Project Report

on

DIABETES PREDICTION USING MACHINE LEARNING

Submitted in partial fulfillment of the
requirement for the award of the degree of

Bachelor of Technology in
Computer Science

GALGOTIAS
UNIVERSITY
(Established under Galgotias University Uttar Pradesh Act No. 14 of 2011)

Under The Supervision of
DR.NITIN MISHRA
AP

Submitted By

Aayushi Saxena
18021011437/18SCSE1010194
Amanya Verma
18021011742/18SCSE1010514

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA
INDIA
MAY, 2022

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA

CANDIDATE’S DECLARATION

I/We hereby certify that the work which is being presented in the project, entitled **“DIABETES PREDICTION USING MACHINE LEARNING”** in partial fulfillment of the requirements for the award of the B.Tech submitted in the School of Computing Science and Engineering of Galgotias University, Greater Noida, is an original work carried out during the period of January, 2022 to May 2022, under the supervision of DR.NITIN MISHRA Associate Professor, Department of Computer Science and Engineering, of School of Computing Science and Engineering , Galgotias University, Greater Noida

The matter presented in the project has not been submitted by me for the award of any other degree of this or any other places.

AAYUSHI SAXENA, 18SCSE1010194

AMANYA VERMA, 18SCSE1010514

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

Nitin

SUPERVISIOR NAME:

ASSOCIATE PROFESSOR

Designation

CERTIFICATE

The Final Project Dissertation Viva-Voce examination of ETE has been held on _**13/5/2022**_ and his/her work is recommended for the exam result for semester 8~~.~~

Signature of Examiner **Signature of Supervisor**

Signature of Project Coordinator **Signature of Dean**

Date: May, 2022

Place: Greater Noida

Abstract

Nowadays most of the people is suffering from diabetes. Diabetes is a chronic disease or group of metabolic disease where a person suffers from an extended level of blood glucose in the body, which is either the insulin production is inadequate, or because the body's cells do not respond properly to insulin. The constant hyperglycemia of diabetes is related to long-haul harm, brokenness, and failure of various organs, particularly the eyes, kidneys, nerves, heart, and veins. The objective of this research is to make use of significant features, design a prediction algorithm using Machine learning and find the optimal classifier to give the closest result comparing to clinical outcomes which will help in detection of diabetes in the patients before it becomes fatal.

The proposed system focuses using algorithms combinations shown above in the block diagram The base classification algorithms are: Decision tree, Random forest, Support Vector Machine, Logistic Regression, KNN for accuracy authentication.

Here we are using Machine Learning Algorithms to predict the data and the algorithms we will use are: Decision Tree, Random Forest, Logistic Regression , SVM Algorithm, KNN Algorithm

The proposed approach will use different classification and ensemble methods and implemented using python. These methods will be standard Machine Learning methods used to obtain the best accuracy from data. Overall we will use best Machine Learning techniques for prediction and to achieve high performance accuracy.The main aim of this project is to design and implement Diabetes Prediction Using Machine Learning Methods and Performance Analysis of that methods. It uses various classification and ensemble learning method in which SVM, Knn, Random Forest, Decision Tree, Logistic Regression used.

Table of Contents

List of Table

List of Figures

Acronyms

B.Tech.	Bachelor of Technology
M.Tech.	Master of Technology
BCA	Bachelor of Computer Applications
MCA	Master of Computer Applications
B.Sc. (CS)	Bachelor of Science in Computer Science
M.Sc. (CS)	Master of Science in Computer Science
SCSE	School of Computing Science and Engineering

CHAPTER-1

Introduction

Diabetes or DM (DM) could also be a collection of metabolic issues known by high glucose levels over a drawn-out period of our time. polygenic disorder (DM) is printed as a bunch of metabolic disorders within the main caused by abnormal endocrine secretion and/or action . Symptoms of high aldose incorporate excessive excreting, frequently feeling thirsty and enlarged hunger . If not treated on time, polygenic disorder can cause serious health issues during a person like acidosis, hyperosmolar hyperglycemic state, or even end in death. this might end in period complications similarly as vas upset, brain stroke, failure, ulcers among the foot, and eye complications . polygenic disorder is caused once the secreter among the body is unable to return up with endocrine in enough amounts or once the cells and tissues among the body fail to utilize the insulin created. polygenic disorder exists in three forms: DM Type-1 is characterized by secreter generating endocrine however what's required by the body, a condition together cited as "insulin-subordinate polygenic disorder mellitus" (IDDM).Folks cluttered with type-1 DM want external endocrine amount to create up for the less endocrine created by the secreter. Diabetes Mellitus Type-2 is marked by the body resisting endocrine because of the body cells react otherwise to endocrine than they ancient would. this might ultimately end in no endocrine among the body. this will be otherwise cited as "non-insulin subordinate diabetes mellitus" (NIDDM) or "adult starting polygenic disorder". this type of polygenic disorder is commonly found in of us with high BMI or folks who lead degree inactive manner. physiological condition polygenic disorder is that the third principle structure that is observed throughout physiological condition. Generally, for a conventional person, aldose levels vary from seventy to 99 milligrams per dl. a private is taken into consideration diabetic providing the quick aldose level is found to be over 126 mg/dL. among the observe, a private having Associate in Nursing aldose concentration of 100 to at least one hundred twenty-five mg/dL is taken into consideration as pre-diabetic .

Such a private is vulnerable to the event of kind 2 polygenic disorder. Over the years, it's been found that people with the following health characteristics face a bigger risk

•A Body Mass Index value larger than 25

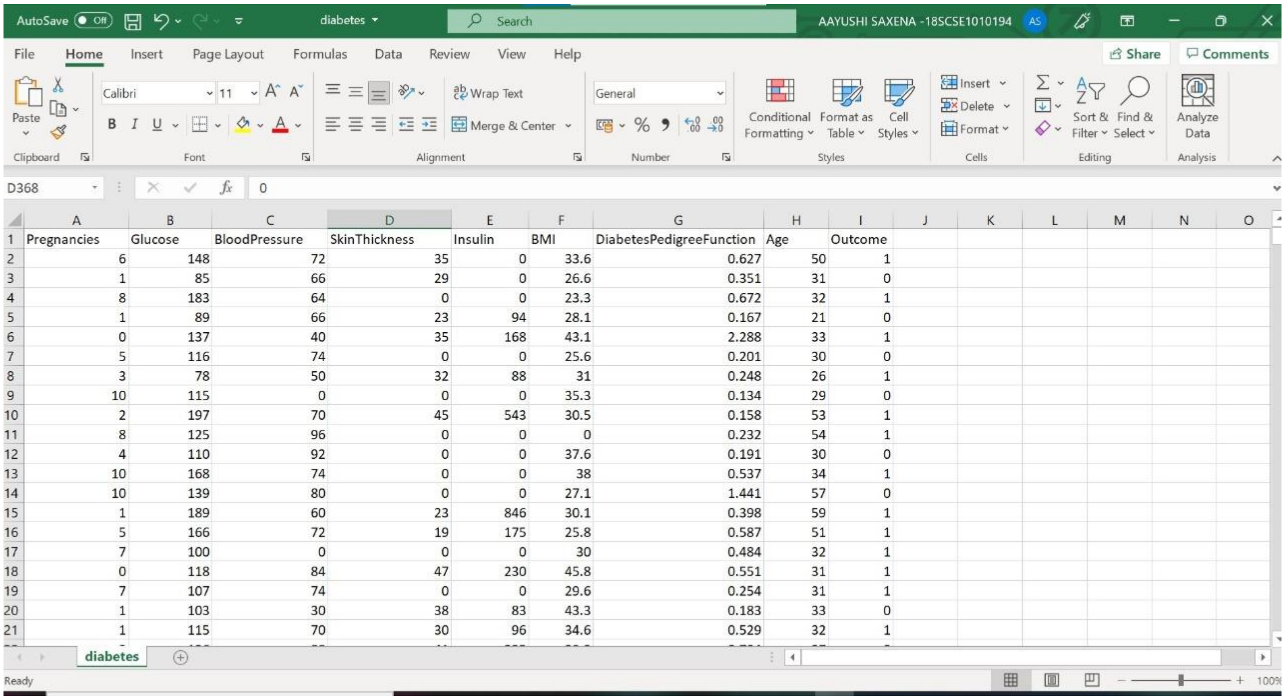
•Members of the family full of polygenic disease

•People World Health Organization have suffered from polycystic ovary disorder at intervals the past.

