Utilizing Machine Learning Techniques to Predict Type 2 Diabetes Onset: A Comparison of Logistic regression, Naive Bayes, and Deep Learning Models

Estéfano Vidal.

May 2023

Abstract. This project utilizes machine learning techniques, specifically Logistic regression, Naive Bayes, and Deep Learning, to predict the onset of Type 2 Diabetes based on various features such as lifestyle, self-perception of health, disease history, and other medical conditions. An exploratory data analysis was performed on a dataset comprising over 253,000 entries. To train the machine learning models, the dataset was balanced in a 1:1 ratio between diabetics and non-diabetics. The results of this project indicate that the Logistic regression algorithm outperforms Neural network and Naive Bayes models for this particular problem.

Keywords: Machine Learning · Deep Learning · Data Analysis.

1 Introduction

Diabetes is a chronic metabolic disorder that affects millions of people worldwide. The disease is characterized by an elevated level of blood glucose, resulting from either a deficiency of insulin secretion or insulin resistance. The long-term complications of diabetes can be severe, including cardiovascular disease, kidney failure, blindness, and neuropathy. Therefore, prevention and early diagnosis are crucial in managing the disease and preventing its complications.

Machine learning has emerged as a promising tool in healthcare, particularly in the field of diabetes diagnosis and prophylaxis. Machine learning algorithms can analyze large volumes of data and identify patterns that are difficult for humans to detect, leading to more accurate and timely diagnosis, as well as showing the tendency for developing the disease. In this report, we will explore and compare how effective machine learning algorithms are, specifically Logistic regression, Naive Bayes, and Deep Learning.

2 Methodology

2.1 Data Source and Characteristics

The dataset was obtained from the Center for Disease Control and Prevention - 2014 Behavioral Risk Factor Surveillance System Survey Data and Documentation (https://www.cdc.gov/brfss/annual_data/annual_2014.html) and can be publicly accessed. This dataset was modified to include only men and non-pregnant women over 20 years of age, and included 21 independent variables, both nominal and numerical, with the dependent variable being clinically diagnosed with diabetes or pre-diabetes.

Number	Variable	Type	Values
1	High blood pressure	Nominal	Yes, No
2	High cholesterol	Nominal	Yes, No
3	Checked cholesterol (last year)	Nominal	Yes, No
4	BMI	Numerical	12 to 98
5	Smoker	Nominal	Yes, No
6	Suffered stroke	Nominal	Yes, No
7	Has hearth disease	Nominal	Yes, No
8	Does physical activity	Nominal	Yes, No
9	Has fruits once a day	Nominal	Yes, No
10	Has vegetables once a day	Nominal	Yes, No
11	Heavy drinker	Nominal	Yes, No
12	Has insurance	Nominal	Yes, No
13	Skipped doctor because of cost (last year)	Nominal	Yes, No
14	Overall health perception	Nominal	1 to 5
15	Days being mentally unwell (last month)	Numerical	0 to 30
16	Days being physically unwell (last month)	Numerical	0 to 30
17	Has difficulty walking	Nominal	Yes, No
18	Sex	Nominal	Male, Female
19	Age	Nominal	1 to 13
20	Education	Nominal	1 to 6
21	Income	Nominal	1 to 8

Table 1: Description of Variables. Variable 14: Health perception, ranging from "Excellent" (1) to "Poor" (5). Variable 19: Age groups, coded as follows: 1 = 20-24, 9 = 60-64, 13 = 80 or older. Variable 20: Education level, coded as follows: 1 = Never attended school or only kindergarten, 2 = Elementary, 6 = College graduate. Variable 21: Income level, coded as follows: 1 = less than \$10,000, 5 = less than \$35,000, 8 = \$75,000 or more

2.2 Data Preprocessing

The dataset contained a mixture of categorical and numerical data, mostly binary values. However, there were also non-dichotomous data with non-Gaussian distributions, so it was necessary to apply preprocessing to the database. This preprocessing consisted of quantile normalization and feature scaling, which were performed using the Python library "Sklearn 1.2.2."

2.3 Data Cleaning and Exploration

The dataset used in this project contained a few outliers, which were handled using techniques such as imputation and removal. Exploratory Data Analysis (EDA) was performed to understand the relationships between the features and the target variable. Visualizations such as box plots, histograms, and heatmaps were used to identify any correlations or patterns within the data.

The exploratory data analysis revealed significant differences between individuals with and without diabetes. Diabetics had a significantly higher Body Mass Index (BMI) and were more likely to experience various health issues, such as high blood pressure, high cholesterol levels, heart disease, stroke history, and difficulty walking. Lifestyle was another discriminating factor; diabetics generally had a lower tendency to consume at least one fruit or vegetable per day and to exercise, while having a higher tendency to smoke. Surprisingly, non-diabetics were more prone to heavy drinking, although this correlation may not necessarily indicate a causal relationship. Socioeconomic factors were also related to the disease. The median yearly income level for diabetics was around $35,000, while for those without the disease, it was $50,000. It is unclear whether this relationship is causal or which direction it goes. However, it might be related to the general well-being of the person and their ability to function. According to the analysis, diabetics have a worse self-perception of their general health, including physical and mental health. They also show a higher frequency of illness. These factors could potentially impact job opportunities. Finally, there was no significant difference in the tendency to develop the disease based on sex.

