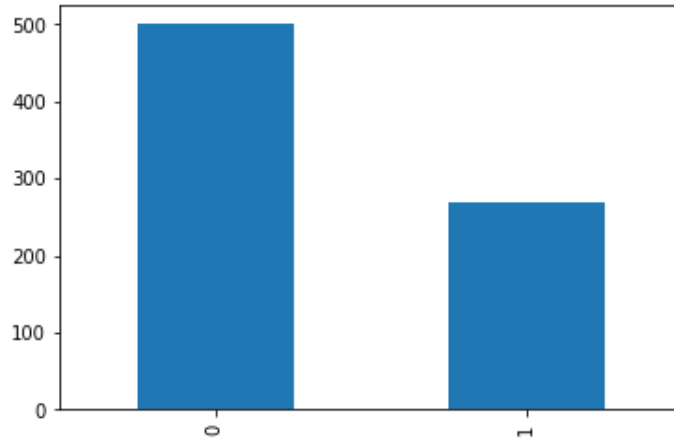


Fig.3.14. Outcome dataset

Fig. 3.14 shows the distribution of outcome attribute in the dataset.

3.2.3 Data scaling

The data is scaled using mean and standard deviation of each attribute,

$$z = \frac{x - \mu}{\sigma}$$

where, μ =mean, σ =std deviation and x =attribute value.

3.3 Data splitting

Entire data was divided into training and testing data as shown in fig3.15. Two-thirds of main dataset was the training data and the rest one-third was used for testing. The training data is the dataset which is given to the model in the beginning for model's

training purpose, which is, to learn from the dataset about the input attributes and the output attribute. The testing data is the dataset which given to the model to model after the training of the model is complete, to check of efficiently the model is working.



Fig.3.15. Data Splitting

So, here the testing data is $1/3^{rd}$ of the whole data set, and the training dataset is $2/3^{rd}$ of the whole dataset, as shown in table 3.9.

Table 3.9. Verification of data splitting.

Dataset	Percentage of data w.r.t original dataset
Dataset before splitting	768 (100%)
Training Dataset	512 (66.6 %)
Testing Dataset	256 (33.3%)

3.4 MACHINE LEARNING METHODOLOGIES

3.4.1 K-Nearest Neighbor

Aim of the algorithm is to find the class for the given input. It is a supervised machine learning algo. K is the number of neighbors with which we will compare the given input. The input will be assigned to the class whose maximum number of data will be near to the input itself. And the calculation is done with the help of Euclidean distance KNN formula:

$$f(x) = \frac{1}{k} + \sum_{x \in N_K(x)} y_i$$

Fig. 3.16. shows the learning curve of KNN model which represents the training score (red line) and cross-validation (green line) of the KNN model. The learning curve

graphically represents the performance score of the model while being trained and while being cross validated. In this study for model validation confusion matrix have been used.

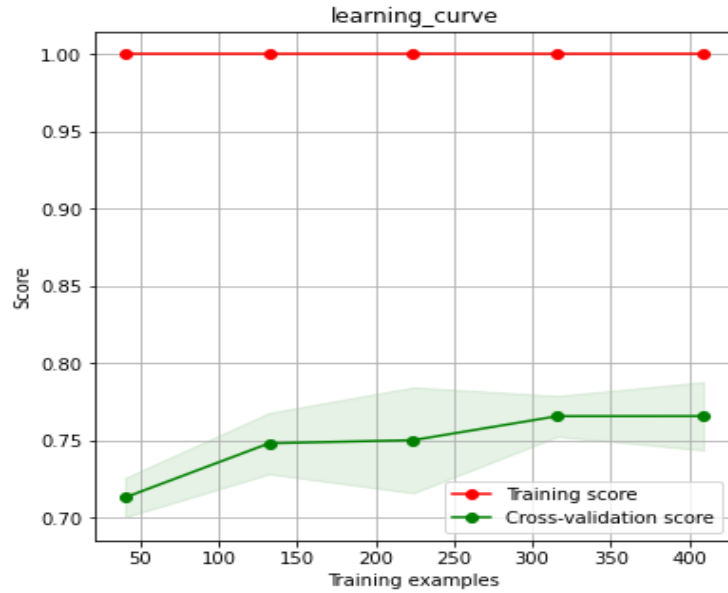


Fig.3.16. Learning curve of KNN.

