**A Machine Learning Approach to Diabetes Risk Prediction Using Clinical Data**

**Birmingham City University**



**Abul Basit - 23164795**

**Contents:**

**Introduction – 3**

**Dataset Description – 4**

**Dataset overview + Statistics –5**

**Data Exploration and Visualisation-7**

**Processing Data-12**

**Model Evaluation-14**

**Conclusion-16**

**References -16**

**1. Introduction**

**Domain Description**

Chronic diseases such as diabetes over time will heavily rely on data driven tools to optimise early diagnosis and treatment of such diseases. Diabetes affects 500 million adults globally; hence, early detection is critical to prevent any serious and severe complications. Based on clinical data, using methods such as predictive analysis through machine learning allows healthcare professionals to identify at risk individuals. This project focuses on using a classification model to predict diabetes risk using patient attributes, which allows for me informed and timely clinical decisions.

**Problem Definition**

Diabtes is a major global health challenge, it affects millions gloabally; early diagnosis is essential to prevent major complications. This project adressess a real-world knowledge discovery problem; predicting diabetes risk using clinical data. Using the Pima Indians Diabetes dataset, the ask involves applying a tree classifcation model to identify pattern that distimguish diabetic from non-diabetic patients. The main aim is to support prevention through accurate, data-driven risk prediction.

**Literature Review**

Diabetes remains as a serious global health concern, with over 537 million adults affected in 2021, through analysis, diabetes is projected continued growth (WHO, 2021). As a response to tackle this, researchers are increasingly applying machine learning to improve detection and prevention. Alghamdi et al. (2017) found that models such as decision trees, logistic regression can effectively predict diabetes outcomes through structured clinical data, using the Pima Indians Dataset as an example. Likewise, Kavakiotis et al. (2017) portrayed decision trees as highly interpretable models in industries such as health care- noting value of transparency in clinical decision making.

**2. Dataset Description**

This project uses Pima Indians Diabetes Dataset, which is available on Kaggle, which was originally provided by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) or via GitHub (Jbrownlee, 2018)

Includes:

768 patients

9 attributes: 8 input features and 1 target variable

All patients are female of Pima Indian heritage, all aged 21 or older. This dataset has been used widely across diabetes risk modelling studies, this also includes physiological variables such as glucose levels, BMI and age.

All attributes are numerical, however, certain zero values in features like glucose, insulin and BMI are medically implausible and represent missing data. During processing, these must be handled appropriately.

The target variable (Outcome) is binary:

* 0= non-diabetic
* 1= Diabetic

This dataset also shows a moderate class imbalance, with approximately 65% of instances classified as non- diabetic and 35% as diabetic. Therefore, this imbalance may influence the performance of classification models and must be valued when evaluating results. The dataset is well suited for binary classification, due to its structure and format, using models such as decision trees, logistic regression, and others.

A screenshot of a medical information

AI-generated content may be incorrect.

**3. Dataset overview and Basic statistics**

**1. Importing Libraries**

The correct libraries must be imported to the notebook along with the correct dataset.

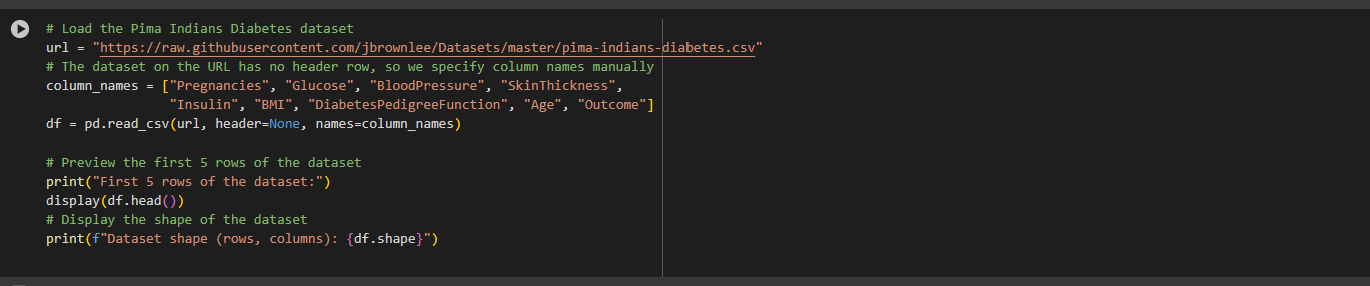
A black background with white text

AI-generated content may be incorrect.