In order to reduce multicollinearity, it was necessary to eliminate variables that exhibited strong positive or negative correlations. To identify these variables, a heatmap was created to display those that were highly correlated. Additionally, an analysis of the variance inflation factor (VIF) was conducted. Ultimately, seven variables were removed due to high VIF values. These variables were: 'Income', 'Education', 'Age', 'General health', 'Has insurance', 'Had cholesterol checked at least once', and 'BMI'. All of the exploratory data analysis graphs can be found in the supplementary materials accompanying this report at the end.

2.4 Algorithms

Logistic regression. Logistic regression is a widely used statistical method for binary classification problems, where the task is to predict the probability of a binary outcome based on a set of predictor variables. The approach involves modeling the relationship between the predictor variables and the binary outcome using a logistic function, which maps the input values to a probability value between 0 and 1. This logistic function has an S-shaped curve and allows us to model the probability of the binary outcome as a function of the predictor variables. The Logistic regression model estimates the parameters of the logistic function based on the training data, using an optimization algorithm that minimizes the difference between the predicted probabilities and the actual binary outcomes. The model's output is a probability score, which can be converted into a binary prediction by applying a threshold value. It was implemented using the Python library "Sklearn 1.2.2.".

Naive Bayes. Naive Bayes is a frequently utilized classification technique in machine learning, which involves computing probabilistic outcomes by tallying frequencies and then merging them with the information provided in the dataset. This algorithm leverages Bayesian theorem and assumes that all attributes are unrelated to each other and rely solely on variable values of classes. However, in real-world applications, this conditional independence assumption rarely holds true, leading to less accurate results. It was implemented using the Python library "Sklearn 1.2.2.".

Neural network. A Neural network is a type of machine learning model inspired by the structure and function of the human brain. It is composed of interconnected nodes or "neurons" that are organized into layers, and it can learn to recognize patterns and relationships in data by adjusting the strength of the connections between neurons. This Neural network is composed of 14 input neurons and 2 output neurons, as it is a binary classification problem. It also includes two hidden layers, each consisting of 70 neurons. The "Relu" activation method is used for the hidden layers and the "Softmax" activation method is used for the output layers. An "Adam" optimizer and a "Sparse Categorical Crossentropy" loss function were selected as they have yielded the best results in previous experiments, an early stop parameter was also implemented to prevent overfitting. It was implemented using the Python library "Tensorflow 2.12.0"

2.5 Metrics

Four different metrics were utilized to assess the performance of the algorithms employed in this project. The formal definition of each metric is given below. Here, TP represents True Positive, FP denotes False Positive, TN represents True Negative, and FN stands for False Negative.

Precision. Precision measures the proportion of TP over the total number of predicted positive results (TP + FP). A high precision score indicates that the model has a low false positive rate, meaning that it rarely predicts a positive result when the actual result is negative.

$$Precision = \frac{TP}{TP + FP}$$

Recall. Also known as sensitivity, it evaluates the ability of a model to correctly identify all positive instances in a dataset. It is calculated as the ratio of TP to the sum of TP and FN, this represents the total number of positive instances that the model failed to identify. A high recall score indicates that the model is good at identifying positive instances, while a low score suggests that it may be missing some positive instances.

$$Recall = \frac{TP}{TP + FN}$$

Accuracy. It measures the percentage of correctly predicted outcomes out of all the predictions made. More formally, accuracy is defined as the ratio of the number of correct predictions (TP and TN) to the total number of predictions. A high accuracy score means that the model is making correct predictions on a high percentage of the data points.

$$Accuracy = \frac{TP + TN}{TP + FP + TP + TN}$$

F-Meassure. It evaluates the performance of a classification model. It is a harmonic mean of precision and recall, which takes into account both false positives and false negatives. It combines precision, which is the ratio of true positives to the total number of positive predictions, and recall, which is the ratio of true positives to the total number of actual positive cases. A high F-measure score indicates that the model has both high precision and high recall, which means it is effectively identifying both positive and negative cases.

