Final Project Report

Diabetes Prediction using classification method

Project Supervisor
Respected Mam Warda Fiaz

Submitted By

Irfan Ullah Khan BC190400474

Software Projects & Research Section,
Department of Computer Sciences,
Virtual University of Pakistan

CERTIFICATE

This is to certify that Irfan Ullah Khan (BC190400474) have worked on and completed their Software Project at Software & Research Projects Section, Department of Computer Sciences, Virtual University of Pakistan in partial fulfillment of the requirement for the degree of BS in Computer Sciences under my guidance and supervision.

In our opinion, it is satisfactory and up to the mark and therefore fulfills the requirements of BS in Computer Sciences.

Supervisor / Internal Examiner

Warda Fiaz
Supervisor,
Software Projects & Research Section,
Department of Computer Sciences
Virtual University of Pakistan

_____
(Signature)

External Examiner/Subject Specialist

Sheikh Israr Ahmad

_____
(Signature)

Accepted By:

_____
(For office use)

EXORDIUM

In the name of Allah, the Compassionate, the Merciful.

Praise be to Allah, Lord of Creation,
The Compassionate, the Merciful,
King of Judgment-day!

You alone we worship, and to You alone we pray for help,
Guide us to the straight path

The path of those who You have favored,

Not of those who have incurred Your wrath,
Nor of those who have gone astray.

DEDICATION

To

The Last Prophet

Hazrat Muhammad (ﷺ)

All words dedicated with respect to
LOVING PARENTS
Whose love and prays always accompanies like a shining star whenever was in darkness.

.

ACKNOWLEDGEMENT

First of all, I am great full to the ALLAH Almighty the most Merciful and Beneficent who guides us in darkness and helps I difficulties.

All respect for his prophet (P.B.U.U) whose gracious favor and blessings enable me to complete this research successfully.

I regarded it a great honor and privilege to express my deepest sense of gratitude and appreciation to my learned worthy and honorable guide Mam Warda Fiaz for his great guidance encouragement excellent attitude and sincere personal involvement throughout the study.

I am also thankful to all those who helped me in collection of relevant material.

Irfan Ullah Khan

PREFACE

Welcome to the project report on "Diabetes Prediction using Classification Method." This report presents an in-depth exploration of the application of classification methods in the prediction of diabetes, aiming to revolutionize the early detection and management of this prevalent chronic condition.
The implementation of predictive analytics in healthcare, particularly in the domain of diabetes prediction, holds significant promise in identifying individuals at risk and facilitating timely intervention. This report aims to elucidate the fundamental concepts of classification algorithms, data preprocessing, feature selection, model evaluation, and the ethical considerations inherent in leveraging predictive analytics in healthcare.
The system detailed in this report is envisioned to be invaluable for healthcare professionals, data scientists, researchers, and anyone intrigued by the intersection of technology and healthcare. It is designed to equip readers with the knowledge and insights necessary to harness classification methods for diabetes prediction, ultimately contributing to improved healthcare outcomes for individuals at risk of developing diabetes.
I trust that this project report will serve as a comprehensive guide for understanding and implementing predictive analytics in the context of diabetes, fostering enhanced healthcare practices and outcomes.
The duration of this project is 1 year provided by the university and I have successfully completed my project in the given time.

1 Table of Contents

CHAPTER 1

Problem statement and challenges

1.1 Introduction

Diabetes Prediction employing classification technique concentrates on the harnessing of machine learning and data science strategies to come up with a predictive model when it comes to diagnosing diabetes in the patients. Manufactured from a holistic data set; this is aimed at supporting both the patient and healthcare professional in finding and diagnosing diabetes accurately; it is really remarkable. The project aims at the following: bring in and look through the data, conduct Exploratory Data Analysis so as to get useful insights, preprocess the data, train the model through the Neural Network algorithm, test its performance, and then evaluate and compare the various models of SVM, Decision Tree, and Logistic Regression. To predict diabetes status, assess model performance using metrics like - Accuracy, - Precision, - Recall, and - F1 score, as well as to identify the most precise model for the prediction is the main purpose. The project will also place the utmost significance on conserving the highest efficacy of the model for future proceedings, which would mean the improvement of diabetes diagnosis by machine learning.

There are two types of diabetes: Type 1 diabetes and Type 2 diabetes. Type 1 diabetes is a health problem frequently diagnosed in children or young adults when their immune system starts attacking and gradually destroying insulin-developing cells in the pancreas and thus they have to receive regular insulin injections. Type 2 diabetes, on the other hand, is the most common and it is found among people and typically older adults, the body's inability to utilize insulin properly, which is known as insulin resistance being one particular feature. Type 2 diabetes can be mostly supplemented by lifestyle changes, medicine, and sometimes insulin therapy.

1.2 Project problem

The project seeks the proper model of machine learning in determining the symptoms of diabetes with the help of relevant data from the patients. To do that, the model will be advanced by using the Neural Network model, SVM, Decision Tree, and Logistic

Regression. The project is designed in the way a typical machine learning life cycle is followed, data import, exploratory data analysis, data preprocessing, model training, testing, and evaluation. The objective determines which model gives the highest precision in predicting diabetes and deposits the best-performing model for the future.

