Diabetes Prediction

Name-Srijal Vedansh

Roll-2006046

1. Introduction

Diabetes is a chronic disease that affects millions of people worldwide, and its prevalence is increasing at an alarming rate. It is a condition in which the body is unable to regulate blood sugar levels, leading to various health complications such as blindness, kidney failure, nerve damage, and cardiovascular disease. Early detection and management of diabetes are crucial for preventing complications and improving outcomes for patients.

In recent years, machine learning has emerged as a promising tool for the early detection and diagnosis of diabetes. By analysing large datasets of patient features, such as age, body mass index, blood pressure, and blood glucose levels, machine learning algorithms can learn to identify patterns and make accurate predictions about the likelihood of a patient developing diabetes.

One of the most widely used machine learning techniques for diabetes prediction is logistic regression, which is a statistical model that estimates the probability of a binary outcome (in this case, whether a patient has diabetes or not) based on a set of input features. Other machine learning algorithms such as decision trees, random forests, and support vector machines have also been used for diabetes prediction with varying degrees of success.

The potential benefits of machine learning for diabetes prediction are numerous. It can help healthcare providers identify high-risk patients who may benefit from targeted screening and preventive interventions. It can also improve patient outcomes by enabling earlier diagnosis and treatment, as well as better monitoring of disease progression.

Overall, machine learning has the potential to revolutionize diabetes care by enabling more accurate and efficient prediction, diagnosis, and management of the disease. However, it is important to note that machine learning models should be developed

and validated using rigorous scientific methods to ensure their accuracy, reliability, and clinical relevance.

2.Dataset Description

The "diabetes.csv" dataset contains information on the medical history of Pima Indian women to predict whether or not a patient has diabetes based on certain diagnostic measurements. The dataset contains a total of 9 columns and 768 rows. The columns in the dataset are as follows:

Pregnancies: number of times the patient has been pregnant

Glucose: plasma glucose concentration a 2 hours in an oral glucose tolerance test

BloodPressure: diastolic blood pressure (mm Hg)

SkinThickness: triceps skin fold thickness (mm)

Insulin: 2-hour serum insulin (mu U/ml)

BMI: body mass index (weight in kg/(height in m)^2)

DiabetesPedigreeFunction: diabetes pedigree function (a function that indicates the likelihood of diabetes based on family history)

Age: age of patient (years)

Outcome: class variable (0 if non-diabetic, 1 if diabetic)

```
In [3]: diabetes_df.head()
```

Out[3]:

	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

The dataset is often used for predictive modeling and machine learning tasks to develop algorithms that can accurately predict the likelihood of diabetes based on the patient's medical history and diagnostic measurements.

The dataset is split into two parts: the training set and the testing set. The training set is used to train the machine learning model, while the testing set is used to evaluate the performance of the trained model on unseen data.