In fig. 3.17 confusion matrix of KNN model is shown. In which true positive values are 142, true negative values are 53, FN values are 36 and FP values are 25.

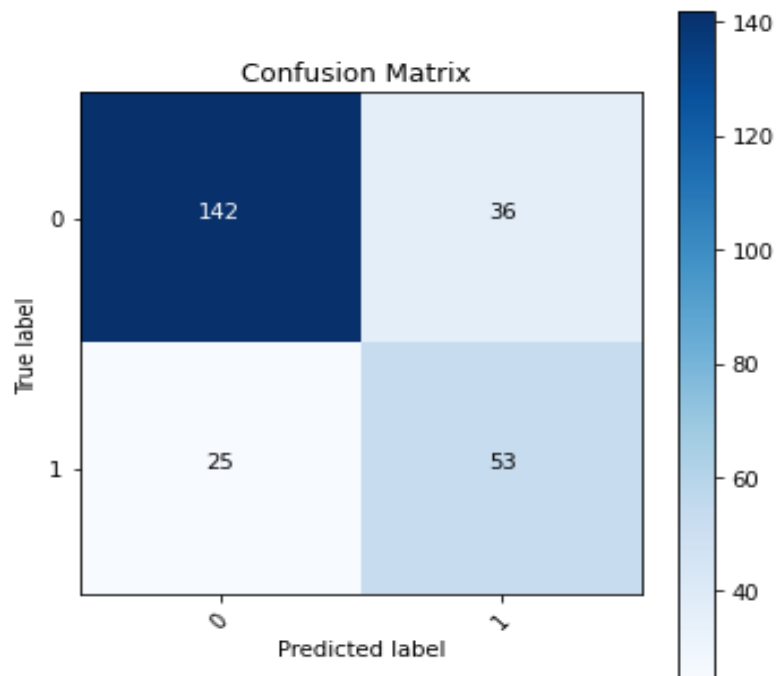


Fig. 3.17 Confusion matrix of KNN.

3.4.2 Support Vector Machine

It is a labelled training data algorithm that creates a hyperplane that separates the points according to their classes. This hyperplane can be seen in 2D space as a plane splitting line into two pieces, one for each segment. Linear SVM is a technique for generating a classifier that can distinguish between labelled datasets. Given two sorts of points, it tries to maximise the margin geometrically. The letter 'Z' is utilised to solve the problem of maximum margin and the reparability limitation.

$$\underset{\gamma, w, b}{\operatorname{argmin}} \frac{1}{2} \|\mathbf{w}\|^2$$
$$s.t. y^j (\mathbf{w}^a x + c) \geq 1, j = 1, 2, \dots, n$$

Fig 3.18 shows the learning curve of support vector machine model which represents the training score (red line) and cross-validation (green line) of the support vector machine model. The learning curve graphically represents the performance score of the model while being trained and while being cross validated. In this study for model validation confusion matrix have been used.

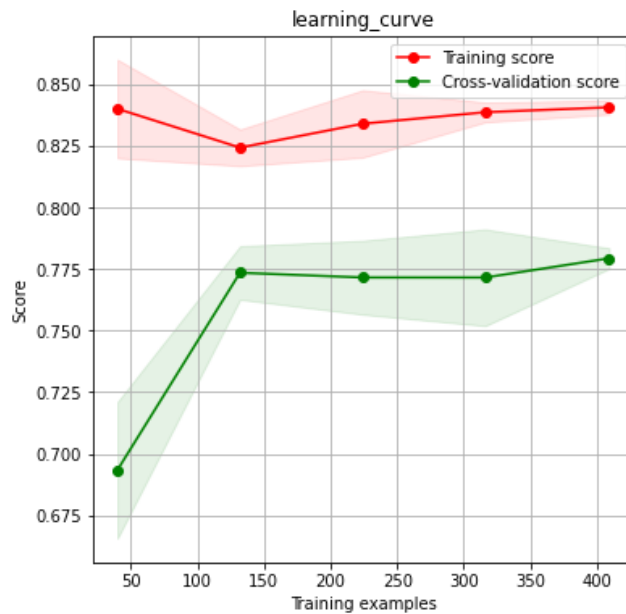


Fig. 3.18 Learning Curve of SVM.