• People happiness to ethnic groups like African yank, or Native yank ,or Spanish yank, or Asian-pacific over 45 years.

•Having associate degree inactive manner

When a doctor diagnoses that someone has prediabetes, they suggest the individual higher their manner. Adopting fitness regime associates degreed an honest diet prepare can facilitate forestall polygenic disease. This analysis aims to figure out the danger of development of polygenic disease in an exceedingly person. thus during this study, we tend to used supply Regression, K- Nearest Neighbors ,Decision Trees, Random Forest and SVM machine learning classification algorithms square measure used and evaluated on the PIDD dataset to seek out the prediction of polygenic disease throughout a patient. Experimental performance of all the half-dozen algorithms is compared on numerous measures and achieved smart accuracy . This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.



D368 | 0

	A	B	C	D	E	F	G	H	I
1	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
2	6	148	72	35	0	33.6	0.627	50	1
3	1	85	66	29	0	26.6	0.351	31	0
4	8	183	64	0	0	23.3	0.672	32	1
5	1	89	66	23	94	28.1	0.167	21	0
6	0	137	40	35	168	43.1	2.288	33	1
7	5	116	74	0	0	25.6	0.201	30	0
8	3	78	50	32	88	31	0.248	26	1
9	10	115	0	0	0	35.3	0.134	29	0
10	2	197	70	45	543	30.5	0.158	53	1
11	8	125	96	0	0	0	0.232	54	1
12	4	110	92	0	0	37.6	0.191	30	0
13	10	168	74	0	0	38	0.537	34	1
14	10	139	80	0	0	27.1	1.441	57	0
15	1	189	60	23	846	30.1	0.398	59	1
16	5	166	72	19	175	25.8	0.587	51	1
17	7	100	0	0	0	30	0.484	32	1
18	0	118	84	47	230	45.8	0.551	31	1
19	7	107	74	0	0	29.6	0.254	31	1
20	1	103	30	38	83	43.3	0.183	33	0
21	1	115	70	30	96	34.6	0.529	32	1

FIG-1.1 DATASET FROM PIMA INDIAN HERITAGE AVALIABLE ON KAGGLE

The dataset contains seven sixty eight instances and nine features. The dataset features are:

1. Total number of times pregnant	2. Glucose/sugar level
3. Diastolic blood pressure	4. Body mass index
5. Skin fold thickness in mm	6. Insulin value in 2 hour
7. Hereditary factor	8. Age of patient

TABLE 1.1- DATASET FEATURES

According to World Health Organization (WHO), diabetes affects 8.5% of adults over the age of 18 and is also responsible for 1.6 million deaths worldwide(2012). According to (WHO) World Health Organization about 422 million people suffer from diabetes, particularly from low or idle-income countries. And this could be increased to 490 billion up to the year of 2030Diabetes caused because of obesity or high blood glucose level, and so forth. It affects the hormone insulin, resulting in abnormal metabolism of crabs and improving the level of sugar in the blood. Diabetes occurs when the body does not make enough insulin.

However, the prevalence of diabetes is found in various countries like Canada, China, India, etc. The population of India is now more than 100 million so the actual number of diabetics in India is 40 million. Diabetes is the major cause of death in the world. Early prediction of diseases like diabetes can be controlled and save human life. To accomplish this, this work explores the prediction of diabetes by taking various attributes related to diabetes disease. For this purpose we use the Pima Indian Diabetes Dataset, we apply various Machine Learning classification and ensemble Techniques to predict diabetes. The primary source of energy for our body is glucose. It comes from the food we eat.

Insulin is produced in the pancreas beta cells. But when the pancreas beta cells become weak, or when our body becomes insulin resistant, Insulin deficiency occurs. The glucose does not reach the cells and stays in our blood Thus, increasing the glucose level. It can be reversed by engaging in physical activities like walking, running, or others and adopting a nutrient-rich diet.

To know diabetes, we need to understand Blood Glucose Action

The primary source of energy for our body is glucose. It comes from the diet we take. Insulin is a hormone present in our body that helps absorb glucose by our body cells. Glucose enters the cells with the help of insulin and is further used for energy generation and body metabolism.

Insulin is produced by the pancreas beta cells. But when the pancreas beta cells become weak, or when our body becomes insulin resistant, insulin deficiency occurs. Due to insulin deficiency, the glucose does not reach the cells and stays in our blood. As a result, it causes an increase in blood glucose levels. Insulin Deficiency can be reversed by engaging in physical activities and adopting a nutrient-rich diet.

When the blood glucose levels, also known as blood sugar levels in the body, are too high, it causes diabetes.

Normal Blood Sugar Levels of An Adult:

1. 72-99 mg/dl before eating
2. 80-133 mg/dl after eating

High Blood Sugar Levels significantly above 140 mg/dl cause diabetes.

Types of Diabetes:

Type-1 Diabetes: in this type of diabetes our immune system attacks the beta cells of our pancreas. This produces insulin in the pancreas. However, the immune system attacks the beta cells and destroys them and the Beta cells become weak. Thus, failing to produce enough insulin.

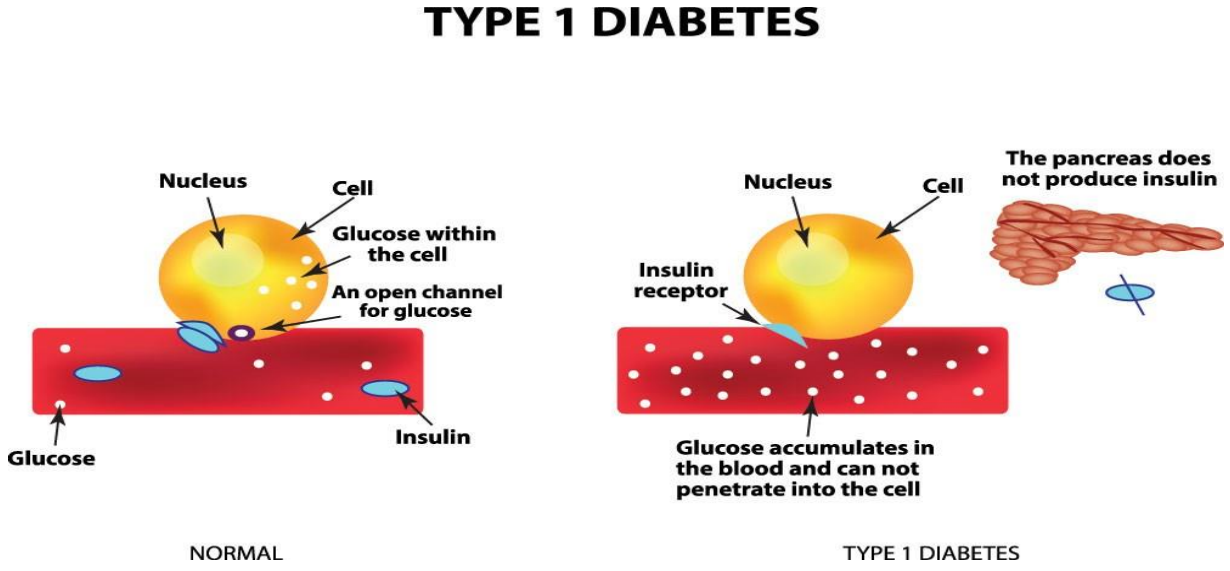

TYPE 1 DIABETES
Nucleus
Cell
Glucose within the cell
An open channel for glucose
Glucose
Insulin
Nucleus
Cell
The pancreas does not produce insulin
Insulin receptor
Glucose accumulates in the blood and can not penetrate into the cell
NORMAL
TYPE 1 DIABETES

Fig- 1.2 How type-1 diabetes occur

Type-1 diabetes causes are mainly genetic, wherein the immune system attacks insulin-reducing cells. Exposure to viruses and chemical toxins can also cause type 1 diabetes.