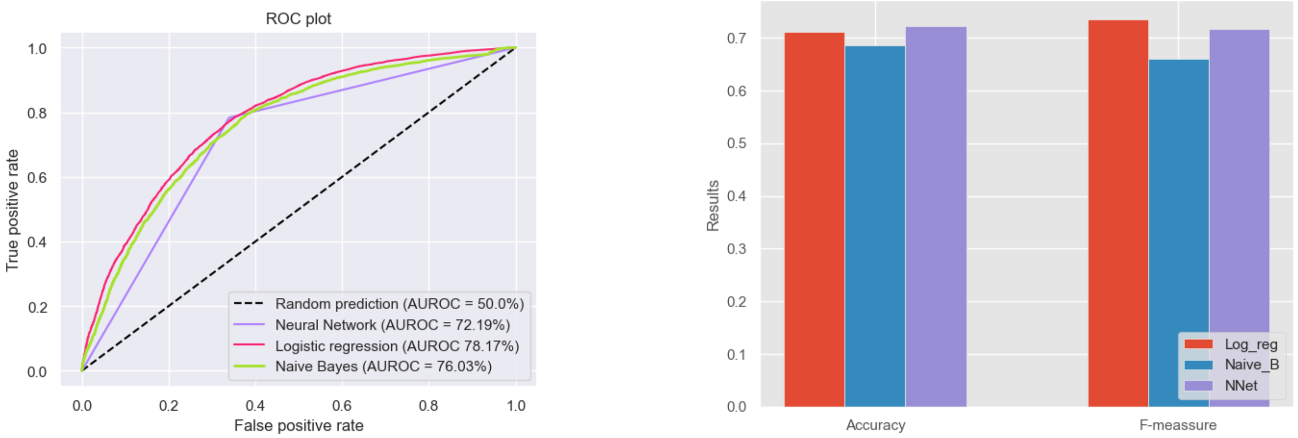
$$F - Meassure = \frac{2 * Recall * Precision}{Precision + Recall}$$

3 Results

The area under the receiver operating characteristic curve (AUROC) values for each machine learning model are as follows: 78.17% for Logistic regression, 76.03% for Naive Bayes, and 72.19% for Neural network (Fig. 1.a). The Logistic regression method proved to be the most effective for this particular problem, with the ability to predict the incidence of diabetes with an average precision score of 76.1% and a recall score of 71.1% . In contrast, the Neural network scored an average precision score of of 64.4% and a recall score of 78.1%. Naive Bayes had the worst performance, with an average precision score 73.5% and a recall score of 60.1% (Fig. 2)

Remarkably, the Neural Network halted training after a mere 9 epochs, as evidenced by the accuracy curve of the validation batches dropping below that of the training set, and the corresponding rise of the loss curve of the validation set (Fig. 3). The absence of an early stopping parameter might have led to overfitting of the model.

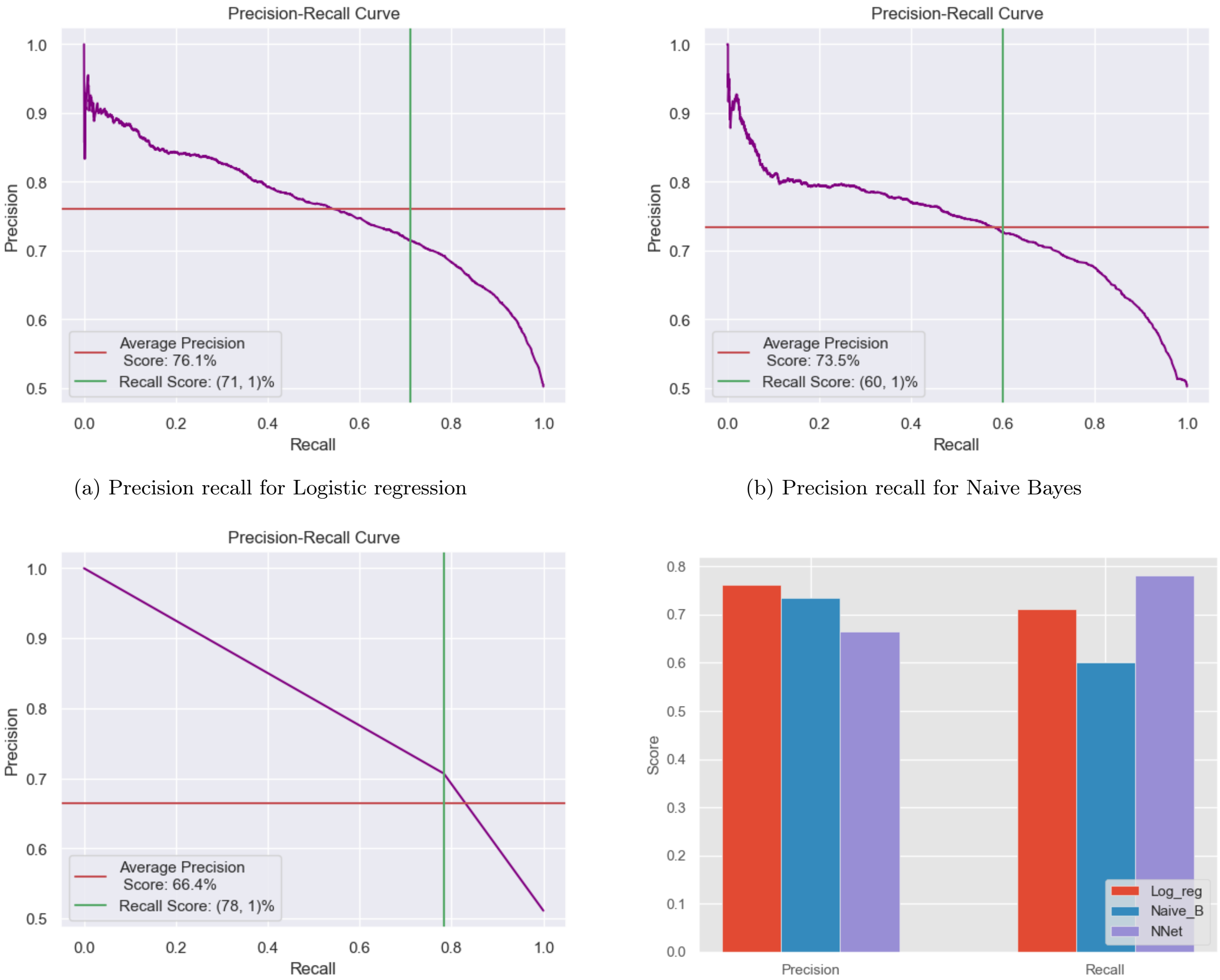
These results provide valuable insights into the performance of various machine learning models for disease diagnosis and prophylaxis. The confusion matrix for each model can be found in the supplementary materials accompanying this report at the end.