**2. Loading the Dataset**

The dataset is loaded directly from GitHub repository into a pandas Data frame (df). Datasets contain 768 rows and 9 columns (8 features and 1 target).

* Pregnancies: Number of times pregnant
* Glucose: Plasma glucose concentration (mg/dL)
* Blood Pressure: Diastolic blood pressure (mm Hg)
* Skin Thickness: Triceps skinfold 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 which scores likelihood of diabetes based on family history)
* Age: Age in years
* Outcome: Class label (0 = no diabetes, 1 = diabetes)



To understand the structure and type of data, first 5 rows are given.

print(df.head())

A screenshot of a computer

AI-generated content may be incorrect.

**A screenshot of a computer

AI-generated content may be incorrect.**

*Figure* 1

Due to the dataset containing 768 patient records and 9 variables (8 features and 1 binary target) as seen in *Figure 1*. Hence, meeting the assignment requirements of having over 100 rows and at least 4 attributes.

Dataset information-

The df.info() output shows that’s all columns are numerical (int64 or float64), with no missing values indicated at import. However, some features have invalid zero values – representing missing data. For example, Glucose, BloodPressure etc, this cannot truly be zero for a living person. Dataset contains 768 entries for each column.

Statistical summary

From df.describe(), the ranges are noted for each feature:

* Pregnancies: range from 0 to 17, an average of ~3.8. Pregnancy counts of 0 was common due to many women in the study not being pregnant.
* Glucose: Min value= 0, but glucose cannot be 0 based in a real physiological context. Presence of 0 (minimum) in Glucose (and other columns like BloodPressure, SkinThickness, Insulin , BMI) shows missing or not accurately measured values encoded as 0- meaning glucose is ~121.7 and max= 199.
* BloodPressure: Min= 0, (invalid as a real blood pressure) mean= ~69, max= 122 mm Hg.
* SkinThickness: Min= 0, (missing data likely), mean ~20, max 99mm.
* Insulin: Min 0 (many missing), mean ~ 80 but, standard deviation is high and the 75th percentile= 127, while max= 846, indicating a long-tail distribution.
* BMI: Min 0 (missing), mean ~32, max ~67.1
* DiabetesPedigreeFunction: Varies from 0.078-2.42, mean ~0.47, (0 is a valid value for pedigree function, hence no issue).
* Age: 21-81 years old, mean~ 33.
* Outcome: Mean outcome= ~0.349, suggesting that around 34.9% of instances are diabetic positive (outcome=1); this aligns with class counts, roughly around 268/768 patients have diabetes).

Hence, it is clear from drawing results from these statistics that we need to handle zero values in certain columns as missing data.

**4. Data cleaning (Handling zero values as missing)**

As seen in the statistical overview many features contain 0 which are not valid physiological values. We will treat these zeros as missing values and thus be handled appropriately.

A screenshot of a computer program

AI-generated content may be incorrect.

As seen in this image, some columns have some numbers and will be marked as missing, by marking these as missing, they can be handled properly though imputation, instead of treating 0’s as actual values.

However, some columns such as Pregnancies, DiabetesPedigreeFunction and Age are unaffected as 0 is valid value for Pregnancies, others didn’t have zeros outside valid range.

**4. Data Exploration and Visualisation**

An Exploratory Data Analysis (EDA) will be performed to understand the dataset distributions and relationships- this will be done through visualising class balance, distributions of key features and correlations between features.

**A screen shot of a computer screen

AI-generated content may be incorrect.**

*Figure 2: Bar plot of Outcome class distribution (0 = No Diabetes, 1 = Diabetes)*

**4.1 Class distribution**

The Outcome class distribution (Fig 2) shows the data set is imbalanced. 65% of patients have outcome 0 (no diabetes) whilst 35% of patients align with outcome 1 (diabetes). Roughly this is 500 no- diabetes and 268 diabetes cases; this imbalance (35% positive rate) means baseline accuracy ( always predicting the majority class 0) would ~ 65%. This model should aim to improve significantly on this baseline.

A screenshot of a computer screen

AI-generated content may be incorrect.**4.2 Distribution of Glucose**

*Figure 3: Histogram of plasma glucose concentration among patients (with Kernel Density Estimate).*

Fig 2 shows the distribution of Glucose levels in the dataset. The distribution is roughly unimodal and rightly skewed.

* Peak around 100-130 mg/dl, which is close to the normal range for glucose.
* The tail extends towards higher glucose values ( up to ~200), most likely patients with diabetes
* Presence of a tail suggests that a subset of patients have significant elevated levels of glucose
* Dropped missing glucose values (0), from this plot for accuracy.