In fig 3.19 confusion matrix of KNN model is shown. In which true positive values are 146, true negative values are 48, FN values are 41 and FP values are 21.

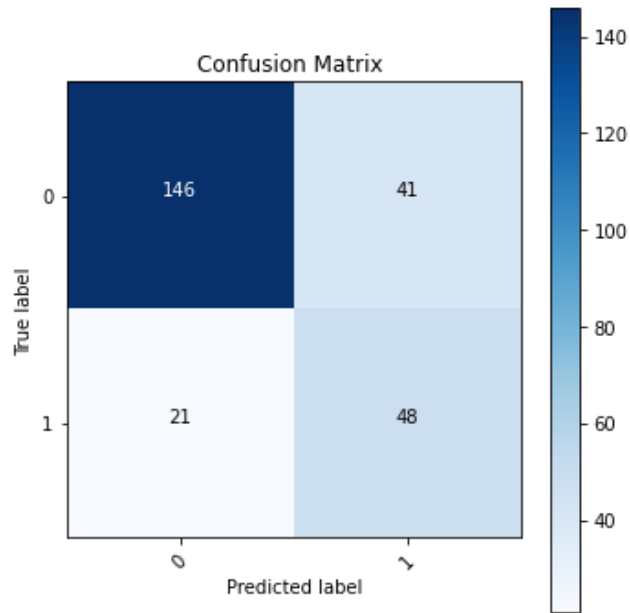


Fig. 3.19 Confusion Matrix of SVM.

3.4.3 Logistic Regression

It's an algorithm for calculating binary outcomes like zero and one (in our case diabetic or nondiabetic). A linear regression is ineffective for categorizing a binary variable because it predicts continuous values that are beyond the range.

$$\frac{p}{(1-p)} = e^{b_0+b_1x} \text{ or } \frac{1}{1+e^{-(b_0+b_1x)}}$$

Fig. 3.20 shows the learning curve of logistic regression model which represents the training score (red line) and cross-validation (green line) of the support vector machine model. The learning curve graphically represents the performance score of the model while being trained and while being cross validated. In this study for model validation confusion matrix have been used.

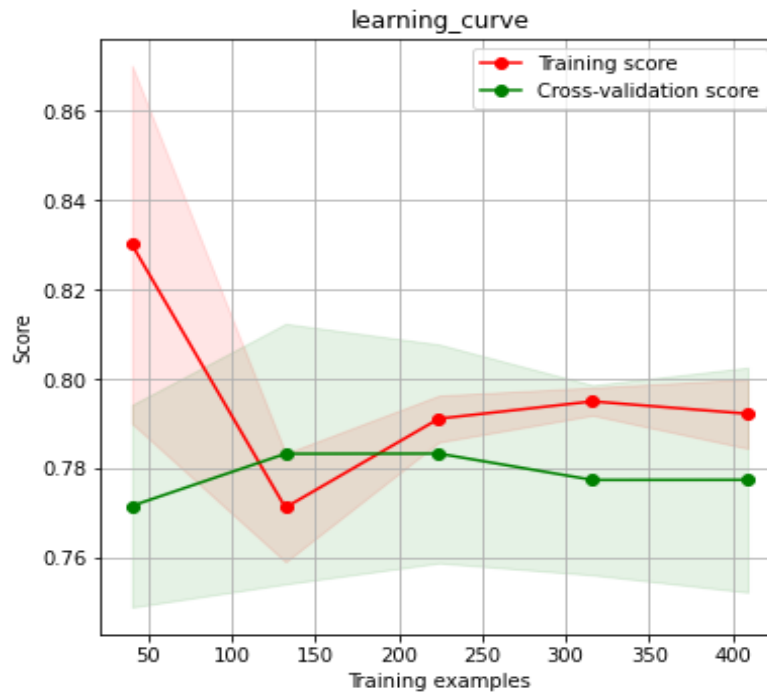


Fig. 3.20 Learning Curve of Logistic Regression.