1.3 Purpose

This project is aiming to create a tool for early detection and prevention of diabetes. Effective prediction of the likelihood of diabetes in people, doctors can act with toys such as changing food, medication, and regular monitoring which are for preventing or delaying the diabetes-related complications.

1.4 Scope

This project encompasses the acquisition and preprocessing of pertinent health data, the selection of suitable classification algorithms, the training and assessment of the predictive model, and the deployment of the model for practical application. An essential aspect involves validating the model's accuracy and dependability using real-world data to ensure its practical efficacy in clinical environments.

1.5 Summary

Using machine learning and data science techniques in this project leads to the development of a predictive model for diagnosing diabetes, thus assisting patients and health providers in the accurate identification of diabetes. Diverse data used in the project includes data importation, exploratory data analysis, pre-processing, and model training with algorithms like Neural Network, SVM, Decision Tree, and Logistic Regression.

Goals comprised in this project are: diabetes status prediction, model performance assessment, and identification of the most precise model. Importance of saving the best-performing model is highlighted, thereby contributing to improvements in diabetes diagnosis

using machine learning techniques. This study concentrates on Type 1 and Type 2 diabetes in order to predict its presence accurately hence providing possibilities for early detection and prevention. At this juncture where one can intervene through appropriate measures if there are any signs recurrently raised concerning an individual's health status in relation to specific diseases like diabetes mellitus that affect everyone regardless of age or sex, specific recommendations should be given.

CHAPTER 2
Proposed methodology

2.1 Introduction

The proposed methodology will use classification techniques to predict diabetes by categorizing data into different groups or classes according to certain characteristics. Classification is a machine learning method that groups data into classes based on some specific criteria. In the case of predicting diabetes, it will apply classification algorithms to some data available to predict how likely someone will be diagnosed with diabetes.

2.2 Purpose

The reason of employing classification techniques for diabetes prediction is developing a reliable and accurate model that can help in timely detection and intervention for people who may develop this disease in the future. Its aim is to enhance the effectiveness of diabetes prediction through use of machine learning algorithms, which consequently allow for proactive healthcare measures as well as personalized interventions aimed at reducing disease burden

2.3 Scope

The scope of this methodology includes:
Data collection: Obtaining data that are connected with diabetes, such as demographic data, lifestyle history and medical records. Selecting Features: Identifying the key factors that predict the risk of contracting diabetes. Model building: Creation and training of classification models like logistic regression, decision tree or support vector machine aimed at studying diabetes.
Model evaluation: How well does a prediction model perform? Some metrics we use in model evaluation include accuracy, sensitivity (recall), specificity or precision as well as AUC (area under the curve).
Interpretation and application: A model's results are interpreted so that there are actionable insights for health care personnel, while at the same time enabling various individuals to make informed decisions on how they could prevent or manage diabetes. Let us develop a strong and credible approach through these key areas of methodology

2.4 Dataset Detail

Detail:
Since there are many open datasets across the web, I am going to use the Pima Indians Dataset from the UCI Machine Learning Repo since it's reliable and contains much more relevant information in regard to the

various attributes used in detection and prediction of this disease such as Glucose level, Insulin, Age, Body Mass Index(BMI), Diabetes Pidegree Function and many more.

You can find this dataset at the link below:

https://www.kaggle.com/datasets/akhilalexander/diabeticprediction

The dataset consist of several medical predictor (Independent) variable and one target variable (outcome).

A predictor variable includes: Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, age, and Outcome.

All about Dataset:

Name of the columns with their meanings

1) **Pregnancies:** Number of times pregnant

2) **Glucose:** Plasma glucose concentration a 2 hours in an oral glucose tolerance test

3) **Blood Pressure:** Diastolic blood pressure (mm Hg)

4) **Skin Thickness:** Triceps skin fold thickness (mm)

5) **Insulin:** 2-Hour serum insulin (mu U/ml)

6) **BMI:** Body mass index (weight in kg/(height in m)^2)