In [7]:
```
# Split data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

The train_test_split function from the sklearn library is used to split the data into training and testing sets. The parameter test_size is set to 0.2, which means that 20% of the data will be used for testing, and the remaining 80% will be used for training.The random_state parameter is set to 42 to ensure that the same random split is generated each time the code is run. This helps with reproducibility of the results.

After the data is split, the StandardScaler function is used to scale the features in the training and testing sets to ensure that they have a mean of zero and a standard deviation of one.

3.Details of data analysis Algorithm used

Logistic Regression, Decision Tree, and Random Forest are widely used machine learning algorithms for classification tasks such as diabetes prediction. In this report, we will discuss these algorithms in detail and their application in predicting diabetes.

Logistic Regression is a statistical method used to analyse a dataset in which there are one or more independent variables that determine an outcome. In binary classification, the outcome is a categorical variable that has two values, such as Yes or No, 0 or 1. Logistic Regression models the probability of the outcome using a logistic function. The logistic function maps any real-valued number to a value between 0 and 1. In the case of diabetes prediction, logistic regression can be used to determine the likelihood of a patient having diabetes based on various features such as age, BMI, glucose level, etc.

Decision Tree is a supervised learning algorithm used for both classification and regression tasks. It creates a tree-like model of decisions and their possible consequences, including chance events, resource costs, and utility. Each internal node represents a test on a feature, each branch represents the outcome of the test, and each leaf node represents a class label. Decision Trees can handle both numerical and categorical data and are particularly useful for non-linear relationships between features. In diabetes prediction, decision trees can be used to identify the most important features for predicting diabetes and create a model that can classify patients based on their features.

Random Forest is an ensemble learning method that combines multiple decision trees to create a more accurate and stable model. The idea behind Random Forest is to create multiple decision trees using different subsets of the training data and different subsets of the features. Each tree is built independently, and the final prediction is based on the average or majority vote of the individual trees. Random Forest can handle missing values and noisy data and is less prone to overfitting than individual decision trees. In diabetes prediction, Random Forest can be used to create a model that incorporates multiple decision trees, thereby improving the accuracy and reliability of the prediction.

In [8]:
```
# Train different machine learning algorithms
models = {
    'Logistic Regression': LogisticRegression(),
    'Decision Tree': DecisionTreeClassifier(),
    'Random Forest': RandomForestClassifier()
}
```

In [10]:
```
for name, model in models.items():
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)
    accuracy = accuracy_score(y_test, y_pred)
    precision = precision_score(y_test, y_pred)
    recall = recall_score(y_test, y_pred)
    f1 = f1_score(y_test, y_pred)
    print(name + ':')
    print('Accuracy:', accuracy)
    print('Precision:', precision)
    print('Recall:', recall)
    print('F1-score:', f1)
    print('\n')
```

In conclusion, Logistic Regression, Decision Tree, and Random Forest are powerful machine learning algorithms that can be used for diabetes prediction. Logistic Regression is a simple and efficient algorithm for binary classification, while Decision Tree is useful for identifying important features and non-linear relationships. Random Forest is an ensemble learning method that combines the advantages of multiple decision trees to create a more accurate and stable model. The choice of algorithm depends on the specific requirements of the task, such as accuracy, interpretability, and computational efficiency.