(a) ROC plot comparing the three machine learning models

(b) General Comparison of the Accuracy and F-Meassure scores

Fig. 1: (a) General ROC plot. (b) Comparison between the Accuracy and F-Meassure metrics for the three models



(a) Precision recall for Logistic regression

(b) Precision recall for Naive Bayes

(c) Precision recall for Neural network

(d) General Comparison of the Precision and Recall scores

Fig. 2: (a, b, c): Precision-Recall curves for Logistic regression, Naive Bayes, and Neural network, respectively. (d) Comparison between the Precision and Recall metrics for the three models

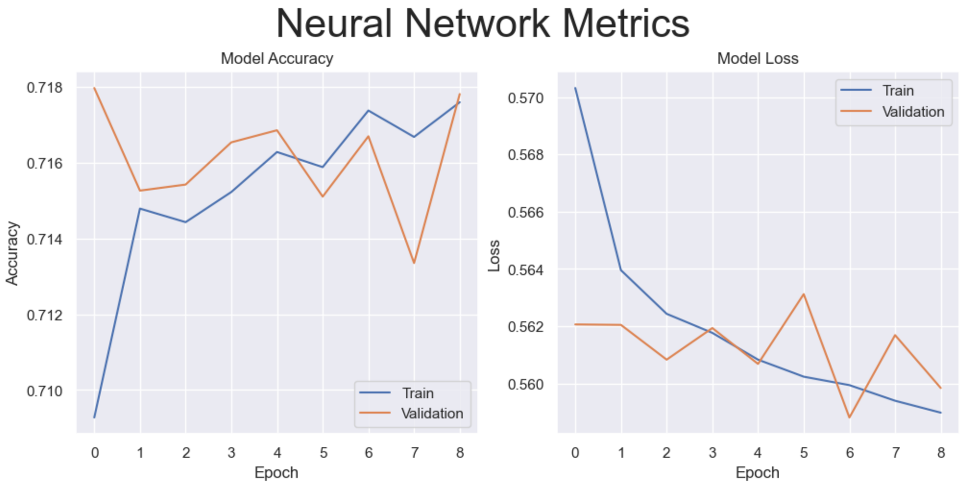
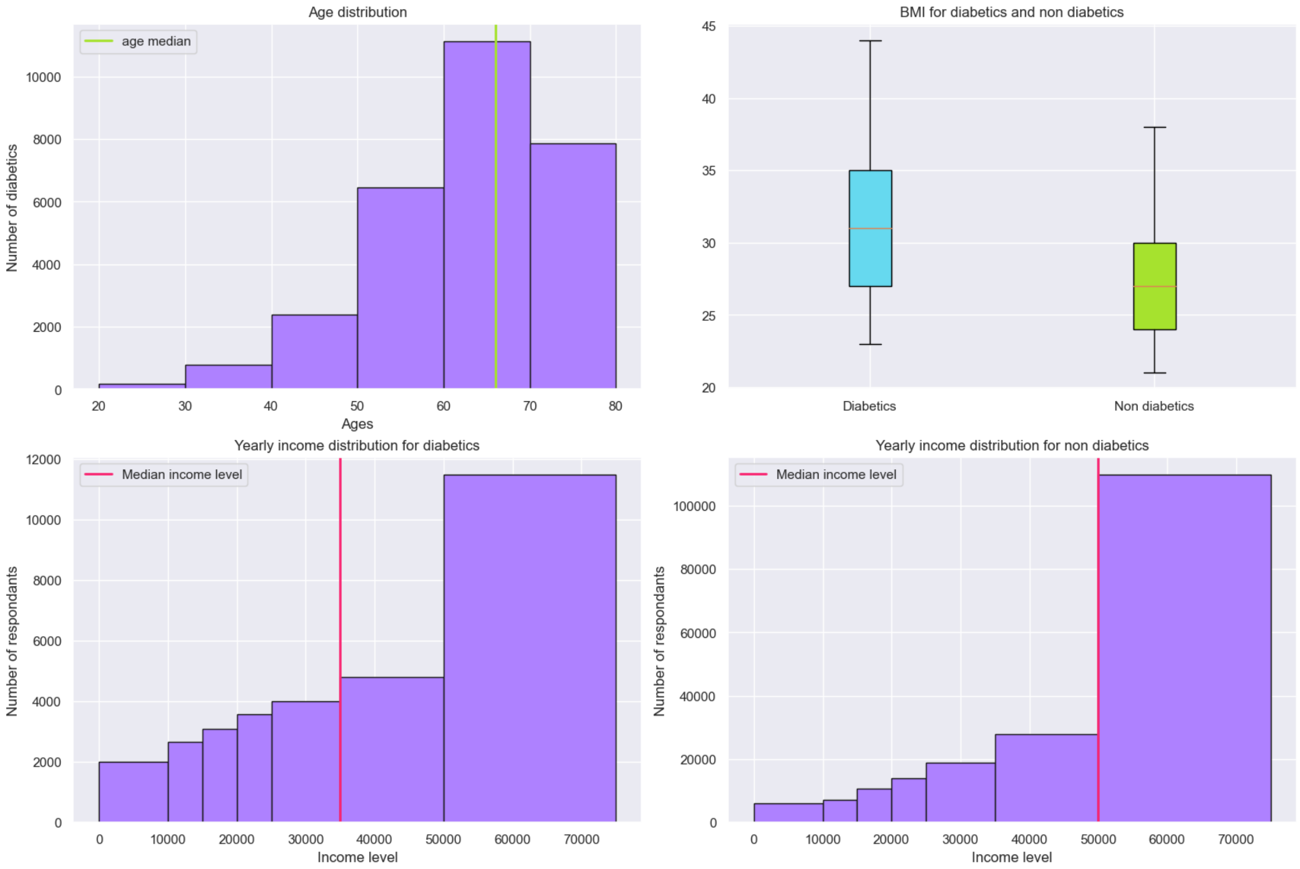