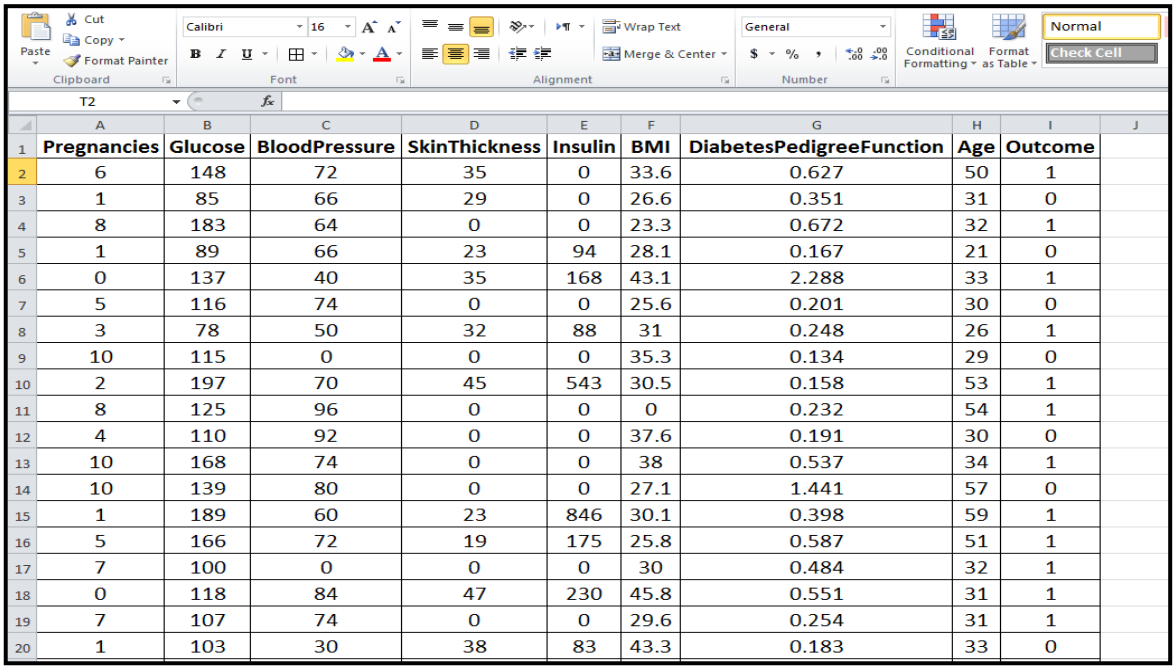
7) **Diabetes Pedigree Function:** Diabetes pedigree function

8) **Age:** Age (years)

9) **Outcome:** Class variable (0 or 1)

268 of 768 are 1 …. (Diabetic)

500 of 768 are 0 …. (Non-Diabetic)



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

Dataset of Diabetes Prediction

2.5 Proposed Methodology

Every software development methodology is a platform for carrying out specific steps that are meant to help in the growth and sustenance of software. Many software development tactics have been in play from the genesis of information technology. Some of the common existing methodologies in the IT industry include.

- **Problem Definition:**

  Define diabetes prediction problem and its importance for medicine. It expresses the importance of developing an accurate and dependable model for early detection.

- **Literature Review:**

  Conduct in-depth research on diabetes forecasting with respect to current literature using machine learning. While tackling diabetes diagnosis Neural Networks, SVM, Decision Trees, and Logistic Regression patterns have been explored by many scholars. Review methodologies, difficulties, consequences from similar studies."

- **Objective Refinement:**:

  Using knowledge from reading, refine the deliverables of this study. These should be in line with the current trends in the field and should cover all the areas that were left unaddressed.

- **Data Collection**:

  Get the diabetic dataset from Kaggle using the link that will be provided to you. It is important that this dataset is able to represent a broad range of situations and can be used for training different types of machine learning algorithms as well.

- **Exploratory Data Analysis (EDA):**

  Try using statistical methods and data visualization tools to familiarize yourself with the dataset. Find out how the various features are distributed, try spotting patterns and most importantly figure out the possible indicators for diabetes.

- **Data Pre-processing:**

  Divide data set into 70% for training and 30% for testing purposes. This involves different preparation steps such as normalizing variables or imputing missing values that may be present in your dataset."

- **Model Training:**

  A Neural Network is trained to predict diabetes using the training dataset. Optimize hyper parameters and monitor training progress.

- **Model Application:**

Use SVM, Decision Tree, and Logistic Regression models in the training data. The results are then compared with those obtained from the Neural Network model.

- **Evaluation Metrics:**

Evaluate how well the model performs by confusion matrix metrics such as Accuracy, Precision, Recall, F1 score and also more in-depth analysis like assessing on the common falsehood. Find out which model has the highest Accuracy and decide whether it is the best one.

- **Results Analysis:**

Analyze and interpret the results, making comparisons between different models. Identify strengths and limitations for each model in predicting diabetes.

- **Model Selection:**

Choose the best model based on its evaluation metrics for future usage.

- **Save Model:**

Implementing a mechanism to save selected models for future predictions ensures efficiency and quick access to forecast capability."

- **Ethical Considerations:**

Discuss the ethical considerations around data privacy, consent and predictive models being responsibly used; make sure you follow all ethical standards while using the data of patients.

- **Documentation and Reporting:**

In the documentation process, make sure you have taken note of every research step which includes methodologies used, decisions made during the process as well as their outcomes; give a detailed report showing research findings, model performance as well as possible future work recommendations.