This distribution suggests many patients have a normal to moderately elevated level of glucose, whilst some have high levels- risk factor for diabetes.

**A screen shot of a graph

AI-generated content may be incorrect.4.3 Distribution of BMI**

*Figure 4: Histogram of Body Mass Index values for patients.*

Fig 4 shows the BMI distribution. It appears to be ordinary with only a slight right skew:

* Centre of distribution of this appears to be between BMI 30-35, showing many participants are overweight/ obese ( BMI ≥ 25 is overweight, ≥ 30 is obese). Mean BMI is ~32)
* Tail extending above BMI 40- extremely high BMI values
* No extremely low BMI values removed BMI=0 entries (missing) from the plot.

According to Fig 4, BMI distribution suggests a generally overweight population- aligning as a risk factor for diabetes. Higher BMI= Higher risk.

**A screenshot of a computer screen

AI-generated content may be incorrect.4.4 Distribution of Number of pregnancies**

*Figure 5: Bar chart of frequency of different pregnancy counts*

* 1 pregnancy and 0 pregnancy are the most common.
* As number of pregnancies increases the count generally decreases.
* 0,1 and 2 pregnancies are visible spikes.

Most women have had few or no pregnancies. Pregnancies could be a relevant feature, as having multiple pregnancies may be linked to health factors affecting diabetes risk, though it could be also emphasised by age.

**4.5 Correlation Heatmap**

It is useful to see how all features correlate with each other, thus by computing the Pearson correlation between all pairs of features. For this, all missing values will be temporarily filled with the median of each column, to compute correlations without dropping too many rows.

**A screenshot of a computer

AI-generated content may be incorrect.**

*Figure 6: Heatmap of Pearson correlation among all features (including Outcome).*

* Outcome shows the strongest positive correlations( highest correlations) with Glucose and a lower correlation to others, hence, high glucose= strong correlation with having diabetes (which is correct medically speaking)
* Pregnancies correlate positively with Age; older women tend to have more pregnancies.
* BMI has positive correlation with SkinThickness, logical, as higher body fat leads to higher skinfold thickness
* Insulin + Glucose, show positive correlations, generally people with high glucose levels= high insulin.
* Some correlations are very weak= features are largely independent. (e.g., Insulin + Pregnancies).

No pair has extremely high correlations, value near 1, to cause multicollinearity concerns in this data set; moderate correlations align with standard expectations.

**6. Preprocessing Data**

Now to prepare the data for remodelling, including handling missing data (imputation) and feature scaling. Decision trees do not require scaling, however, will include for completeness and future modelling consideration. Dataset will be split into training and test sets to analyse model’s performance against unseen data.

**A screenshot of a computer program

AI-generated content may be incorrect.**

An 80/20 split was used for training and testing. The stratify=y parameter ensures the class distribution is preserved on both splits- both train and test have ~ 35% positives, avoiding a skewed split. For reproducibility, random\_state was used.

A screenshot of a computer

AI-generated content may be incorrect.

**Imputation:** Missing values were filled, marked as NAN for Glucose, BloodPresuure, Skin Thickness, Insulin, BMI, with the median of the respective feature from the training set. With the use of training- set medians prevents data leakage into the data set. After Imputation, neither X\_train nor X\_test has missing values. Median imputation was chosen due these features having skewed distributions or outliers- median more robust than mean.

**Scaling:** Applying standardised to features (subtract mean and divide by standard deviation). Each feature in x\_train\_scaled = mean ~0 and std ~1. Using scale fitted on training data to transform test data. Scaling is not necessary but done for completeness.

Now, data is clean and pre-processed, ready for modelling.

**7. Model Training (Decision Tree Classifier)**

Training a Decision Tree Classifier to predict diabetes. A decision tree splits data based on feature values to make predictions. Using default parameters for the tree and set a random\_state for reproducibility.

*Figure 7: Simplified Decision Tree for Diabetes Prediction*

A diagram of a diagram

AI-generated content may be incorrect.

The decision tree classifier was trained with default parameters, the full tree had a depth of 14, 111 leaf nodes- showing a complex model. For, interpretability, a simplified version of the tree with a max depth of 3 is shown, highlighting most influential decision paths.

* Orange nodes- mostly non- diabetic
* Blue nodes- Mostly Diabetic
* Grey boxes- deeper branches
* Glucose is the most important predictor- used first
* Splits into other important splits- BMI, Insulin and Age
* The tree splits participants based on these values to predict whether they are diabetic or not.