Fig. 3: Accuracy and loss values for the training and validation sets used by the Neural network

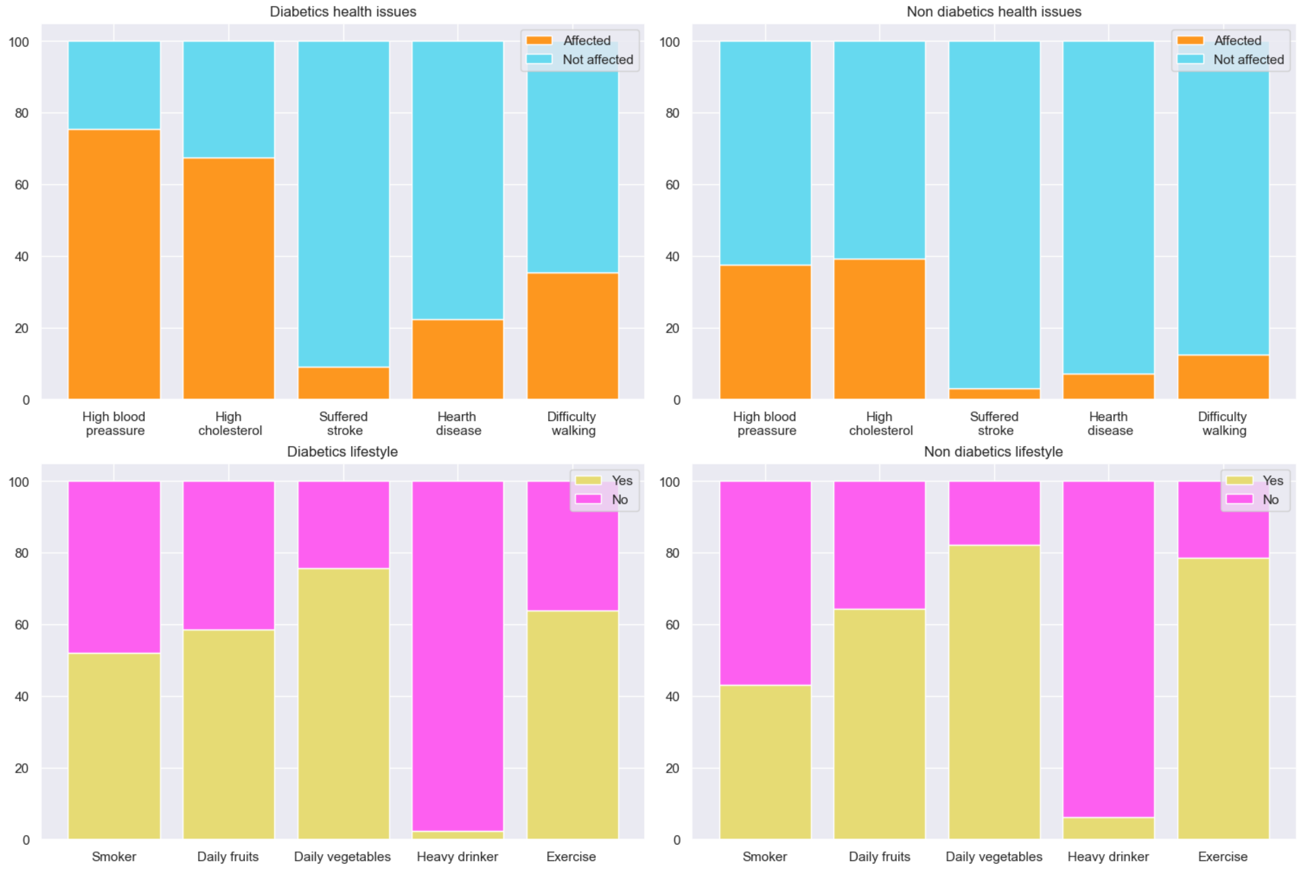
4 Conclusion and Future Work

This project showcases the usefulness of machine learning for disease prediction by using three different methods, with Logistic regression being the most effective outperforming the other models in most metrics. The Naive Bayes algorithm was outperformed in both accuracy and F-measure by both the Logistic Regression and Neural Network algorithms. The Neural Network slightly outperformed the Logistic Regression in terms of accuracy; however, the Logistic Regression exhibited better F-measure, which compensated for the small difference in accuracy (Fig 1.b). Therefore, overall, the Logistic Regression method