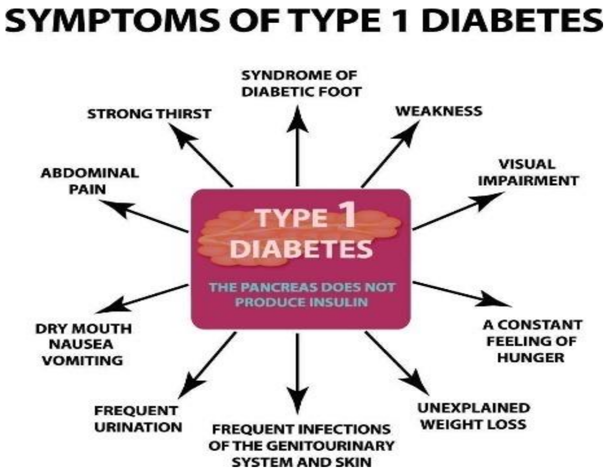

SYMPTOMS OF TYPE 1 DIABETES
SYNDROME OF DIABETIC FOOT
STRONG THIRST
WEAKNESS
ABDOMINAL PAIN
VISUAL IMPAIRMENT
TYPE 1 DIABETES
THE PANCREAS DOES NOT PRODUCE INSULIN
DRY MOUTH NAUSEA VOMITING
A CONSTANT FEELING OF HUNGER
FREQUENT URINATION
FREQUENT INFECTIONS OF THE GENITOURINARY SYSTEM AND SKIN
UNEXPLAINED WEIGHT LOSS

Fig-1.3 Symptoms of type 1 Diabetes

If Type-1 diabetes is left untreated it can cause vision loss, nerve damage foot pain and, in worse case scenarios leg amputation is also needed, Healing of wounds would become slow, heart diseases, kidney failure etc.

Type-2 Diabetes: In Type -2 Diabetes, the body cells become resistant to insulin which then decreases glucose absorption by the cells. The high blood glucose levels cause pressure on increasing insulin production and therefore, weaken the pancreatic cells and the body becomes insulin resistant. Poor Lifestyle, Poor eating choices, family history, age, obesity are the main causes of Type-2 Diabetes.

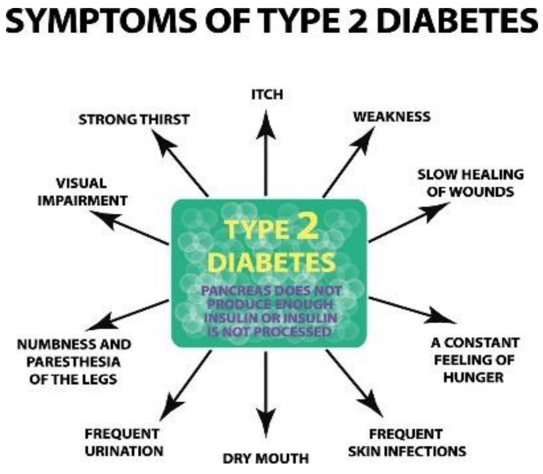

SYMPTOMS OF TYPE 2 DIABETES
ITCH
STRONG THIRST
WEAKNESS
VISUAL IMPAIRMENT
SLOW HEALING OF WOUNDS
TYPE 2 DIABETES
PANCREAS DOES NOT PRODUCE ENOUGH INSULIN OR INSULIN IS NOT PROCESSED
NUMBNESS AND PARESTHESIA OF THE LEGS
A CONSTANT FEELING OF HUNGER
FREQUENT URINATION
DRY MOUTH
FREQUENT SKIN INFECTIONS

Fig-1.4 Symptoms of type-2 diabetes

If Type-2 Diabetes is left untreated following complication arise: 32% of people with type-2 diabetes are prone to cardiovascular risks like heart attacks.

29% of type 2 diabetic patients have stage 3 diabetic ketoacidosis complications, Hearing Impairments, Memory loss etc.

Chronic Type 2 Blood Sugar Symptoms Are:

- Skin infections mainly yeast infections
- Numbness in the feet
- Dark skin patches under armpit, neck, thighs
- Foot pain, slow healing wounds, and nerve damage.

Type 2 Blood Sugar Symptoms In Women

- Painful Urination
- Irregular Periods
- Weight Gain
- Depression

Type 2 Blood Sugar Symptoms In Men

- Skin Infection
- Vision Loss
- Slow Healing Wounds.
- Erectile Dysfunction.
- Frequent Urination.

The following tests help in the diagnosis of type-1 and type-2 Diabetes:

- **A1C (Glycated Haemoglobin test):** It is a blood test that checks the blood glucose level average for the past three months. If the A1C value > 6.5, it is an indication of diabetes.

- **Random blood sugar test:** A blood sample can be taken at any time. If the blood sugar value is > 200mg/dL and the patient has signs of frequent urination, thirst, etc., it indicates diabetes.
- **Fasting blood sugar test:** For this test, a blood sample from the patient is taken after an overnight fast. Blood glucose levels > 126mg/dL indicates diabetes. Generally, these tests are repeated twice or thrice for the confirmation of the diseases.

DKA Or Diabetic Ketoacidosis: Ketoacidosis diabetes occurs when insulin level gets low, and the body starts to burn fat for energy. DKA is a severe complication of type-1 diabetes. But it can also occur in patients with type 2 diabetes. Due to the low insulin level, extreme ketones produce and cause the blood to become more acidic.

The liver releases ketones after burning body fat for generating energy when insulin is deficient. Therefore, a high level of ketones leads to DKA.

Ketones levels > 1.6 mmol/L for a longer time indicate Diabetic ketoacidosis.

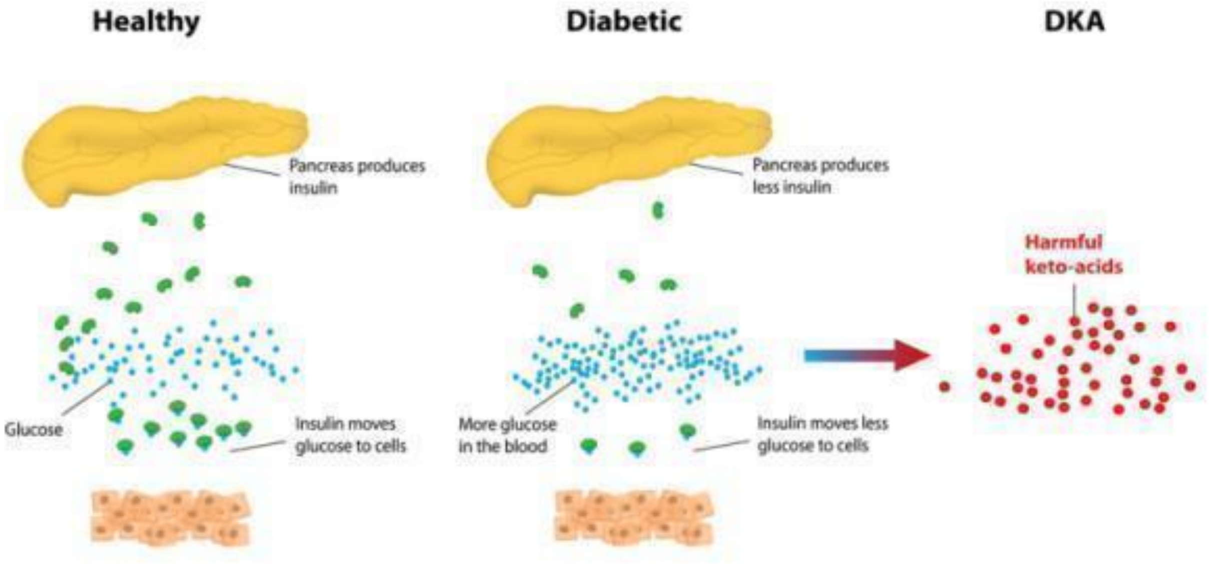

Healthy
Diabetic
DKA
Pancreas produces insulin
Pancreas produces less insulin
Harmful keto-acids
Glucose
Insulin moves glucose to cells
More glucose in the blood
Insulin moves less glucose to cells

Fig- 1.5 Diagrammatic representation of DKA

Causes Diabetic Ketoacidosis: Hormone disorders, diuretics and corticosteroids, alcohol consumption, pancreatic, diseases like pneumonia, and urine infection all can trigger DKA.