Algorithms used in this Project:

a) Logistic Regression

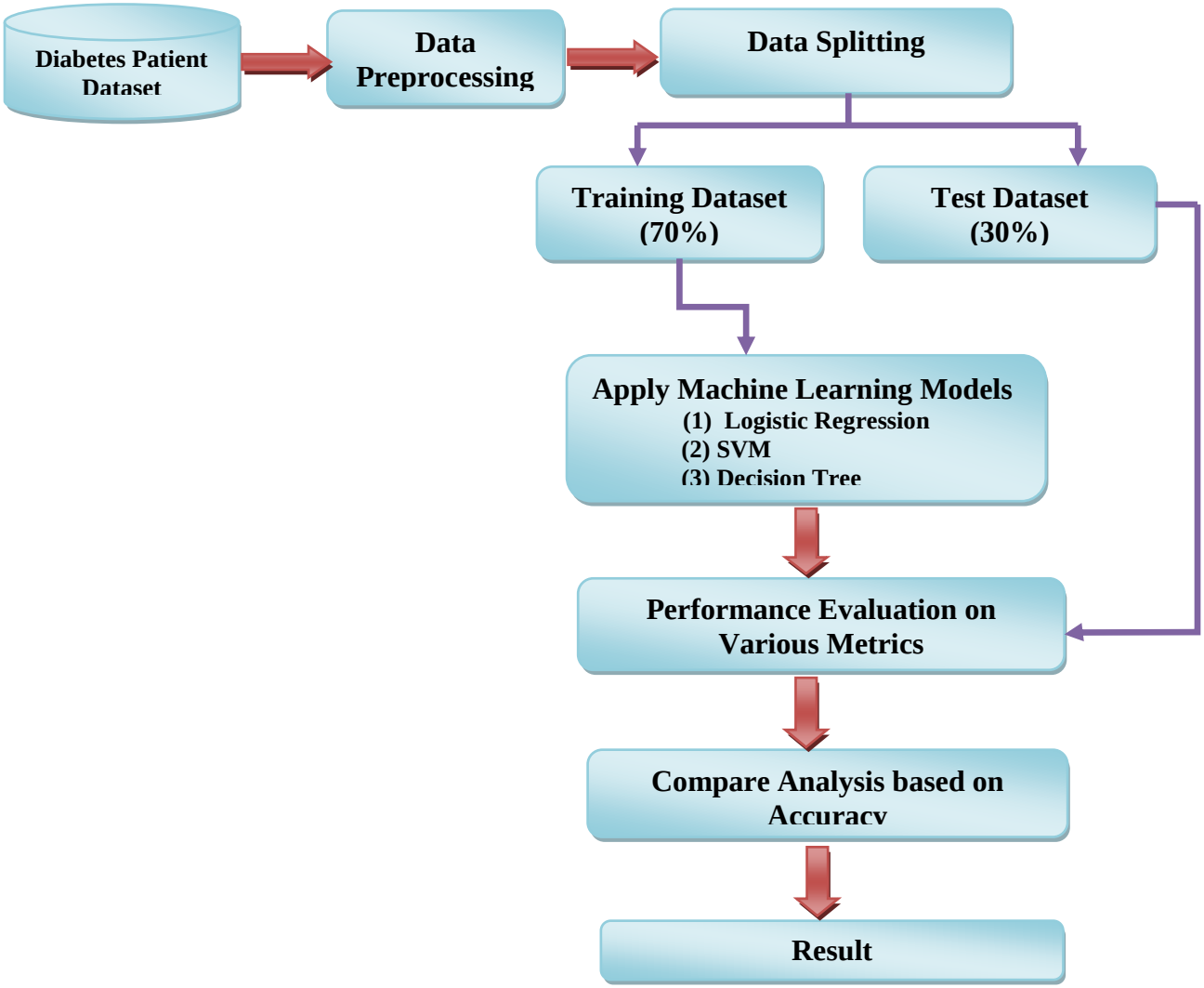
This is when the machine learning algorithm or AI is at work to sort an outcome between "zero" and "one".

b) Support Vector Machine (SVM):

This is machine learning technique that attempts to logically divide samples into categories according to some criterion.

c) Decision Tree:

At each Node of tree, the optimal condition splits our data into two based on some feature



2.6 Data Pre-processing

The data pre-processing steps for diabetes prediction using classification methods entail essential procedures that are meant to increase the accuracy and efficiency of predictive models. In diabetes prediction context, data pre-processing helps in classifying algorithms dataset preparation where classified. Below is a brief description of how these activities are done:

- ✓ **Data Cleaning: Data Integration**
- ✓ **Data Transformation**
- ✓ **Data Reduction**

- ✓ **Data Discretization**
- ✓ **Data Normalization**

The data mining and machine learning procedure will be more effective and accurate by carrying out such data preprocessing tasks. It is necessary to perform this step on data to obtain information that may be considered accurate as well as meaningful by the consuming public.

The predictive models for diabetes classification using machine learning methods can have their accuracy, efficiency and robustness during predicting diabetes outcomes improved if the steps of careful data pre-processing listed below are followed correctly.

CHAPTER 3

Implementation

3.1 Implementation

Here's a detailed breakdown of the implementation process that is involved in implementing this project which includes various main steps for building and testing the diabetes prediction model.

3.1.1 Data Acquisition

Get the diabetes dataset at this location (https://www.kaggle.com/datasets/akhilalexander/diabeticprediction). Get it into the system that it can be processed further.