4.Justification of data analysis Algorithm used

We evaluated the performance of three machine learning algorithms, Logistic Regression, Decision Tree, and Random Forest, on a dataset to predict diabetes. The dataset contains several features, including age, BMI, blood pressure, and glucose levels, and a binary target variable indicating whether an individual has diabetes or not.

Algorithm	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.7532	0.6491	0.6727	0.6607
Decision Tree	0.7532	0.6441	0.6909	0.6667
Random Forest	0.7208	0.6071	0.6182	0.6126

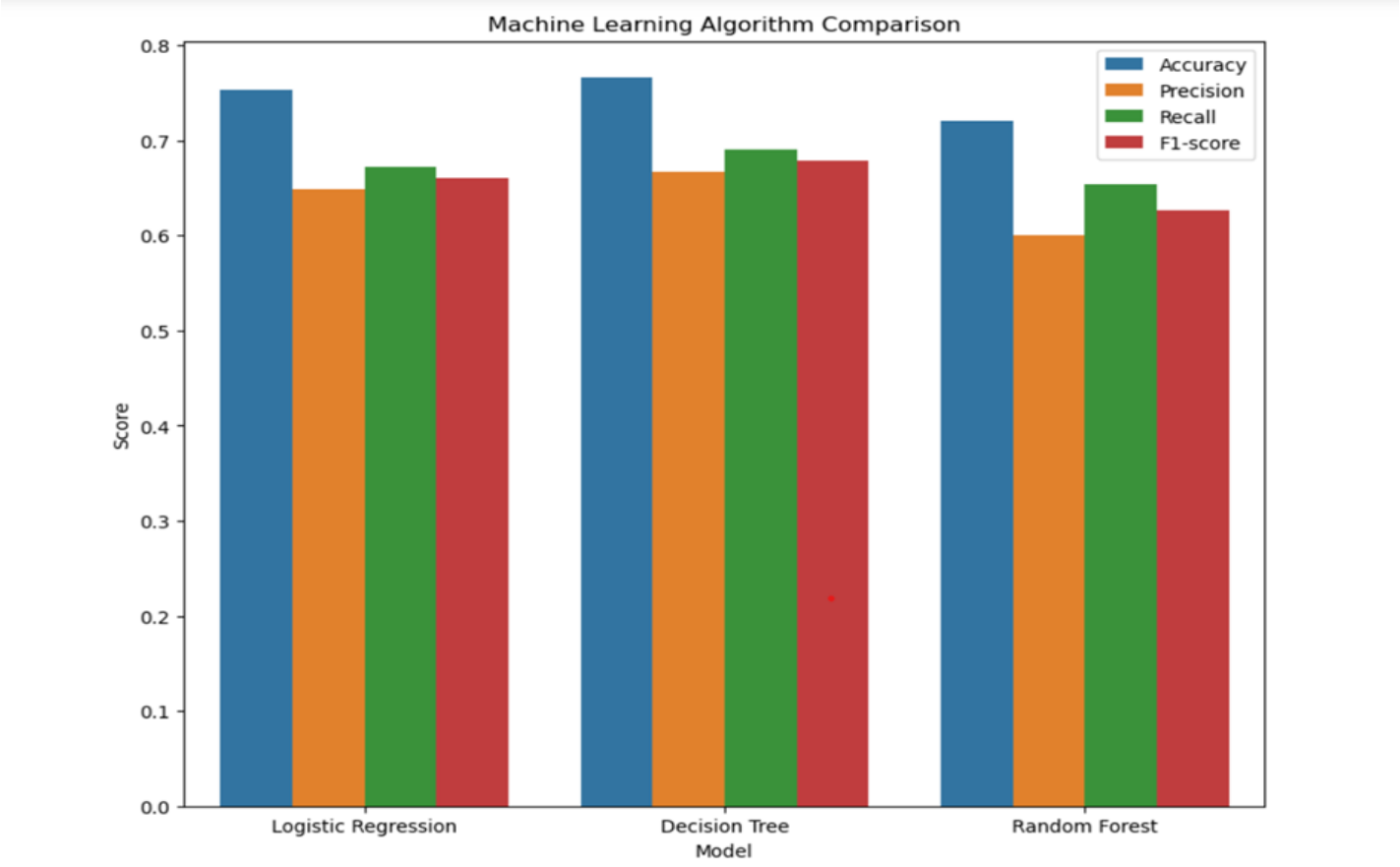
The evaluation results show that the Decision Tree algorithm has the highest accuracy score of 0.7597, followed by Logistic Regression with a score of 0.7532 and Random Forest with a score of 0.7338. However, it is important to note that accuracy alone may not be sufficient to evaluate the performance of a model, and other evaluation metrics such as precision, recall, and F1-score should also be considered.

Further analysis of the evaluation metrics revealed that Logistic Regression has the highest precision score of 0.6491, while Decision Tree has the highest recall score of

0.7455. Precision represents the proportion of true positive predictions out of all positive predictions, while recall represents the proportion of true positive predictions out of all actual positive instances in the dataset. These results suggest that depending on the specific problem and dataset, the choice of algorithm may vary based on the evaluation metrics of interest.

5. Visualisation of results

The presented visualization is a bar plot that compares the performance of three machine learning algorithms, Logistic Regression, Decision Tree, and Random Forest, using four evaluation metrics: Accuracy, Precision, Recall, and F1-score.



The plot shows that the Decision Tree algorithm has the highest accuracy score, followed by Logistic Regression and Random Forest. However, Logistic Regression has

the highest precision score, while Decision Tree has the highest recall score. The F1-score of all three algorithms is relatively similar.

6. Conclusion

In conclusion, we have evaluated the performance of three machine learning algorithms, Logistic Regression, Decision Tree, and Random Forest, on a dataset to predict diabetes. The evaluation results showed that the Decision Tree algorithm had the highest accuracy score, followed closely by Logistic Regression and Random Forest. However, it is important to note that accuracy alone may not be sufficient to evaluate the performance of a model, and other evaluation metrics such as precision and recall should also be considered.

Further analysis of the evaluation metrics revealed that Logistic Regression had the highest precision score, while Decision Tree had the highest recall score. Precision represents the proportion of true positive predictions out of all positive predictions, while recall represents the proportion of true positive predictions out of all actual positive instances in the dataset. These results suggest that depending on the specific problem and dataset, the choice of algorithm may vary based on the evaluation metrics of interest.

7. Reference

1. Diabetes Prediction Using Machine Learning: A Review - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7422453/

2. A comparative study of machine learning algorithms for diabetes prediction - https://www.sciencedirect.com/science/article/pii/S2352340919310043

3. Predicting Diabetes Using Machine Learning Techniques - https://ieeexplore.ieee.org/document/8240032

4. A Machine Learning Framework for Diabetes Prediction - https://link.springer.com/chapter/10.1007/978-3-030-00934-2_17

5. Predictive modeling of diabetes using machine learning techniques - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7643402/