Diabetic Ketoacidosis Symptoms

- High blood sugar levels (hyperglycemia)
- High ketone levels in the urine
- Rapid and fruity-smelling breath
- Nausea or Vomiting.
- Muscle Aches.
- Headaches.
- Weakness.
- Belly Pain

- Shortness of breath
- Confusion

Having ketone levels in the range of 1.0 to 3.0 or > 3.0 for a longer time.

Why Should You Worry About DKA?

DKA can be a life-threatening condition. It can cause kidney damage and renal failure. If left untreated for a longer time, diabetic ketoacidosis can lead to coma or death.

Diagnosis Of Diabetes Ketoacidosis:

These blood tests check for blood glucose levels, blood acidity levels, and ketones levels.

Blood sugar levels above 200 mm/dL, Ketones levels > 3.0, and high blood acidity levels indicate diabetes Ketoacidosis.

Gestational Diabetes: Gestational diabetes occurs during pregnancy in some women when their blood sugar levels go high due to pregnancy.

It usually develops in the 24th week of pregnancy. Thus, doctors recommend blood glucose level tests during 24th to 28th week of pregnancy to rule out or diagnose gestational diabetes.

Blood Glucose Meters
143
mg/dL
condition in which blood sugar (glucose) levels become high during pregnancy and usually disappears after giving birth
Gestational Diabetes

Fig-1.6 What Happens during pregnancy

Causes Of Gestational Diabetes Symptoms

Weight gain during pregnancy, family history of diabetes, and hormonal changes are the leading causes of Gestational diabetes.

Gestational Diabetes Symptoms

- Weight gain
- Frequent urination
- Unusual thirst
- Sugar in urine
- Blurred vision
- Fatigue
- Skin infection, especially vaginal and bladder infection

Risks Associated With Gestational Diabetes When Left Untreated:

- Mothers who have gestational diabetes are at higher risk of having C-section delivery.
- Women with gestational diabetes are at higher risk of delivering an obese baby.
- Gestational diabetes can cause future type 2 diabetes in such women in future. It can thus cause type 2 diabetes complications when left unmanaged.
- Untreated, gestational diabetes can lead to premature delivery and can cause baby death after birth.
- Gestational diabetes, when left unmanaged, can be the cause of breathing problems in babies.

Diagnosis Of Gestational Diabetes:

To diagnose gestational diabetes, doctors generally ask for a blood glucose level test. These tests are conducted around the 24th to 28th week of pregnancy. But if you have a family history of diabetes, doctors may carry out these tests at the early stages of pregnancy.

An oral glucose tolerance test is used to diagnose gestational diabetes. In the oral glucose tolerance test, first, a blood sample is drawn after 1hour of eating. If the blood glucose levels are > 140mg/DL, then a 3-hour test is recommended. Glucose syrup is given, and after 3 hours of taking that glucose syrup, the blood glucose levels are repeated. Any value > 140mg/Dl indicates gestational diabetes.

Doctors shall repeat these tests in the 4th week to confirm gestational diabetes. If the values of the blood glucose levels are still high and > 140 mg/dL, it is diagnosed as gestational diabetes.

Early detection of Diabetes will lead to an early reversal of Diabetes. One should not ignore early symptoms and not hesitate to take expert advice. Health coaches along with Diabetologists will assist you during the Diabetes Reversal Program. Regular follow-ups and weekly challenges without disturbing your daily schedule will help you normalize blood sugar levels in the comfort of your house

Hypoglycemia: Hypoglycaemia is a medical condition where the blood glucose level falls below the normal levels. It is also known as low glucose condition. If left untreated it can lead to severe health problems like coma.

When the blood sugar levels fall below 70mg//dL it is an indication of Hypoglycaemia or low sugar levels.

Symptoms Of Hypoglycemia Are:

- Hunger,
- Disorientation,

    - Too much sweating,
    - Anxiety,
    - Feeling sleepy and dull,
    - Sweating,
    - Irregular heartbeat,
    - Confusion,
    - Hand tremor.
- This is a serious condition. When left untreated it can lead to seizures and coma. Hypoglycaemia is a life-threatening disease. One in every 25 patients suffering from type 1 diabetes dies due to severe hypoglycaemia.

- Treatment Of Hypoglycemia
- Eat a few grams (15) of food like a fast-acting carbohydrate. Due to fast-acting carbohydrates, blood glucose levels increase. Consulting a Diet Coach helps you know about fast-acting carbohydrate foods available in your kitchen. Diet Coach also tells the quantity one should consume and how to maintain healthy blood sugar level throughout life naturally..

Diabetes is a disease that is manageable through proper lifestyle, weight loss programs, diet control, and proper treatment. Some guidelines to help you to manage diabetes efficiently are as follows:

- Test your blood glucose level regularly. You can do this at home through glucometers.
- Maintain a proper diabetic diet chart and a healthy lifestyle.
- Do exercise regularly to manage body weight.
- Take a low-carb diet to maintain proper weight.
- Eat food that is high in fiber. As a result of it, you shall have proper weight management.
- Eat more fruits, vegetables, and grains.
- Lose weight if you are overweight.
- If you have diabetes, you should cut more oily food from your diet to control weight gain.
- Avoid Smoking.
- Drink lots of water.
- Don’t walk barefoot as it may cause damage to foot skin and cause foot pain and foot ulcers.
- Avoid processed food and juices.
- Avoid Sugary drinks, and sugar syrup as they may cause hyperglycemia.
- Always keep with you an apple or a few raisins if you have complaints of hypoglycemia.

- Measure your sugar levels, especially after workouts, stress, and exercises.
- Don’t ignore any signs or symptoms that can cause complications. As a result, the situation may become life-threatening.
- Don’t take self-medication. Always consult the doctor before taking any medicine. Also, consult a doctor in case of any issues.
- Avoid taking an overdose. Due to this you may create health issues.

Our project help in diagnosis of diabetes or we can say we predict diabetes in patients by using Data collected from them and finding the accurate machine learning algorithm which will give us the best results. Machine Learning Is a method that is used to train computers. 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., however it’s tough to choose best technique. Thus for this purpose we apply popular classification and ensemble methods on dataset for prediction.

CHAPTER-2

LITERATURE SURVEY

In this section, some closely related works are discussed briefly. In most of the reseach works, Pima Indians Diabetes Dataset (PIDD) have been used by many researchers for diabetes prediction. Various supervised machine learning algorithms were used to predict diabetes (Kaur & Kumari, 2020). Radial basis function (RBF) kernel SVM, artificial neural network (ANN), multifactor dimensionality reduction (MDR), linear SVM and k-NN are some of them to mention. Based on p value and odds ratio (OR), Logistic Regression (LR) has been used to recognize the risk factors for diabetes (Maniruzzaman et al., 2020). Four classifiers have been adopted to predict diabetic patients, such as NB, DT, Adaboost, and RF. Partition protocols like- K2, K5, and K10 were also adopted, repeating these protocols into 20 trails. For the performance measurement of the classifiers, accuracy (ACC) and area under the curve (AUC) were analyzed.