is better suited for this particular problem. Further work could involve exploring other statistical methods and improving the deep learning algorithm, another option to improve the quality of the results is to take into consideration different variables. The insights gained from this project can also provide valuable information about factors related to diabetes, such as lifestyle, medical history, physical and mental health, and socioeconomic status.

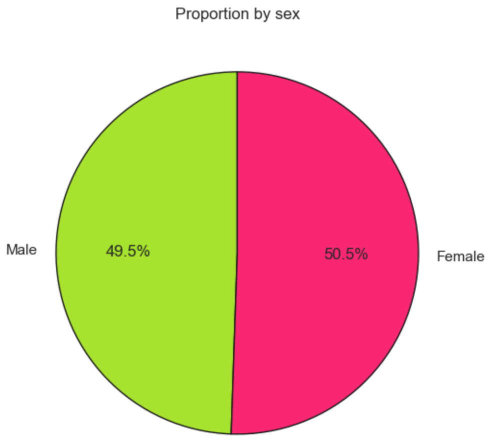
5 Supplementary Materials



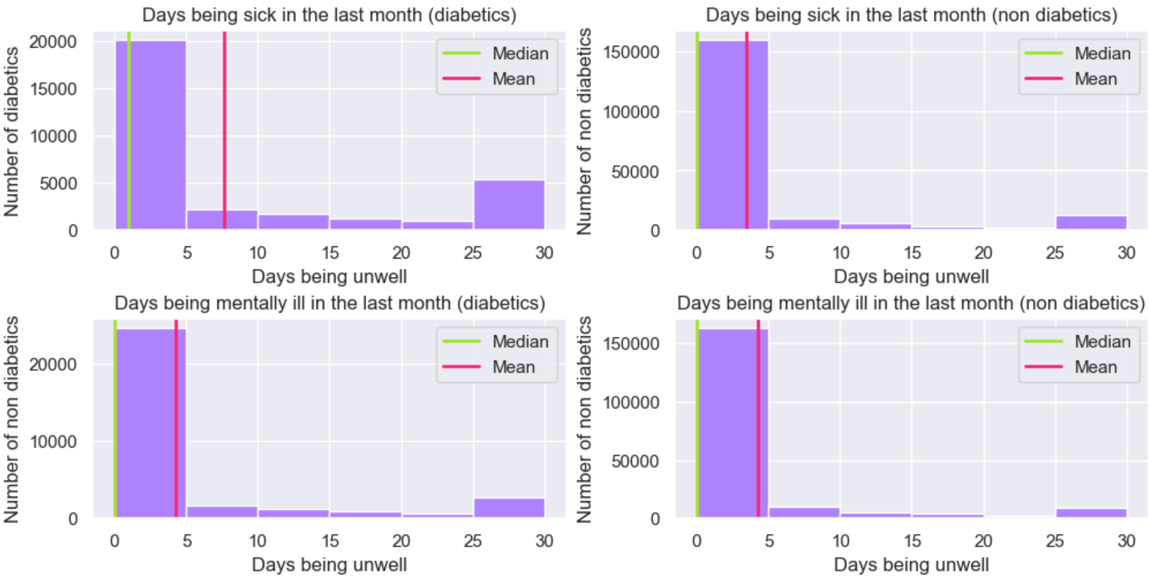
Complementary Fig. 1: From left to right: Age distribution for diabetics, BMI comparison between diabetics and non-diabetics, Yearly income distribution for diabetics, Yearly income distribution for non-diabetics