In fig. 3.21 confusion matrix of KNN model is shown. In which true positive values are 140, true negative values are 47, FN values are 42 and FP values are 27.

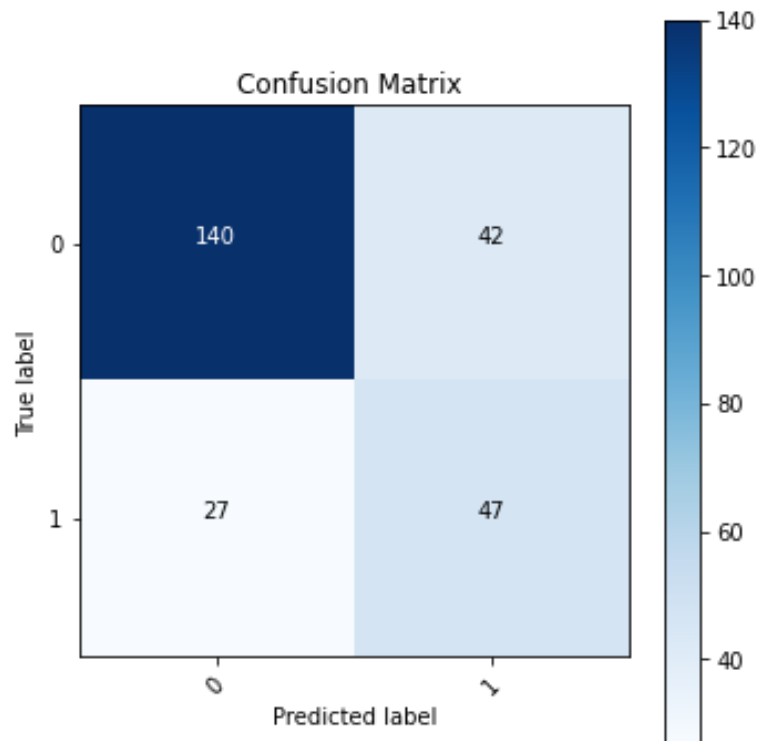
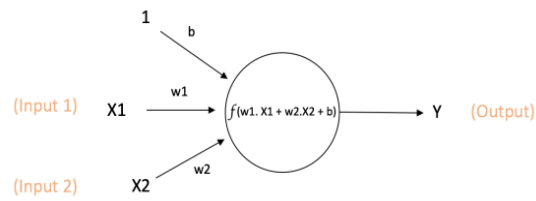


Fig. 3.21 Confusion Matrix of Logistic Regression.

3.4.4 Artificial Neural Networks

The output layer, hidden layer, and input layer are the three layers of an ANN, which are made up of interconnected neurons. The hidden layer has multi-layered structure. The nodes in successive layers are all linked together. Every neuron has an activation function, which is a transformation function that is applied to the node before it is sent to the next layer as input. The result of a node is computed as:



$$\text{Output of neuron} = Y = f(w1.X1 + w2.X2 + b)$$

CHAPTER 4

RESLUTS

In this research, we have performed diabetes prediction on PIMA Indian dataset, to predict diabetes a person is diabetic or not. First data is preprocessed by eliminating all the NaN(Not a number) values, by replacing them by the mean or the median of the respective attribute. Then the prediction was made by using four different ML algorithms KNN, SVM, LR and ANN.

4.1 RESULTS OF KNN

Table 4.1 shows the training score of the KNN model. And table 4.2 shows the testing score of the KNN model.

Table 4.1 Training Score of KNN

Accuracy	1.0
Precision	1.0
Recall	1.0
F score	1.0

Table 4.2 Testing Score of KNN

Accuracy	0.76171875
Precision	0.6794871794871795
Recall	0.5955056179775281
F score	0.6347305389221557

4.2 RESULTS OF SVM

Table 4.3 shows the training score of the SVM model. And table 4.4 the testing score of the SVM model.