Kopitar et al. (2020) showed a comparison of widely utilized regression models such as Glmnet, RF, XGBoost, LightGBM for predicting type 2 diabetes mellitus. The goal of this work was to examine if innovative machine learning methodologies gave any advantages in early prediction of impaired fast glucose and fasting plasma glucose (FPGL) levels compared to classic regression techniques.

For the prediction of diabetic patients, Maniruzzaman et al. (2020) have chosen four classifications such as naive bays (NB), decision tree (DT), adaboost and random forest. These methods were also implemented by three types of partition protocols (K2, K5, and K10). These classifers' performances are measured with precision (ACC) and curve surface (AUC).

A hybrid model to detect type 2 diabetes was suggested by Albahli (2020). In order to extract unknown, hidden property from the dataset and to obtain more exact results, we use K-mean clustering, which is followed by the execution of a Random Forest and XGBoost classifier.

Yahyaoui et al. (2019) suggested a Machine Learning Techniques (ML) DSS for anticipating diabetes. They compared traditional machine learning with approaches to the deep learning. The authors applied the classifiers most typically used for a standard machine learning method: SVM and the Random Forest (RF). In contrast, they used a full-scale neural network (CNN) for Deep Learning (DL) to forecast and identify patients who suffer from diabetes.

Zou et al. (2018) predicted diabetes using the decision tree, random forests, and neural network. The dataset is collected from the Luzhou physical exams in China. The PCA was applied to reduce the dimension of the dataset. They selected several ML approaches to execute independent test to verify the universal applicability of method.

Supervised machine learning models which explore data-driven approaches were used to identify patients with diabetes diseases (Dinh et al., 2019). A complete research was conducted based on the National Health and Nutrition Examination Survey (NHANES) dataset. To develop models for cardiovascular, prediabetes, and diabetes detection, they have used all available feature variables within the data. Using various time frames and set of features within the data, different machine learning models, namely Support Vector Machines, logistic regression, gradient boosting and random forest were evaluated for the classification.

In Choubey et al. (2017) the authors used NBs for the classification on all the attributes. Afterwards GA was used as an attribute selector and NBs used the selected attributes for classification. The experimental results show the performance of this work on PIDD and provide better classification for diagnosis. Three specific supervised machine learning methods are used by Joshi and Chawan (2018), namely SVM, Logistic regression and ANN. His goal for research was to predict diabetes patients and he has also proposed an effective model for the prior detection of diabetes disease. Rajeswari and Prabhu (2019) focused on machine learning classification algorithms for predicting diabetes disease with more accuracy. Their study in SVM classification algorithm achieved highest accuracy. Various measures have been used to calculate the performance of classification algorithms.

An intelligent model using machine learning practices is developed (Nilashi et al., 2017) to identify diabetes disease. This model is constructed using approaches like clustering, removal of noise and classification, each of which made use of SOM, PCA and NN, respectively. The adaboost and bagging ensemble techniques are used to detect diabetes (Perveen et al., 2016). Along with standalone data mining technique, a base learner is used to identify patients with diabetes mellitus, namely J48 (c4.5) decision tree that makes use of multiple diabetes risk factors. In the Canadian Primary Care Sentinel Surveillance Network, three different ordinal adult groups are selected for classification. Experimental result shows that, the adaboost ensemble method shows better performance than both bagging and standalone J48 decision tree. For diagnosing T2DM, Kazerouni et al. (2020) has taken in consideration four different classification models, namely SVM, K-NN, ANN and LR. A comparison is done among these algorithms to measure the diagnostic power of this algorithms. The algorithms are performed on six LncRNA variables and demographic data. K.VijiyaKumar(2019) 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. Nonso Nnamoko 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.

Tejas N. Joshi 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. Deeraj Shetty 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. Muhammad Azeem Sarwar 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. 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 improved. Hence a system is required as Diabetes Prediction is important area in computers, to handle the issues identified based on previous research. Yasodhaet al.[1] uses the classification on diverse types of datasets that can be accomplished to decide if a person is diabetic or not. The diabetic patient's data set is established by gathering data from hospital warehouse which contains two hundred instances with nine attributes. These instances of this dataset are referring to two groups i.e. blood tests and urine tests. In this study the implementation can be done by using WEKA to classify the data and the data is assessed by means of 10-fold cross validation approach, as it performs very well on small datasets, and the outcomes are compared. The naïve Bayes, J48, REP Tree and Random Tree are used. It was concluded that J48 works best showing an accuracy of 60.2% among others.

Aiswaryaet al. [2] aims to discover solutions to detect the diabetes by investigating and examining the patterns originate in the data via classification analysis by using Decision Tree and Naïve Bayes algorithms. The research hopes to propose a faster and more efficient method of identifying the disease that will help in well-timed cure of the patients. Using PIMA dataset and cross validation approach the study concluded that J48 algorithm gives an accuracy rate of 74.8% while the naïve Bayes gives an accuracy of 79.5% by using 70:30 split. Gupta et al. [3] aims to find and calculate the accuracy, sensitivity and specificity percentage of numerous classification methods and also tried to compare and analyse the results of several classification methods in WEKA, the study compares the performance of same classifiers when implemented on some other tools which includes Rapidminer and Matlabusing the same parameters (i.e. accuracy, sensitivity and specificity). They applied JRIP, Jgraft and BayesNet algorithms. The result shows that Jgraft shows highest accuracy i.e 81.3%, sensitivity is 59.7% and specificity is 81.4%. It was also concluded that WEKA works best than Matlab and Rapidminner. Lee et al. [4] focus on applying a decision tree algorithm named as CART on the diabetes dataset. after applying the resample filter over the data. The author emphasis on the class imbalance problem and the need to handle this problem before applying any algorithm to achieve better accuracy rates. The class imbalance is a mostly occur in a dataset having dichotomous values, which means that the class variable have two possible outcomes and can be handled easily if observed earlier in data preprocessing stage and will help in boosting the accuracy of the predictive model. after applying the resample filter over the data. The author emphasis on the class

imbalance problem and the need to handle this problem before applying any algorithm to achieve better accuracy rates. The class imbalance is a mostly occur in

a dataset having dichotomous values, which means that the class variable have two possible outcomes and can be handled easily if observed earlier in data preprocessing stage and will help in boosting the accuracy of the predictive model.