3.1.2 Exploratory Data Analysis (EDA)

Perform Exploratory Data Analysis (EDA) in order to understand more about the dataset. Figure out some of the facts, ongoing processes, sequences and associations present within this dataset by studying summary statistics. Based on some of the above steps, draw diagrams for better comprehension about various properties in this data."

3.1.3 Data Preprocessing

To deal with any missing values, use proper techniques such as imputation. Transform categorical data into formats that can work with machine learning algorithms. Ensure numerical features are on similar scales. Divide the dataset into training (70%) and testing (30%) subsections. Train the Neural Network model with the training data. Optimize hyper parameters for improving performance in Neural Networks.

3.1.4 Model Evaluation

Application of the trained Neural Network model on test data Performance evaluation of the model using metrics such as Precision, Recall, Accuracy and F1 score Construction of a confusion matrix to evaluate how well the model performs in terms of prediction ability 6. Comparative Analysis Train other models with the algorithms SVM, Decision Tree and Logistic Regression. Using the same set of metrics as employed in step 5; assess how well each model performs.
Identify the best model out of all models for diabetes prediction so that its productivity is maximized 7. Model Selection and Saving Choose the most accurate model specifically in predicting whether a person has diabetes or not save the chosen mode for future utilization and deployment the primary objective of this project is to enable patients and healthcare providers detect and prevent diabetes as early as possible by creating the best possible diabetes prediction tool By following this implementation process, the project aims to develop an accurate and reliable diabetes prediction model that can assist patients and healthcare professionals in early detection and prevention of diabetes.

3.2 Tools and Language

To implement the diabetes prediction project a combination of tools and programming languages that meet machine learning and data science requirements will be used. Here is a list of essential tools and languages that will be required in the study:

3.2.1 Tools:

1. **Windows 7 or higher**
2. **Google Chrome Browser**
3. **Anaconda**
4. **Python**: Python will serve as the primary programming language for developing the machine learning models due to its extensive libraries for data manipulation, analysis, and modeling (e.g., Numpy, Pandas, and Scikit-learn).
5. **Excel:** We used Excel for dataset.
6. **Jupyter Notebook**: Jupyter Notebook provides an interactive environment for code development, data visualization, and documentation, making it ideal for exploratory data analysis and model prototyping.

3.2.2 Language:

Python: Python is a versatile and powerful programming language that has a wide range of applications and a promising future scope. Here's why Python is considered one of the best programming languages:

- Diverse Applications: Data science, web development, AI, automation, IoT, scientific computing
- Easy to Learn and Use: Simple, intuitive syntax, emphasis on readability
- Extensive Libraries and Ecosystem: Vast community-developed tools and frameworks
- Cross-Platform Compatibility: Runs on Windows, macOS, Linux
- Growing Demand and Job Opportunities: High and increasing adoption across industries

Python's wide-ranging capabilities, ease of use, and thriving ecosystem make it one of the best programming languages, with a promising future outlook across various domains.

- **Pandas**:
- **Scikit-learn**: Scikit-learn will be utilized for implementing machine learning algorithms like SVM, Decision Tree, and Logistic Regression, enabling model training and evaluation.
- **Matplotlib and Seaborn**: These visualization libraries will be employed for creating informative plots and graphs to visualize data patterns and model performance.
- **NumPy** : NumPy is a powerful Python library for scientific computing.

Python is broad in scope and one of the most versatile, easy to use, and widely applied programming languages because of its extensive ecosystem. More importantly, as Python's popularity increases, it appears likely that there will be brighter days ahead, particularly considering the field of data science, machine learning and web development.

3.3 DATASET DESCRIPTION

The objective of this project is to predict whether patient has diabetes or not.

The dataset consist of several medical predictor (Independent) variable and one target variable (outcome).

A predictor variable includes: Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, age, and Outcome.

All about Dataset:

Name of the columns with their meanings

1) **Pregnancies:** Number of times pregnant

2) **Glucose:** Plasma glucose concentration a 2 hours in an oral glucose tolerance test

3) **Blood Pressure:** Diastolic blood pressure (mm Hg)

4) **Skin Thickness:** Triceps skin fold thickness (mm)

5) **Insulin:** 2-Hour serum insulin (mu U/ml)

6) **BMI:** Body mass index (weight in kg/(height in m)^2)

7) **Diabetes Pedigree Function:** Diabetes pedigree function

8) **Age:** Age (years)

9) **Outcome:** Class variable (0 or 1)

There are 9 Attributes and 768 Instances.