Table 4.3 Training Score SVM

Accuracy	0.84375
Precision	0.8613138686131386
Recall	0.659217877094972
F score	0.7468354430379747

Table 4.4 Testing Score SVM

Accuracy	0.7578125
Precision	0.6956521739130435
Recall	0.5393258426966292
F score	0.6075949367088608

4.3 RESULTS OF LR

Table 4.5 shows the training score of the Logistic Regression model. And Table 4.6 shows the testing score of the logistic regression model.

Table 4.5. Training Score of LR

Accuracy	0.787109375
Precision	0.75
Recall	0.5865921787709497
F score	0.6583072100313481

Table 4.6. Testing Score of LR

Accuracy	0.73046875
Precision	0.6351351351351351
Recall	0.5280898876404494
F score	0.5766871165644172

4.4 RESULTS OF ANN

Table 4.7 shows the training score of the artificial neural networks model. And table 4.8 shows the testing score of the logistic regression model.

Table 4.7 Training Scores of Artificial Neural Networks

Accuracy	0.80859375
Precision	0.7454545454545455
Recall	0.6871508379888268
F score	0.7151162790697675

Table 4.8. Training Scores of Artificial Neural Networks

Accuracy	0.71484375
Precision	0.6025641025641025
Recall	0.5280898876404494
F score	0.562874251497006

Among all the four algorithms, KNN showed the best accuracy of 76%. **Table 4.9** Shows all the values of accuracy, precision, recall and f1-score of all the four machine learning algorithms.

Table 4.9. Performance measures of all the four algorithms

	Accuracy	Precision	Recall	F1-score
KNN	0.761718 75	0.6794871794871 795	0.5955056179775 281	0.6347305389221 557
SVM	0.757812 5	0.6956521739130 435	0.5393258426966 292	0.6075949367088 608
Logistic Regressi on	0.730468 75	0.6351351351351 351	0.5280898876404 494	0.5766871165644 172
Neural Network	0.738281 25	0.625	0.6178775280898 876	0.6214689265536 723

CHAPTER 5

CONCLUSION AND FUTURE WORK

In this study we ought to resolve the complications occurred during diagnosis of diabetes disease. The study put forwards an light on different machine learning algorithm such as the SVM , KNN , logistic regression and ANN for predicting whether a patient is diabetic or not .It was concluded that out of all KNN performed best with an accuracy of 76% ,hence, it is a better option for classifying complex data.

In future we'll try to come up with much better mechanisms and a much larger data set in order to increase the accuracy to help medical practitioners to treat patients and overcome this deadly disease. To more increase the accuracy of the models one can use new pre-processing techniques on the dataset.