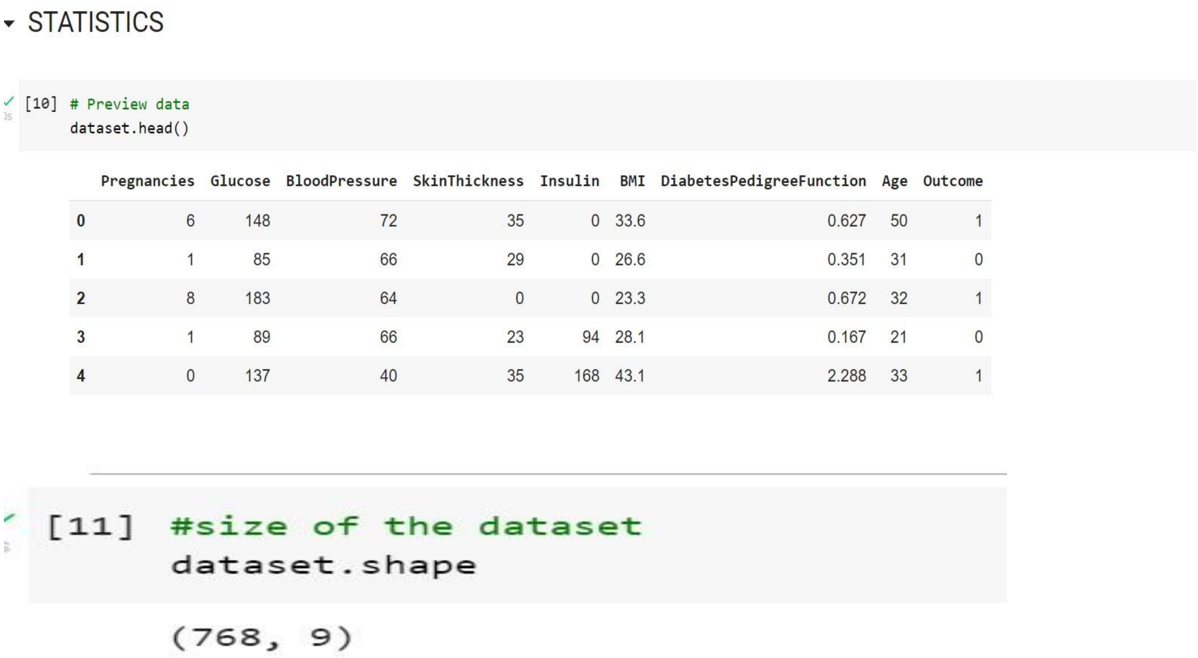
The analysis of related work gives results on various healthcare datasets, where analysis and predictions were carried out using various methods and techniques. Various prediction models have been developed and implemented by various researchers using variants of data mining techniques, machine learning algorithms or also combination of these techniques. Dr Saravana Kumar N M, Eswari, Sampath P and Lavanya S (2015) implemented a system using Hadoop and Map Reduce technique for analysis of Diabetic data. This system predicts type of diabetes and also risks associated with it. The system is Hadoop based and is economical for any healthcare organization. Aiswarya Iyer (2015) used classification technique to study hidden patterns in diabetes dataset. Naïve Bayes and Decision Trees were used in this model. Comparison was made for performance of both algorithms and effectiveness of both algorithms was shown as a result. 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. Humar Kahramanli and Novruz Allahverdi (2008) used Artificial neural network (ANN) in combination with fuzzy logic to predict diabetes. B.M. Patil, R.C. Joshi and Durga Toshniwal (2010) 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. Mani Butwall and Shraddha Kumar (2015) proposed a model using Random Forest Classifier to forecast diabetes behaviour. Nawaz Mohamudally1 and Dost Muhammad (2011) used C4.5 decision tree algorithm, Neural Network, K-means clustering algorithm and Visualization to predict diabetes. It represents taxonomy for Machine Learning Algorithms that can be used for diabetes prediction. The task of choosing a machine learning algorithm includes feature matching of the data to be learned based on existing approaches. Taxonomy of machine learning algorithms is discussed belowMachine learning has numerous algorithms which are classified into three categories: Supervised learning, Unsupervised learning, Semi-supervised learning.

CHAPTER-3

WORKING OF PROJECT

A. **Dataset Description**- the data is gathered from UCI repository which is named as Pima Indian Diabetes Dataset. The dataset have many attributes of 768 patients.



▾ STATISTICS

```
[10] # Preview data
     dataset.head()
```

	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

```
[11] #size of the dataset
     dataset.shape

     (768, 9)
```

Figure 3.1 data processing

A. **Data Visualization-**Data visualization provides an important suite of tools for identifying a qualitative understanding. This can be helpful when we try to explore the dataset and extract some information to know about a dataset and can help with identifying patterns, corrupt data, outliers.

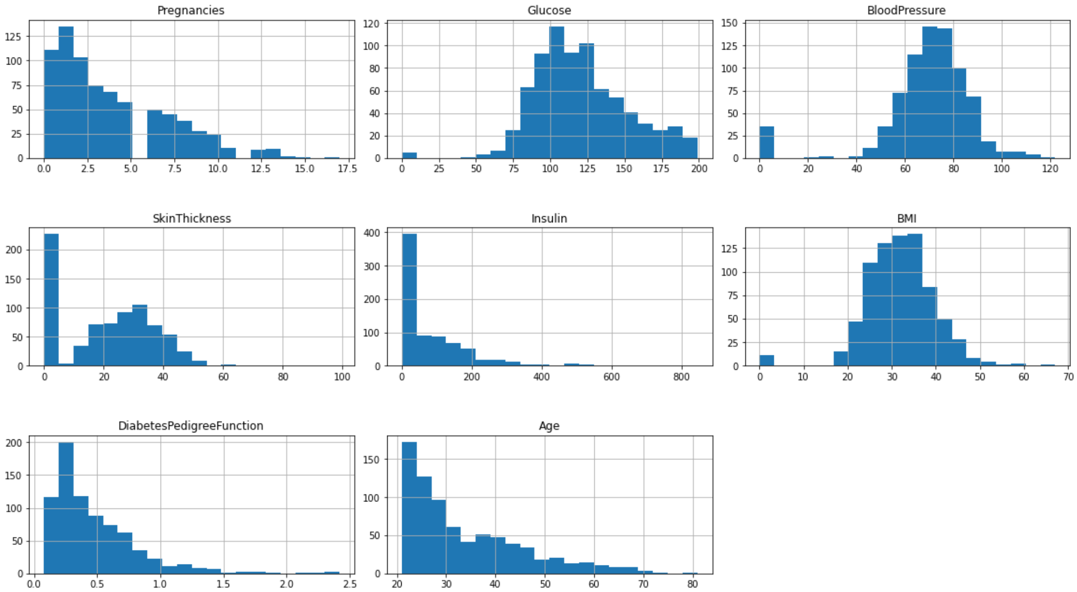
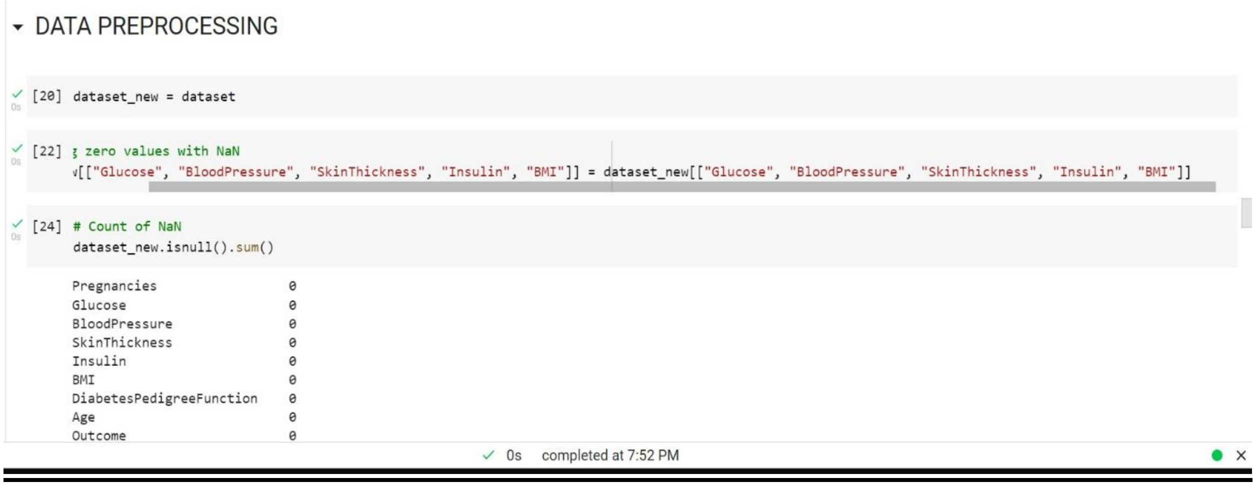


Figure 3.2 Histogram for all the data values in the dataset

B. Data Preprocessing- Data preprocessing is most important process. Mostly healthcare related data contains missing vale and other impurities that can cause effectiveness of data. To improve quality and effectiveness obtained after mining process, Data preprocessing is done. To use Machine Learning Techniques on the dataset effectively this process is essential for accurate result and successful prediction.