Shape of Data

```
In [86]: data.shape
Out[86]: (768, 9)
```

268 of 768 are 1 … (1 indicates Diabetic)

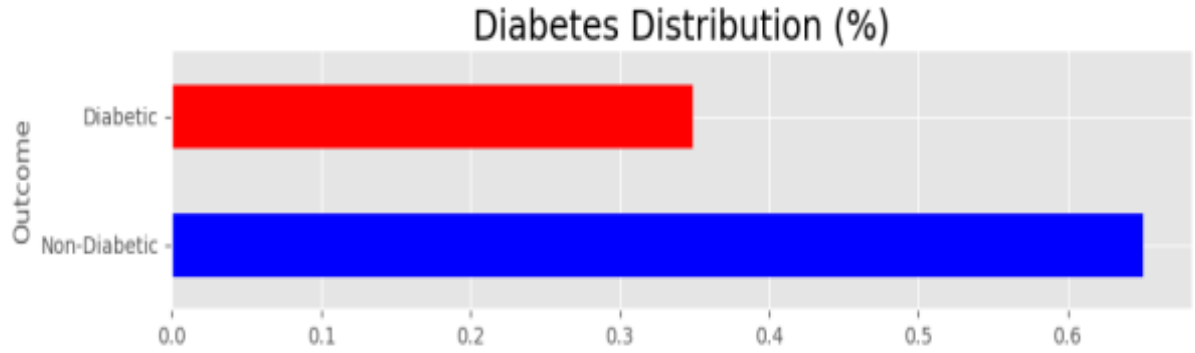
500 of 768 are 0… (0 indicates Non-Diabetic)

Diabetes Distribution

```
#Finding Class Distribution Percentage
print(data['Outcome'].value_counts(ascending=True))
print(data['Outcome'].value_counts(1,ascending=True).apply(lambda x: format(x,
print()
# Plot the bar chart
data['Outcome'].value_counts(normalize=True).plot(kind='barh',figsize=(10, 2),
plt.title('Diabetes Distribution (%)', fontsize=18)
plt.yticks(ticks=[0,1], labels=['Non-Diabetic', 'Diabetic'])
plt.show()
```

```
Outcome
1    268
0    500
Name: count, dtype: int64
Outcome
1    34.895833%
0    65.104167%
Name: proportion, dtype: object
```

thod



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

Dataset of Diabetes Prediction

3.4 Data Pre-processing

The data pre-processing steps for diabetes prediction using classification methods entail essential procedures that are meant to increase the accuracy and efficiency of predictive models. In diabetes prediction context, data pre-processing helps in classifying algorithms dataset preparation where classified. Below is a brief description of how these activities are done:

Data Cleaning: It deals with spotting and rectifying inaccuracies or disparities in records such as absence of figures, unusual things, and copies. Ways to clean data include filling in missing data, deleting data and data conversion.

Data Integration: It involves coming up with a single set of data by bringing together different data from various sources. It becomes a bit of a challenge to merge data because they come in different formats, structures, and meanings. When it comes to data integration, record linkage and data fusion should also be considered as some of the techniques that can be used.

Data Transformation: This entails preparing the data in a way that it can be analyzed, and some of the common methods are normalization, standardization, and discretization. Normalization scales the data i.e., brings the range of the variables to a common standpoint whereas standardization transforms the data such that its mean is zero while the standard deviation is one. Discretization on the other hand is a process through which continuous information is divided into several categories and further grouped.

Data Reduction: Such data reduction methods include extracting features that are used to represent high-level data as in methods like principal component analysis (PCA) and linear discriminant analysis (LDA). Extracting such features usually entails estimation of eigenvectors or singular vectors corresponding to the largest eigenvalues or singular values of the matrix that results from applying one of the multiple aforementioned algorithms.

Data Discretization: This is where data is partitioned into parts which are separate and distinct instead of being continuous. Basically, discretization is always used for algorithms where categorical input is required like in machine learning and data mining. Ways of making data discrete include but aren't limited to equal width binning technique, clustering or even equal frequency binning technique.

Data Normalization:

Normalization is only applicable to numerical columns. There are five common normalization methods:

1. Single feature scaling
2. Min-max scaling
3. Z-score normalization
4. Log scaling
5. Clipping

This consist of shuffling the data into a uniform set of data such as 0-1 or -1 , Normalizing is often used for different unit data and large scales Normalization techniques include min – max , z-score and decimal scaling.

The data mining and machine learning procedure will be more effective and accurate by carrying out such data preprocessing tasks. It is necessary to perform this step on data to obtain information that may be considered accurate as well as meaningful by the consuming public.

Normalization is a technique often applied as part of data preparation for machine learning. The goal of normalization is to change the values of numeric columns in the dataset to a common scale, without distorting differences in the ranges of values. For machine learning, every dataset does not require normalization.

Removing Null Values:

There are a few ways to remove null values in Python:

- ✓ **Using dropna() in Pandas**: The dropna() method allows you to drop rows or columns with null values.
- ✓ **Using filter() function**: The built-in filter() function can be used to create a new list without null values.

- ✓ **Using list comprehension**: You can use a list comprehension to create a new list without null values.

By removing null values; you can ensure that your data is clean and ready for further analysis or model training. This can lead to more accurate and reliable results, as well as more efficient computations.

In summary, removing null values is a crucial data cleaning step in Python, as it helps to ensure the quality and reliability of your data and analysis.