REFERENCES

1. Gauri D. Kalyankar, Shivananda R. Poojara and Nagaraj V. Dharwadkar,” Predictive Analysis of Diabetic Patient Data Using Machine Learning and Hadoop”, International Conference On I-SMAC,978-1-5090-3243-3,2017.
2. <https://www.google.com/url?sa=i&url=https%3A%2F%2Ftowardsdatascience.com%2Fcousins-of-artificial-intelligence-dda4edc27b55&psig=AOvVaw37H-piLBV9uq2KvwVDdb65&ust=1653979413380000&source=images&cd=vfe&ved=0CAwQjRxqFwoTCNi0j9DPPhvgCFQAAAAAdAAAAABAE>
3. https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.toptal.com%2Fmachine-learning%2Fsemi-supervised-image-classification&psig=AOvVaw37z328ZmsN83msPW-HQubI&ust=1653979207259000&source=images&cd=vfe&ved=0CAwQjRxqFwoTCLjn_OvOhvgCFQAAAAAdAAAAABAD
4. https://www.google.com/url?sa=i&url=https%3A%2F%2Ftowardsdatascience.com%2Fwhat-are-the-types-of-machine-learning-e2b9e5d1756f&psig=AOvVaw2iX9jyNgf_elefHJXCy-nB&ust=1653979356930000&source=images&cd=vfe&ved=0CAwQjRxqFwoTCOjTz7LPhvgCFQAAAAAdAAAAABAD
5. <https://d3e3a9wpte0df0.cloudfront.net/wp-content/uploads/2019/08/Unsupervised-landing-design-scheme-2.jpg>
6. Humar Kahramanli and Novruz Allahverdi,”Design of a Hybrid System for the Diabetes and Heart Disease”, Expert Systems with Applications: An International Journal, Volume 35 Issue 1-2, July, 2008.
7. A. Kumar Dwivedi, “Analysis of computational intelligence techniques for diabetes mellitus prediction,” *Neural Comput. Appl.*, vol. 13, no. 3, pp. 1–9, 2017.
8. J. Vijayashree and J. Jayashree, “ An Expert System for the Diagnosis of Diabetic Patients using Deep Neural Networks and Recursive Feature Elimination,” *International Journal of Civil Engineering and Technology*, vol. 8, pp. 633-641, Dec. 2017.
9. Dost Muhammad Khan¹, Nawaz Mohamudally², “An Integration of K-means and Decision Tree (ID3) towards a more Efficient Data Mining Algorithm ”, *Journal Of Computing*, Volume 3, Issue 12, December 2011.
10. M. Heydari, M. Teimouri, Z. Heshmati, and S. M. Alavinia, "Comparison of various classification algorithms in the diagnosis of type diabetes in Iran," *International Journal of Diabetes in Developing Countries*, pp. 1-7, 2015.
11. H. Temurtas, N. Yumusak, and F. Temurtas, “A comparative study on diabetes disease diagnosis using neural networks,” *Expert Syst. Appl.*, vol. 36, no. 4, pp. 8610–8615, 2009, doi: 10.1016/j.eswa.2008.10.032.
12. J. Vijayashree and J. Jayashree, “ An Expert System for the Diagnosis of Diabetic Patients using Deep Neural Networks and Recursive Feature Elimination,” *International Journal of Civil Engineering and Technology*, vol. 8, pp. 633-641, Dec. 2017.
13. M. Kumari, Dr. R. Vohra, and A. Arora, “Prediction of Diabetes using Bayesian Network,” *International Journal of Computer Science and Information Technologies*, vol. 5, pp. 5174-5178, 2014.
14. K. Rajesh and V. Sangeetha, “Application of Data Mining Methods and Techniques for Diabetes Diagnosis”, *International Journal of Engineering and Innovative Technology (IJEIT)* Volume 2, Issue 3, September 2012.

15. Mani Butwall and Shraddha Kumar, "A Data Mining Approach for the Diagnosis of Diabetes Mellitus using Random Forest Classifier", *International Journal of Computer Applications*, Volume 120 - Number 8, 2015.
16. L. B. Goncalves and M. M. Bernardes, "Inverted Hierarchical Neuro-Fuzzy BSP System: A Novel Neuro-Fuzzy Model for Pattern Classification and Rule Extraction in Databases," in *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 36, no. 2, pp. 236- 248, Mar. 2006.
17. A. Ashiquzzaman, A. K. Tushar, M. Islam, J.-M. Kim et al., "Reduction of overfitting in diabetes prediction using deep learning neural network," arXiv preprint arXiv:1707.08386, 2017.
18. J. Zhu, Q. Xie, K. Zheng. "An Improved Early Detection Method of Type-2 Diabetes Mellitus Using Multiple Classifier Systems". *Information Sciences*, volume 292, pages 1-14, 2015.
19. B.M. Patil, R.C. Joshi and Durga Toshniwal, "Association Rule for Classification of Type-2 Diabetic Patients", *ICMLC '10 Proceedings of the 2010 Second International Conference on Machine Learning and Computing*, February 09 - 11, 2010.
20. M. Fatima and M. Pasha, "Survey of Machine Learning Algorithms for Disease Diagnostic," *J. Intell. Learn. Syst. Appl.*, vol. 09, no. 01, pp. 1-16, 2017, doi:10.4236/jilsa.2017.91001.