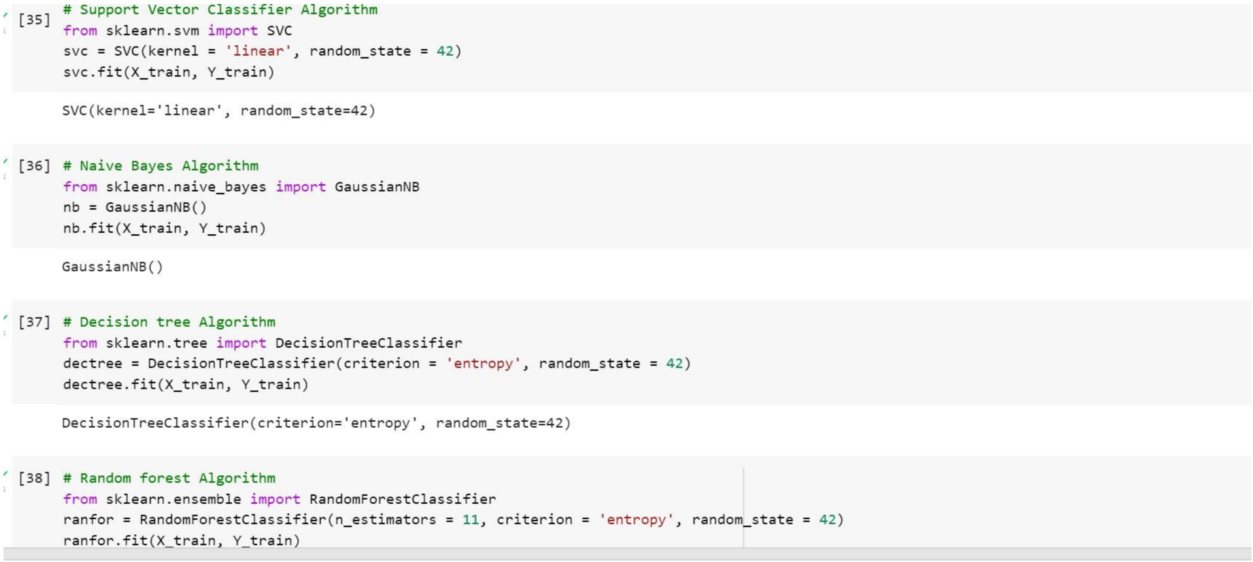
DATA PREPROCESSING

```
[20] dataset_new = dataset

[22] g zero values with NaN
     v[["Glucose", "BloodPressure", "SkinThickness", "Insulin", "BMI"]] = dataset_new[["Glucose", "BloodPressure", "SkinThickness", "Insulin", "BMI"]]

[24] # Count of NaN
     dataset_new.isnull().sum()

     Pregnancies                 0
     Glucose                     0
     BloodPressure               0
     SkinThickness               0
     Insulin                     0
     BMI                         0
     DiabetesPedigreeFunction    0
     Age                         0
     Outcome                     0
```

✓ 0s completed at 7:52 PM

Figure 3.3 data preprocessing techniques

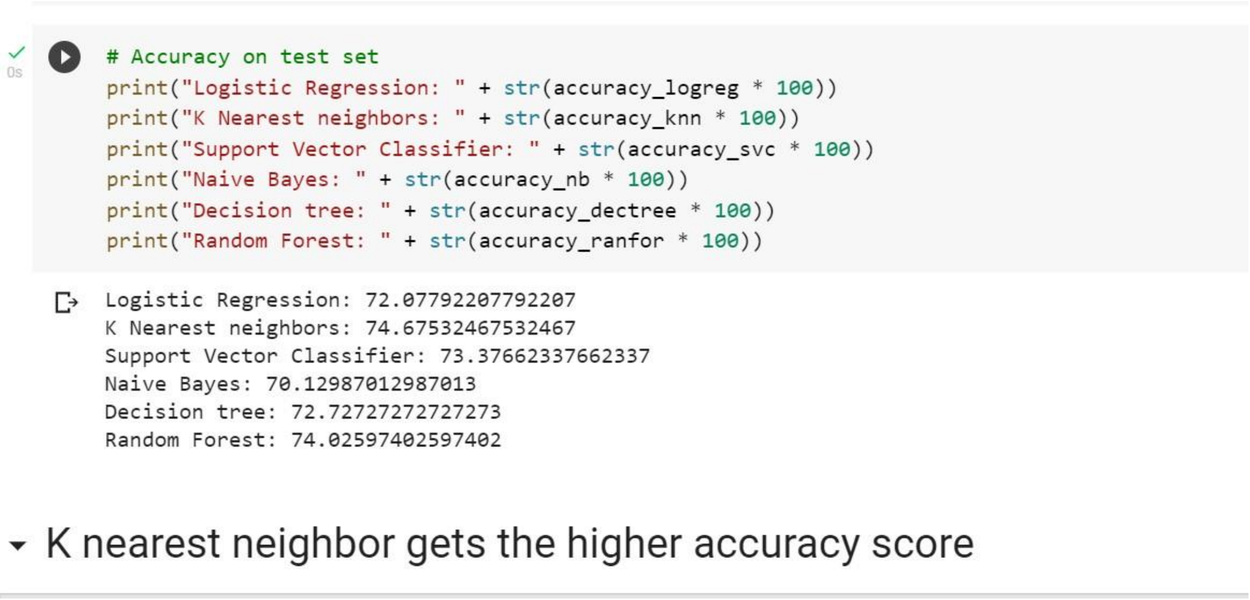
C. Data Modelling- A machine learning model is **a file that has been trained to recognize certain types of patterns**. You train a model over a set of data, providing it an algorithm that it can use to reason over and learn from those data.We used LOGISTIC REGRESSION , K-NEAREST , RANDOM FOREST, SVM, NAÏVE BAYES AND DECISION TREE.



```
[35] # Support Vector Classifier Algorithm
     from sklearn.svm import SVC
     svc = SVC(kernel = 'linear', random_state = 42)
     svc.fit(X_train, Y_train)

     SVC(kernel='linear', random_state=42)

[36] # Naive Bayes Algorithm
     from sklearn.naive_bayes import GaussianNB
     nb = GaussianNB()
     nb.fit(X_train, Y_train)

     GaussianNB()

[37] # Decision tree Algorithm
     from sklearn.tree import DecisionTreeClassifier
     dectree = DecisionTreeClassifier(criterion = 'entropy', random_state = 42)
     dectree.fit(X_train, Y_train)

     DecisionTreeClassifier(criterion='entropy', random_state=42)

[38] # Random forest Algorithm
     from sklearn.ensemble import RandomForestClassifier
     ranfor = RandomForestClassifier(n_estimators = 11, criterion = 'entropy', random_state = 42)
     ranfor.fit(X_train, Y_train)
```

Figure 3.4 Data modelling classification algorithms

D. Evaluation-There are a total of 768 records and 9 features in the dataset. Each feature can be either of integer or float dataype. Some features like Glucose, Blood pressure , Insulin, BMI have zero values which represent missing data. There are zero NaN values in the dataset. In the outcome column, 1 represents diabetes positive and 0 represents diabetes negative



```
# Accuracy on test set
print("Logistic Regression: " + str(accuracy_logreg * 100))
print("K Nearest neighbors: " + str(accuracy_knn * 100))
print("Support Vector Classifier: " + str(accuracy_svc * 100))
print("Naive Bayes: " + str(accuracy_nb * 100))
print("Decision tree: " + str(accuracy_dectree * 100))
print("Random Forest: " + str(accuracy_ranfor * 100))
```

```
Logistic Regression: 72.07792207792207
K Nearest neighbors: 74.67532467532467
Support Vector Classifier: 73.37662337662337
Naive Bayes: 70.12987012987013
Decision tree: 72.72727272727273
Random Forest: 74.02597402597402
```