Out Put of Null value:

Checking Null Values

```
In [93]: data.isnull().sum()

Out[93]: Pregnancies                 0
         Glucose                     0
         BloodPressure               0
         SkinThickness               0
         Insulin                     0
         BMI                         0
         DiabetesPedigreeFunction    0
         Age                         0
         Outcome                     0
         dtype: int64
```

So there is no null value in our project.

CHAPTER 4

Result and discussion

4.1 RESULTS of Algorithms

In this project we have use 3 algorithms which are given below:

1) **Logistic Regression:**
2) **Support Vector Machine (SVM):**
3) **Decision Tree**

Logistic Regression

A logistic regression model is a supervised learning algorithm that is used to classify binary data. It will provide the probability that a data point is binary as a function of one or more predictor variables. For example logistic regression is commonly used in predicting customer churn rates, credit card fraud detection systems and medical diagnosis

The result of this algorithm is :

Algorithm	Precision	Recall	F-1	Accuracy %
Logistic Regression	0.74	0.74	0. 0.74	**74.025974**

Support Vector Machine (SVM)

Support Vector Machine (SVM) is an algorithm for supervised learning which can serve for both regression and classification jobs. In finding the feasible hyper plane which partitions different group maximally, SVMs are employed. SVMs are acknowledged for their suitability towards disparate data while being especially powerful in high-dimensional scenarios. Image recognition, text categorization, and bioinformatics are the typical incidences where they are applied.

The result of this algorithm is:

Algorithm	Precision	Recall	F-1	Accuracy %
SVM	0.73	0.74	0.72	**73.593074**

Decision Tree

A decision tree is an algorithm that tells you if an observation is good or bad. It compares the features of the current record with the branch of the tree and follows it to the bottom. There are many nodes at each stage and various decisions occur at each node. All these decisions lead to a single conclusion about an observation as you want either positive judgment or a negative. However, it recursively partitions the data based on the feature that provides the most information gain leading to a hierarchy of_ decisions sill this decision is reached; hence it is known as recursive partitioning. Decision Trees are intuitive, easy to interpret, and can handle both numerical and categorical data. They are used in various fields such as medical diagnosis, quality control, and decision support_ systems for_ industrial processes.

The result of this algorithm is:

Algorithm	Precision	Recall	F-1	Accuracy %
Decision Tree	0.70	0.68	0.69	**68.396268**

