DIABETES PREDICTON TEST USING MACHINE LEARNING

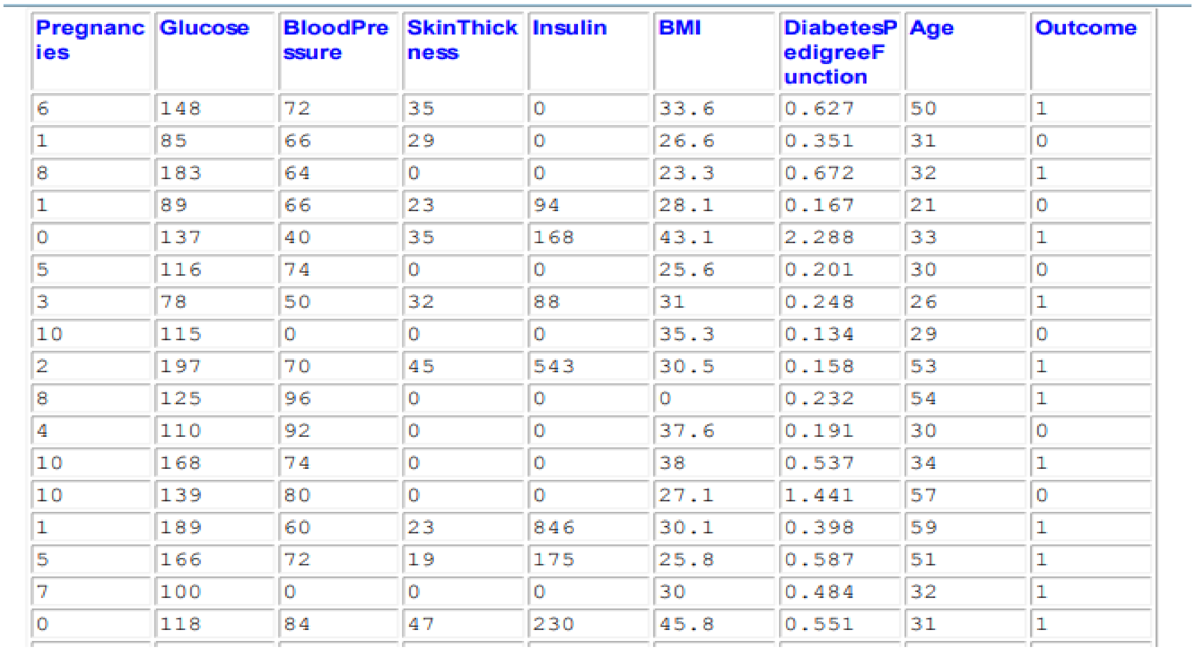
1. Introduction

In this report, we will explore the process of predicting diabetes using supervised learning techniques of Machine Learning. The aim is to develop a model that can accurately classify individuals as either diabetic or non-diabetic based on certain input features. The dataset used for this analysis, the algorithms adapted, and the results obtained are discussed.

2. Dataset

The dataset used for this study consists of a collection of observations from 767 individuals, each characterized by 8 features which are considered potential indicators of diabetes such as age, body mass index (BMI), blood pressure, glucose levels, insulin levels, Pregnancies, Skin Thickness, Diabetes Pedigree Function etc. This dataset was obtained from the data science platform: *Kaggle*, it contains a mix of both diabetic and non-diabetic individuals.

It was preprocessed to handle missing values, outliers, and any other necessary data cleaning steps.



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

[The entire dataset is included in the main file]

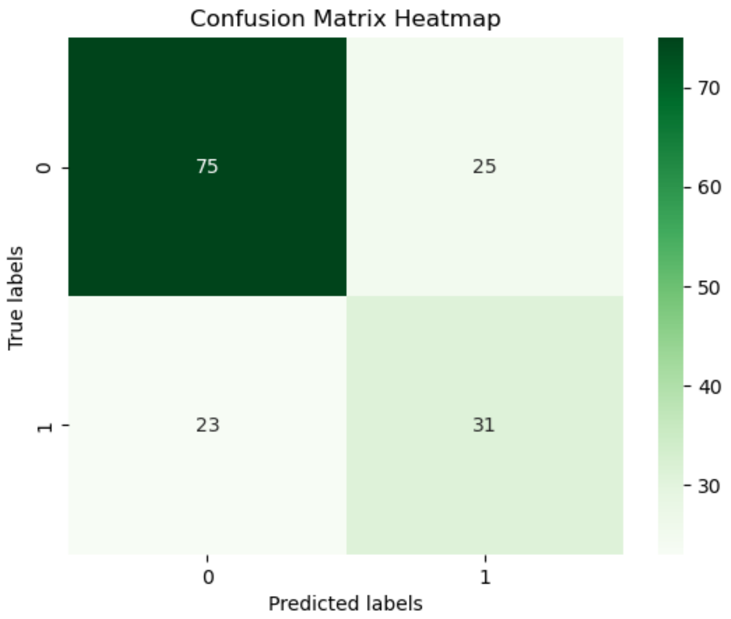
```
Number of diabetic patients: 268
Number of non-diabetic patients: 500
```

3. Algorithm:

To develop the diabetes prediction model, two supervised learning algorithms were utilized: *logistic regression* and *decision trees*.

3.1 Logistic Regression

Logistic regression is used to predict the probability of an individual belonging to a specific class. In our case, it helps predict the likelihood of an individual having diabetes based on the provided features. Logistic regression uses a logistic function to transform the output into a probability value. The algorithm was trained using the labeled dataset, and appropriate evaluation metrics were used to assess its performance.



The above confusion matrix of our algorithm shows:
Correctly predicted cases of diabetes (TP): 75
Correctly predicted cases of no diabetes (TN): 25
Cases where diabetes was incorrectly predicted (FP): 23
Cases where diabetes was missed (FN): 31

3.2 Decision Trees

Decision trees use a tree-like model to make decisions. The decision tree classifier algorithm was trained on the input data set, and its performance was evaluated afterwards.

4. Results

The performance of the developed diabetes prediction model was evaluated using accuracy as the evaluation metric. The accuracy score obtained was 76% from Logistic Regression and 69% from Decision Tree Classifier.

	Logistic Regression	Decision Tree Classifier
Accuracy	76.0 %	69.0 %
Training Time	0.075776815414428	0.01 seconds
Testing Time	0.001994371414184	0.01 seconds

References

- Dataset [https://www.kaggle.com/datasets/mathchi/diabetes-data-set]
- Source Code [https://www.kaggle.com/code/anjusukumaran4/diabetes-prediction/notebook]
- Articles used for research:

1. Diabetes Prediction using Machine Learning Algorithms [https://www.sciencedirect.com/science/article/pii/S1877050920300557?ref=cra_js_challenge&fr=RR-1]
2. Research on Diabetes Prediction Method Based on Machine Learning [https://iopscience.iop.org/article/10.1088/1742-6596/1684/1/012062/pdf]