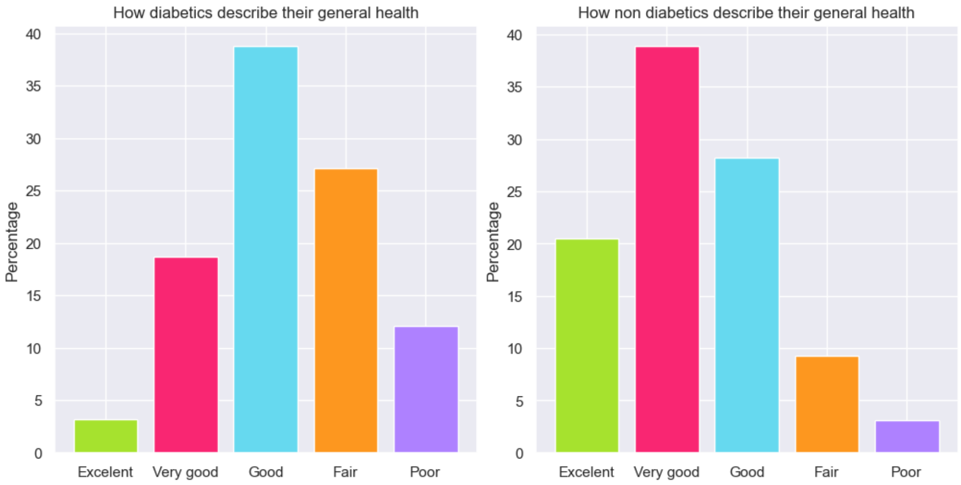
Complementary Fig. 2: Top to bottom: Barchart showing the percentage of diabetics (left) and non-diabetics (right) affected by health issues. Barchart showing the percentage of diabetics (left) and non-diabetics (right) associated with lifestyle habits



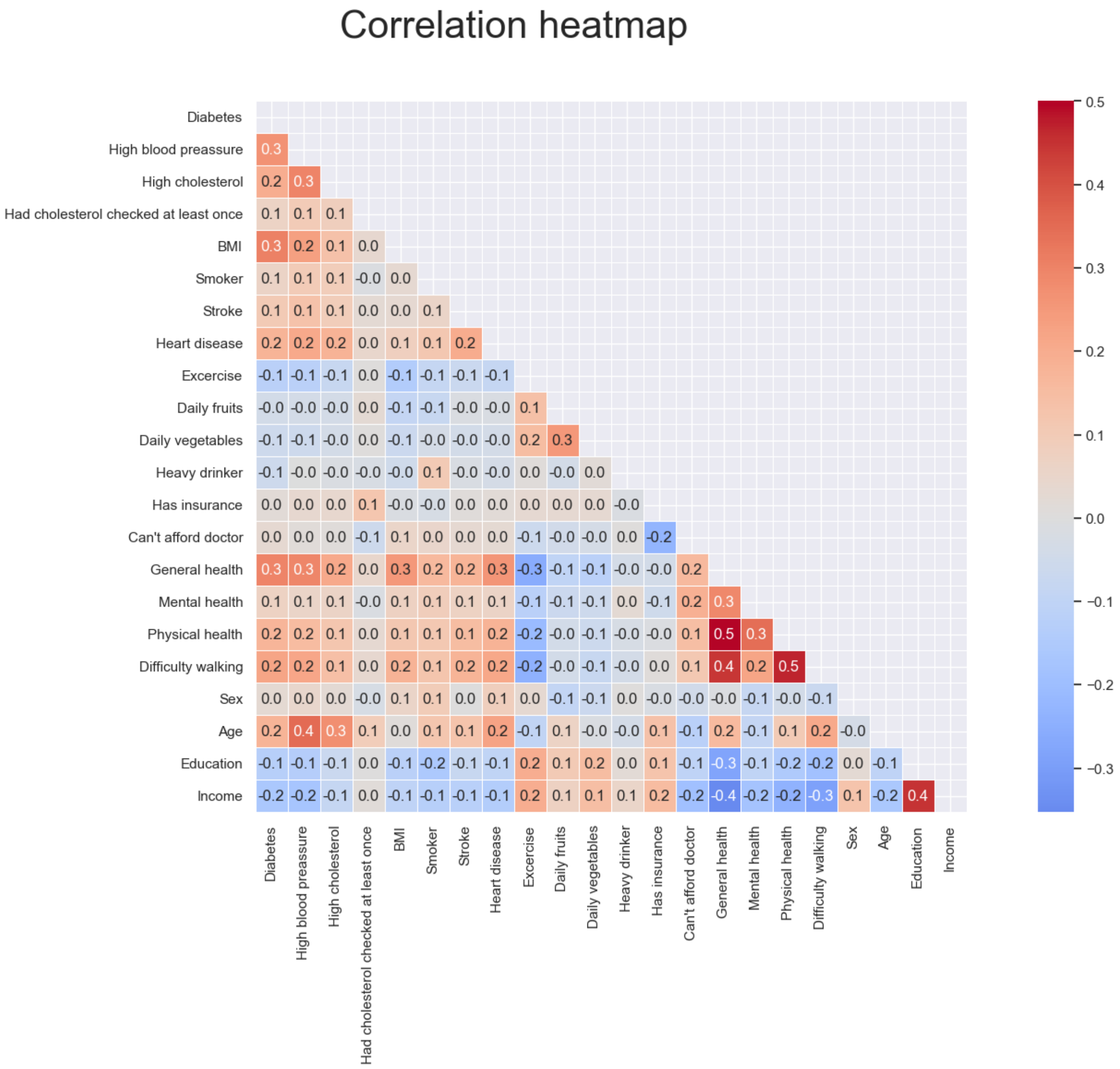
Complementary Fig. 3: Incidence by gender



