**Machine learning For Predictive Analytics – Assessment 2 – Build A Classifier**

**Harry O’Donnell C20379081 TU856**

**Introduction:**

Diabetes dataset shows correlation between medical characteristics and the risk of having diabetes. The dataset is stored in a comma separated values file. This file has 9 columns and 768 rows.

Column names and meaning:

* Pregnancies- number of pregnancies
* Glucose - level of blood glucose
* BloodPressure- blood pressure
* SkinThickness- thickness of the skin
* Insulin- level of insulin in blood
* BMI- calculated BMI
* DiabetesPedigreeFunction- likelyhood of diabetes based on family history
* Age- age of patient
* Outcome- 0 for not being diabetic and 1 for being diabetic

A screenshot of a graph

Description automatically generated

**Data Cleaning and Pre-processing:**

**Issues with the Data:**

I removed the zero values because in certain medical measurements they are not feasible (e.g., zero blood pressure). Replacing them with NaN allows for proper handling during imputation.

Median is chosen over mean for its robustness against outliers. This is important as features like 'Insulin' have wide ranges and potential outliers.

- Checked the missing or zero values in key columns.

A screenshot of a computer

Description automatically generated

- Replaced zero values with NaN and imputed missing data with the median value of each column.

A computer code with red text

Description automatically generated

A close-up of a code

Description automatically generated

- Verified the absence of missing values after imputation.

A screenshot of a computer

Description automatically generated

**4. Data Visualization / Analysis**

- Created a pie chart to visualize the proportion of diabetic and non-diabetic individuals.

A pie chart with numbers and a few percentages

Description automatically generated

This shows a very high percentage of people have diabetes present (35%), I created this Pie chart to get a feel for breakdown of the target feature.

I used a heatmap to display the correlation between all the variables in the dataset..

A blue and white squares with black text

Description automatically generated

Based on the correlation matrix, it’s clear that **Glucose levels** has the greatest effect on the Diabetes outcome.

There’s also a strong correlation between **BMI** and **Age** and the outcome.

There’s also a strong correlation between **insulin** and **glucose** levels and between BMI and **Blood pressure**.

This Box plot shows the relationship between the glucose levels and whether the person is diabetic, this graph successfully shows the strong correlation between the glucose levels and the outcome. There’s a much higher minimum for people who are diabetic which makes sense with what would be expected from a diabetic person, the blood sugar stays higher for a longer period of time.

A diagram of diabetes

Description automatically generated

The bar plot provided shows the relationship between glucose levels and insulin levels. It suggests that as glucose levels increase there is a wide range of insulin responses. When sombeody has diabetes, the body is unable to manage insulin and glucose levels correctly. The plot suggests that individuals with higher glucose levels often have higher insulin levels.

A chart showing a graph of glucose and insulin

Description automatically generated with medium confidence

The bar plot shows a comparison of blood pressure across three BMI categories: underweight/normal, overweight, and obese. The bars represent the average blood pressure within each BMI category. They suggest that blood pressure tends to increase with higher BMI categories.

A graph of different colored rectangular shapes

Description automatically generated with medium confidence

The scatter plots provided represents glucose levels as a function of age in three different BMI categories: underweight/normal, overweight, and obese. Each dot represents an individual's glucose level at their age.

Theres an increase in both the glucose levels and the number of data points when moving to the higher BMI category. Higher BMI is associated with higher glucose levels, which can indicate a greater risk for diabetes.

A graph of bmi and age

Description automatically generated

**Preparing Data for Classifier**

I separated the dataset into features (X) and the target variable (y). The target variable 'Outcome' suggests the presence (1) or absence (0) of diabetes. Which makes it a categorical variable suitable for logistic regression.

The descriptive feature (X) were then normalised using StandardScaler from the scikit-learn library. This normalisation ensures that each feature contributes equally to the model's decision boundary.

- Split the dataset into features and the target variable.

A close-up of words

Description automatically generated

**Choosing the Classifier**

**Logistic Regression** was chosen as the classifier.

I decided to use Logistic Regression as my model because it's well-suited for predicting 'yes' or 'no' outcomes, which in this dataset is if the person has diabetes or not.

This method is great for calculating the likelihood of different outcomes, which helps us make clear choices.

It's also really helpful for understanding which factors, like glucose levels or body weight, are important for predicting diabetes.

**A screenshot of a computer code

Description automatically generated**

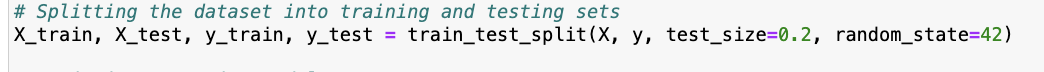
I developed the logistic regression classifier with an increased number of iterations, this was important as it allowed the model more rounds to learn from the data, I set max\_iter to 1000.

**Evaluation**

Testing was performed by splitting the dataset into a training set (80%) and a test set (20%). I trained the logistic regression model on the training set and evaluated it on the test set.

The following metrics were used to test model performance:

* **Accuracy**: Percentage of total correct predictions.
* **Recall**: Ability of the model to find all relevant cases (sensitivity).
* **Precision**: Proportion of positive identifications that were correct.
* **F1 Score**: The Harmonic mean of precision and recall.
* **Confusion Matrix**: Provided a breakdown of true positives, true negatives, false positives, and false negatives.



A black text on a white background

Description automatically generated

**Results**

A screenshot of a computer code

Description automatically generated

A screenshot of a computer

Description automatically generated

The logistic regression model returned the following results on the test set:

* **Accuracy**: 75.32%
* **Recall**: 62%
* **Precision**: 67%
* **F1 Score**: 64%
* **Confusion Matrix**:

**True Negatives**: 82

**False Positives**: 17

**True Positives**: 34

**False Negatives**: 21

The logistic regression model showed a solid performance, correctly predicting diabetes status in 75.32% of cases, which is the accuracy of the model. It successfully identified 62% of the actual diabetic cases, which was shown by the recall. When the model predicted diabetes, it was correct 67% of the time (precision). The F1 score, which balances precision and recall was 64%, which shows a reasonable compromise between the two. But, the model did miss diagnose diabetes in 21 individuals, showing that the model still makes errors.