Data Mining and Foundation of AI

Assessment 1

100571823

# Introduction

The selected dataset for this project was the Healthcare-Diabetes Dataset.

Diabetes is a chronic metabolic disorder which can lead to serious heath complications if left untreated. Identifying the key factors that contribute to positive diagnosis allow for early detection leading to improved patient outcomes.

This report explores to answer the question **“Which three factors are most likely to cause a positive diabetes diagnosis?”**. To answer this question, we will be using a publicly available diabetes dataset, containing varying health indicators such as blood pressure, glucose levels, BMI, insulin levels and age to determine the top three most significant predictors of a diabetes diagnosis using data-driven machine learning.

# Preprocessing & Exploratory Data Analysis

## Data Preprocessing

To create an accurate model capable of providing us meaningful insights, the data needs to be effectively preprocessed. The preprocessing stage ensures a clean and well-structured dataset ready for analysis.

#### Handling Unnecessary Columns

Firstly, the dataset was reviewed and the column containing the patient ID was removed. This was done because the column was not necessary for the creation of the algorithm. The columns purpose was solely that of an identifier and carried no value in regards to the feature importance. Since the analysis focuses on trends across the dataset opposed to individual cases, the column would not enhance the model and could introduce unneeded complexity.

#### Managing Missing values

The data was checked for any missing values and any values found to be missing were imputed using the median data of their respective column, this method was chosen over mean imputation because the median is less sensitive to outliers, therefore preserving the distribution of data within the dataset. Median imputation also ensures that any skewed distributions remain intact by avoiding feature representation distortion.

## Exploratory Data Analysis (EDA)

EDA was conducted to understand the feature distribution, detect any patterns and/or within the dataset. the analysis will help to identify the key data trends in which the algorithm will be trained on.

#### A group of blue and white graphs AI-generated content may be incorrect.Visualizing Feature Distributions

To determine the distribution of datapoints within the dataset histograms, boxplots, bar charts and scatter plots were created to analyze the dataset. The histograms were created in order to show the various distributions of the datapoints to determine if the data needs to be normalized or if there were any missing data values that were not processed properly during the preprocessing stage.

From the created histograms we can see that Glucose, BMI and Age all appear to be more skewed to the right, indicating that most of the values are concentrated on the lower end but there are extreme higher values.

A group of graphs with different colored squares

AI-generated content may be incorrect.

The box plots then were used to show the distribution of the columns separated by their outcome, this allows us to visualize the key factors in the dataset that differ between diabetic and none diabetic individuals.

The generated box plots show that glucose levels are significantly higher in those with a positive diagnosis (1) as well as BMI being higher in those with a positive diagnosis. The charts also show a moderate correlation between BMI and age.

#### Correlation Analysis

Correlation analysis was done on the dataset in order to determine the features with the strongest correlation to a positive outcome, as to make the model training as relevant as possible by prioritizing the strongest correlations to the target variable (positive diagnosis) for model training, features with a low correlation will be discarded from the model. This improves the model by preventing overfitting by reducing redundant features as well as helping feature selection in the model by focusing on the most relevant variables. The pearson correlation matrix was used as it is efficient and effective are measuring the relationship between two variables and quantify the correlation into a number. The closer the number is to 1, the more closely related the variables are to both being present.

A graph of a number of patients with diabetes

AI-generated content may be incorrect.

From the created bar chat it is clear that Glucose, BMI and Age are the top three columns which have the highest correlation results and therefore they will be the three columns that we will be using going forward with the model.

Finally we will create scatter plots to visualize the relationships between the features identified. The purpose if this is to help confirm the known risk factors diabetes established in the dataset, and identifying the multi-variable relationships will benefit the model building.

A graph with blue and orange dots

AI-generated content may be incorrect.

The scatter plots have identified that there is a positive correlation between Glucose and BMI, indicating that as glucose levels increase among positive outcomes, BMI also increases linearly. We can also see that Older individuals tend to have higher glucose levels as well as positive diabetes cases are more frequent among higher BMI and Age ranges. The Scatter plots have helped to identify the multi-variable relationships required to build an accurate model, which marks the end of the preprocessing and EDA stage of the model development.

# Classification analysis and solutions

The problem being addressed is that of a binary classification. This is due to the outcome being predicted as having diabetes (1) or not having diabetes (0) based on various health related datapoints. The output labels are those which are already known therefore making this **a Supervised Binary Problem.** The model cannot be categorized as unsupervised as it deals with data that contains labels. The dataset contains input features paired with output labels and the model is trained based on these labeled examples.

The dataset was split into training and testing sets at a ratio of 80-20. The larger portion allows the model to still learn effectively whilst the smaller portion is used to ensure there is enough data for a fair performance evaluation. This approach avoids overfitting and mimics real world deployment.

To develop an effective predictive model, two solutions will be trained and critically evaluated. Logistic regression and Random Forest Classification. With each solution having two contrasting strengths, we will examine the models performances leading to a justified selection of the final model.