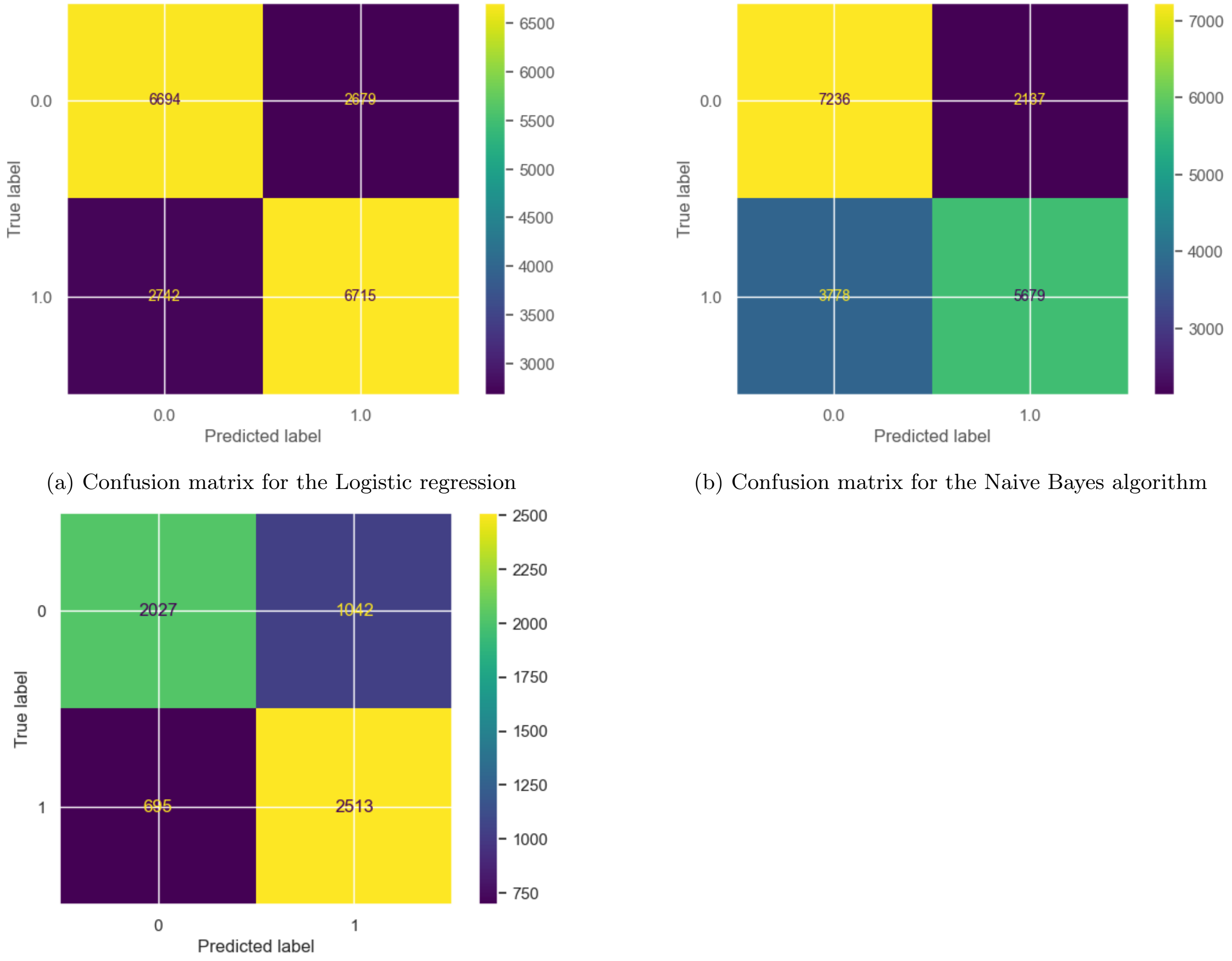
Complementary Fig. 4: From top to bottom: Comparison between diabetics (left) and non-diabetics (right) of the number of days physically ill within the last month. Comparison between diabetics (left) and non-diabetics (right) of the number of days mentally ill within the last month.



Complementary Fig. 5: Self-perception of overall health among diabetics (left) and non-diabetics (right)



Complementary Fig. 6: Heatmap showing correlations between variables



(a) Confusion matrix for the Logistic regression

(b) Confusion matrix for the Naive Bayes algorithm

(c) Confusion matrix for the Neural network

Fig. 7: Confusion matrix for each algorithm