K nearest neighbor gets the higher accuracy score

Figure 3.5 accuracy testing

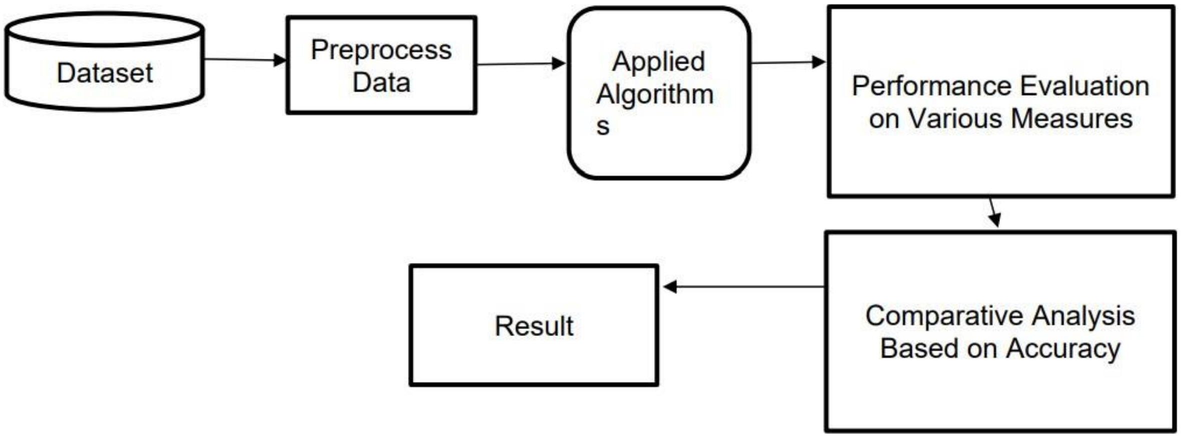

Dataset
Preprocess Data
Applied Algorithm s
Performance Evaluation on Various Measures
Comparative Analysis Based on Accuracy
Result

Fig 3.6 Model Diagram of the Project

CHAPTER -4

RESULT AND DISCUSSION

It is easy to see that there is no single feature that has a very high correlation with our outcome value. Some of the features have a negative correlation with the outcome value and some have positive.

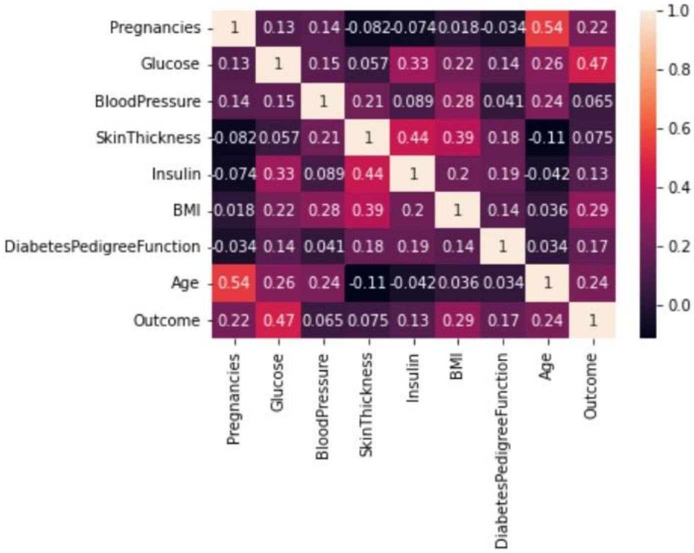


Figure 4.1 Correlation Matrix

It shows how each feature and label is distributed along different ranges, which further confirms the need for scaling. Next, wherever you see discrete bars, it basically means that each of these is actually a categorical variable. We will need to handle these categorical variables before applying Machine Learning. Our outcome labels have two classes, 0 for no disease and 1 for diseases.

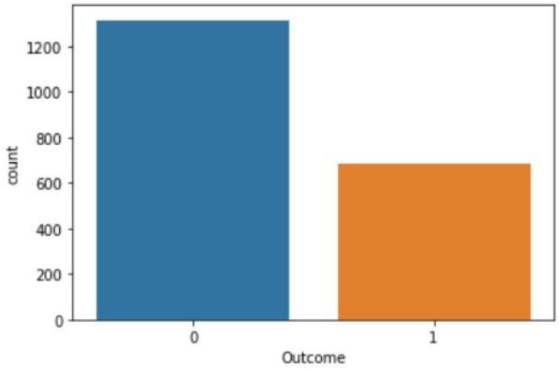


Fig-4.2 Bar Plot For Outcome Class

The above graph shows that the data is biased towards datapoints having outcome value as 0 where it means that diabetes was not present actually. The number of non-diabetics is almost twice the number of diabetic patients.

Procedure of Proposed Methodology

Step1: Import required libraries, Import diabetes dataset.

Step2: Pre-process data to remove missing data.

Step3: Perform percentage split of 80% to divide dataset as Training set and 20% to Test set.

Step4: Select the machine learning algorithm i.e. KNearest Neighbor, Support Vector Machine, Decision Tree, Logistic regression, Random Forest and Gradient boosting algorithm.

Step5: Build the classifier model for the mentioned machine learning algorithm based on training set.

Step6: Test the Classifier model for the mentioned machine learning algorithm based on test set.

Step7: Perform Comparison Evaluation of the experimental performance results obtained for each classifier.

Step8: After analyzing based on various measures conclude the best performing algorithm.

The proposed approach uses different classification and ensemble methods and implemented using python. These methods are standard Machine Learning methods used to obtain the best accuracy from data. In this work we see that random forest classifier achieves better compared to others. Overall we have used best Machine Learning techniques for prediction and to achieve high performance accuracy . . Various Machine Learning Techniques provide efficient result to collect Knowledge by building various classification and ensemble models from collected dataset. Such collecteddata 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. After performing the classification algorithms, we are generating the following results for the different splits of training and testing data: we get results that are close to 74% in the training set and 76-77% in the test results. This depicts that the training set has been trained up to 75% accuracy which means that the data that has been trained has been used to predict the test results which have a 75% average accuracy in the analyzing of the dataset. K-Nearest Neighbor algorithm came out to be the machine learning with highest accuracy.

CHAPTER-5

CONCLUSION AND FUTURE SCOPE

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 affects other organs of human body. Diabetes can be controlled if it is predicted earlier. 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. 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, Knn, Random Forest, Decision Tree, Logistic Regression and Naiye Bayes classifiers are used. And 74% classification accuracy has been achieved. The Experimental results can be asst health care to take early prediction and make early decision to cure diabetes and save humans life. Healthcare professions found it hard to find healthcare data and perform analysis on them due to lack of tools, resources. But using ML, we can overcome this and can perform analysis on real-time data leading to better modeling, predictions. This enhances and improves overall healthcare services. Now, IoTs being integrated with ML in order to make smart healthcare devices that sense if there is any change in the person’s body, health data when he uses the device (Pacemaker, Stethoscope, etc.) and this will notify the person regarding this through an app. This helps in easy monitoring, advanced prediction and analysis thereby reducing errors, saving time and life of people.

Reference

- https://www.ijrte.org/wp-content/uploads/papers/v8i5/E6290018520.pdf
- https://www.analyticsvidhya.com/blog/2021/06/must-known-data-visualization-techniques-for-data-science/
- ijert.org/research/diabetes-prediction-using-machine-learning-techniques-IJERTV9IS090496.pdf
- https://docs.microsoft.com/en-us/windows/ai/windows-ml/what-is-a-machine-learning-model

[1]. Aljumah, A.A., Ahamad, M.G., Siddiqui, M.K., 2013. Application of data mining: Diabetes health care in young and old patients. Journal of King Saud University - Computer and Information Sciences 25, 127–136. doi:10.1016/j.jksuci.2012.10.003.

[2]. Arora, R., Suman, 2012. Comparative Analysis of Classification Algorithms on Different Datasets using WEKA. International Journal of Computer Applications 54, 21–25. doi:10.5120/8626-2492.

[3]. Bamnote, M.P., G.R., 2014. Design of Classifier for Detection of Diabetes Mellitus Using Genetic Programming. Advances in Intelligent Systems and Computing 1, 763–770. doi:10.1007/978-3- 319-11933-5.

[4]. Choubey, D.K., Paul, S., Kumar, S., Kumar, S., 2017. Classification of Pima indian diabetes dataset using naive bayes with genetic algorithm as an attribute selection, in: Communication and Computing Systems: Proceedings of the International Conference on Communication and Computing System (ICCCS 2016), pp. 451– 455.