#### Random Forest

Random Forest was selected to handle non-linear data relationships and enhance predictive performance via a process called ensemble learning. For this study, the model was set with 100 decision trees, using Gini impurity to determine the node splits. Unlike a single decision tree which can overfit data the random forest builds multiple trees on random subsets of data and averages the predictions. A strength of random forest is its feature importance ranking helping to identify the most influential predictors more accurately. Random forest does have a higher computational cost however than some more simple models. *(Parmar, Katariya and Patel, 2018)*

A graph of different colored bars

AI-generated content may be incorrect.

#### Logistic Regression

Logistic regression was selected as a baseline model for its efficiency and interpretability, it assumes a linear relationship between features and the odds of diabetes making it more effective with datasets have a clear decision boundary. Logistic Regression models provide direct coefficients to allow an easy understanding of how each feature influences predictions.

A graph of a number of people

AI-generated content may be incorrect.The model was set to a random state figure of 42, ensuring repeatability and the maximum iterations were set to 1000, which is higher than is usually set for a logistic regression model however the higher figure was set to ensure that the model has enough iterations to reach an optimal solution based on the lower value now being sufficient for analyzing the convergence of larger datasets*. (Nick and Campbell, 2007)*

#### Model Comparison and Evaluation

To assess the effectiveness of the Models, both were evaluated based on accuracy, precision, recall and f1-score. The Random Forest model achieved and accuracy rating of 99.64%, highlighting the superior predictive power the model, able to classify the instances in the dataset. The logistic re accuracy was 77.08% which struggled by comparison to correctly classify positive diabetes cases. The accuracy provides a good measure of the model performance, however it doesn’t fully capture the entire analysis of the model, therefore further analysis is required. Classification reports were generated to better quantify the overall ability of each model.

Random Forest Model Report:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Outcome** | **Precision** | **Recall** | **F1-Score** | **Support** |
| 0 | 0.99 | 1.00 | 1.00 | 363 |
| 1 | 1.00 | 0.99 | 0.99 | 191 |
|  |  |  |  |  |
| **Accuracy** |  |  | 99.64 | 554 |
| Macro avg | 1.00 | 0.99 | 1.00 | 554 |
| Weighted avg | 1.00 | 01.00 | 1.00 | 554 |

Logistic Regression Model Report:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Outcome** | **Precision** | **Recall** | **F1-Score** | **Support** |
| 0 | 0.79 | 0.89 | 0.84 | 363 |
| 1 | 0.72 | 0.55 | 0.62 | 191 |
|  |  |  |  |  |
| **Accuracy** |  |  | 0.77 | 554 |
| Macro avg | 0.75 | 0.72 | 0.73 | 554 |
| Weighted avg | 0.77 | 0.77 | 0.76 | 554 |

A graph showing a comparison of a model

AI-generated content may be incorrect.As we can see by the reports, the Random forest model scores higher across every area of the reports. Given these results it is clear that the Random forest model is the appropriate choice for diabetes prediction as logistic regressions poor recall makes it unsuitable for this particular medical applications.

#### Model Testing

Now the final model has been chosen, we will give it some test data and see the outcome it determines. The Random Forest model was tested on two new patient cases with different health indicators. The first a patient with a moderate glucose levels (120), a BMI of 30 and an age of 25 was classified as non-diabetic with a confidence percentage of 99%. Indicating that the model strongly associates the given data combination with a low risk of diabetes, whereas the second patient it was given had higher glucose levels (150), a BMI of 35 and an age of 65 was diagnosed as diabetic with a 67% positive percentage. These results indicate that the model is working as expected and is effective at determining Which three factors are most likely to cause a positive diabetes diagnosis.

# Conclusion

This report has explored the question of which three factors are most likely to contribute to a positive diabetes diagnosis by applying machine learning techniques to a publicly available diabetes dataset. Through data preprocessing, EDA and correlation analysis and the development of a Random Forest predictive model. To conclude Random Forest is the most effective model for predicting diabetes in this dataset, offering high accuracy and reliability. The findings confirm that Glucose, BMI, and Age are the top three predictors of diabetes, which is consistent with established medical knowledge.

# References

Camizuli, E. and Carranza, E.J. (2018). Exploratory Data Analysis (EDA). *The Encyclopedia of Archaeological Sciences*, pp.1–7. doi:https://doi.org/10.1002/9781119188230.saseas0271.

Gogtay, N. and Thatte, U. (2017). Principles of sample size calculation. *Indian Journal of Ophthalmology*, [online] 58(6), p.517. doi:https://doi.org/10.4103/0301-4738.71692.

Nick, T.G. and Campbell, K.M. (2007). Logistic Regression. *Topics in Biostatistics*, 404(404), pp.273–301. doi:https://doi.org/10.1007/978-1-59745-530-5\_14.

Parmar, A., Katariya, R. and Patel, V. (2018). A Review on Random Forest: An Ensemble Classifier. *International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018*, 26, pp.758–763. doi:https://doi.org/10.1007/978-3-030-03146-6\_86.

Pore, N. (2023). *Healthcare Diabetes Dataset*. [online] www.kaggle.com. Available at: https://www.kaggle.com/datasets/nanditapore/healthcare-diabetes.