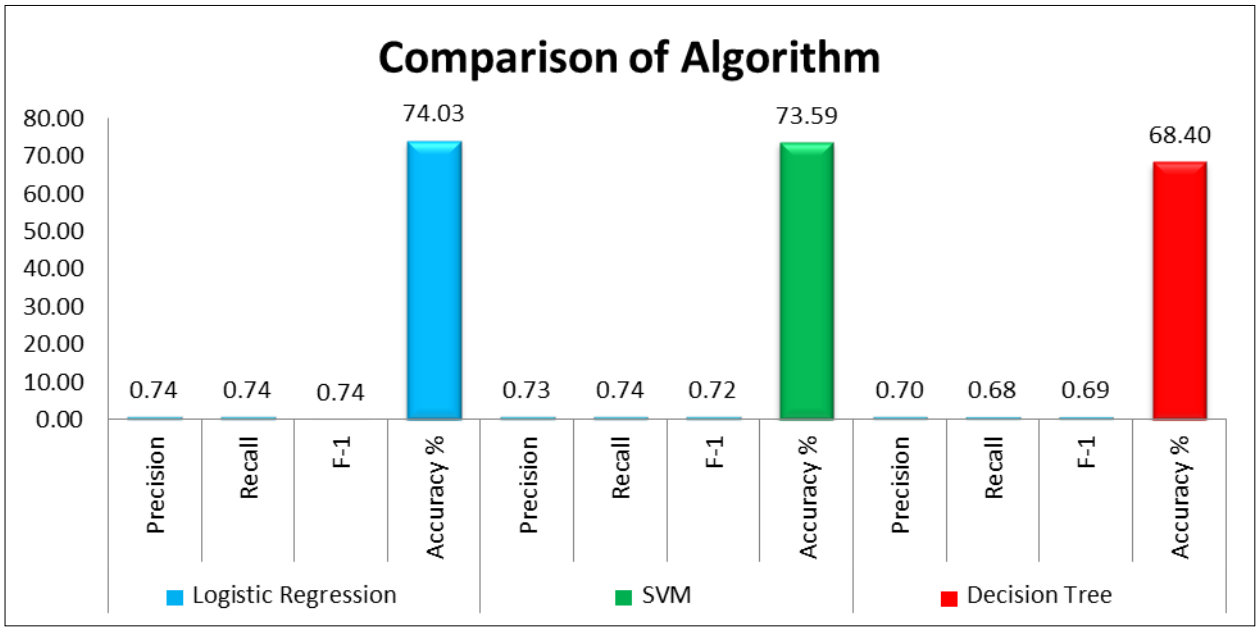
4.2 DISCUSSION FOR BEST

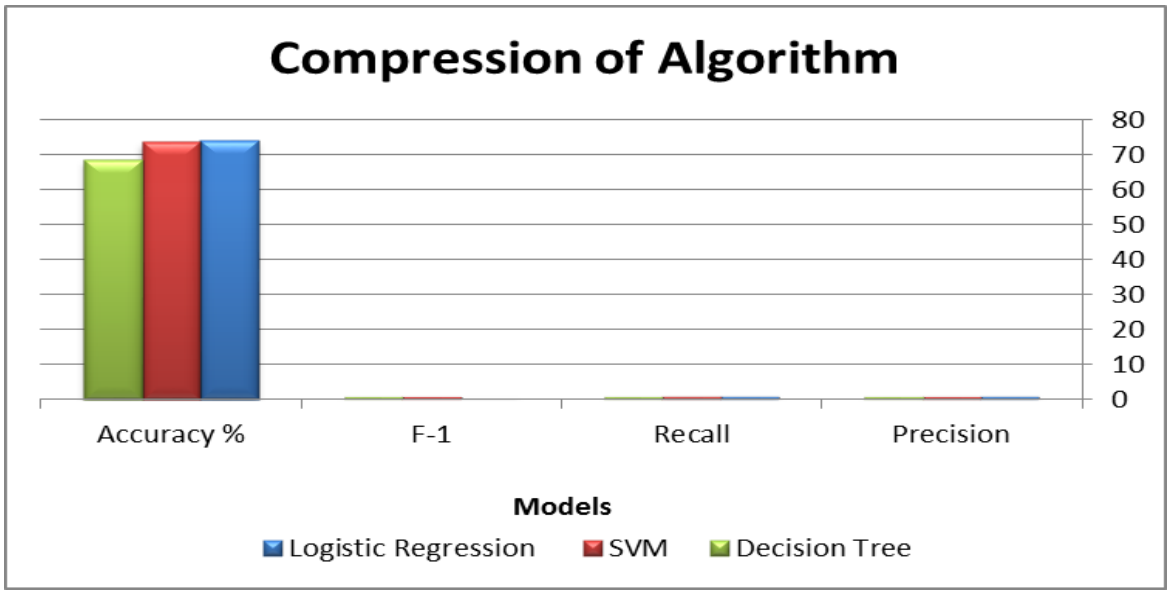
After successfully run of a 3 algorithm we get this result:

Algorithm	Precision	Recall	F-1	Accuracy %
Logistic Regression	0.74	0.74	0. 0.74	**74.025974**
SVM	0.73	0.74	0.72	**73.593074**
Decision Tree	0.70	0.68	0.69	**68.396268**

Logistic Regression provides us with higher accuracy in comparison with other methods. Its accuracy is **74.025974%.** It is a binary classification learning algorithm that is supervised and models the probability of a certain binary output such as yes/no or 1/0 based on a corresponding variable. And it is important to know that the outcome variable being analyzed is binary such as 1 or 0. Therefore, the solution is the most appropriate.

4.3 COMPARISON OF ALGORITHMS (DIAGRAM)





5 REFERENCES

1. www.academia.edu
2. www.slideshare.net
3. www.uml.org
4. *https://www.javatpoint.com/machine-learning-life-cycle*
5. *https://www.kaggle.com/datasets/akhilalexander/diabeticprediction*
6. *https://ocw.vu.edu.pk/Videos.aspx?cat=Computer+Science%2fInformation+Technology+&course=CS607*
7. *https://www.youtube.com/watch?v=_u-PaJCpwiU&list=PLu0W_9lII9ai6fAMHp-acBmJONT7Y4BSG&index=1*
8. *https://vulms.vu.edu.pk/Courses/CS607/Downloads/AI_Complete_handouts_for_Printing.df*
9. *https://www.programiz.com/python-programming*
10. https://www.tutorialspoint.com/python/index.htm
11. *https://www.tutorialspoint.com/python/index.htm*
12. *https://www.w3schools.com/python/*
13. https://machinelearningmastery.com/machine-learning-in-python-step-by-step/
14. https://www.geeksforgeeks.org/machine-learning-with-python/
15. https://www.youtube.com/watch?v=ZftI2fEz0Fw
16. *https://www.simplilearn.com/10-algorithms-machine-learning-engineers-need-to-know-article*
17. *https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/*
18. *https://www.anaconda.com/*
19. *https://www.python.org/*
20. *https://jupyter.org/*
21. *https://www.academia.edu/42929334/Diabetes_prediction_using_classification_algorithms_for_Weka*
22. **CS504 Software Engineering – I**
23. **CS605 Software Engineering – II**
24. **CS403 Database Management System**
25. **CS614 SOFTWARE PROJECT MANAGEMENT**

6 APPENDIX

Chapter 1	PROBLEM STATEMENT AND CHALLENGES
Chapter 2	PROPOSED METHODOLOGY
Chapter 3	IMPLEMENTATION
Chapter 4	RESULT AND DISCUSSION

Author:

Irfan Ullah Khan
bc19040474@vu.edu.pk
programmarself@gmail.com