A computer screen shot of a program

AI-generated content may be incorrect.

* Feature\_names= Labels for each node split using actual column names
* max\_depth=3, only shows top 3 layers to avoid clutter
* class\_names: Names for the target classes.

**8. Model Evaluation**

The model is to be evaluated on the 154 test samples that the model has not seen. Key metrics- Accuracy, Confusion Matrix and a Full Classification report- (precision, recall and F1- score).

**A screenshot of a computer

AI-generated content may be incorrect.**

*Figure 8: Confusion matrix of the Decision Tree classifier on test data.*

**Accuracy:** According to *figure 8*, the model achieves 72.08% accuracy on the test set. This is a significant improvement on the ~65% achieved by this dataset.

**Confusion Matrix** (*figure 8*): The results provided detailed counts:

* True Negatives (TN) – 85 cases predicted correctly as no- diabetes
* True Positives (TP)- 26 correct cases predicted as diabetes
* False Positives (FP)- 28 healthy cases incorrectly predicted as diabetes
* False Negatives(FN)- 28 cases wrongly predicted no diabetes when diabetic

A screenshot of a computer

AI-generated content may be incorrect.

**Interpretation:**

Model achieved a test accuracy of 72.08%, identifying non-diabetic patients is a strength of this model- Precision: 0.75, Recall 0.85. Performance on diabetic patients was significantly lower- Precision: 0.63, Recall: 0.48, portraying model’s ability to under predict positive diabetic cases; evident from confusion matrix FP.

Inaccuracy suggest model is biased towards non- diabetic cases, aligning with datasets 65/35 class imbalance; this highlights a key limitation. Although model is accurate overall- reliability in detecting diabetic patients needs improvement.

Techniques such as : class weighting, oversampling or switching to ensemble models to increase sensitivity to diabetic cases.

**ROC Curve and AUC Score**

ROC curve provides a view of the model’s performance across every classification threshold, compared to confusion matrix’s specific thresholds. AUC (area under curve) gives a value between 0.5 (no skill) – 1.0 (perfect classifier), to summarise this performance.

A screen shot of a computer screen

AI-generated content may be incorrect.

*Figure 9: ROC curve and AUC*

In *figure 9*, the ROC curve generated by the Decision Tree classifier on the Pima Indians Diabetes test set illustrates the model’s predictive ability. The blue line represents how the model performs across all classification thresholds. The dotted line is random guessing baseline, model that has no discrimination. The ROC curve rises above the dotted line, suggests that the model is better than chance. The AUC is 0.67, denoting a small level of discrimination- it beats the AUC of random guessing, which is 0.5, however it is considerably lower than the ideal value 1.0. AUC of 0.67 means the model will correctly rank a randomly chosen diabetic higher than a non-diabetic person, around 67% of the time. This highlights improvements in class separation.

This performance suggests potential limitations in the current model and data. The decision tree may be too simple to capture the complex relationships and the inherent data difficulty- overlapping class distributions. To improve the performance, tuning the tree’s hyperparameters or switching to more powerful ensemble models ( random forests), which have produced better AUC values on similar datasets.

**10. Conclusion**

This report presented a successful, through analysis and modelling process on the Prima Indians Diabetes dataset. Daa was explored, cleaned and pre-processed, including handling missing values. A Decision Tree Classifier was trained, achieving performance above the baseline; it was evaluated using metrics such as accuracy, precision, recall, confusion matric and ROC curve.

Although the model demonstrated moderate predictive performance and interpretability, results show more powerful models are required to improve accuracy. Further improvements may include experiments with other models such as logistic regression, random forest and compare their performance.

Overall, this report demonstrates a clear framework for handling binary classification problems in a structured manner, whilst also portraying the importance of critical evaluation and model selection in real world healthcare applications.

**11. References**

1. Kaggle. (n.d.). *Pima Indians Diabetes Database*. Available at: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database [Accessed 14 May 2025].
2. UCI Machine Learning Repository. (n.d.). *Pima Indians Diabetes Database*. Available at: <https://archive.ics.uci.edu/ml/datasets/pima+indians+diabetes> [Accessed 14 May 2025].
3. World Health Organization (WHO). (2021). *Diabetes*. Available at: <https://www.who.int/news-room/fact-sheets/detail/diabetes> [Accessed 14 May 2025].
4. Alghamdi, M., Al-Mallah, M., Keteyian, S. et al. (2017). Decision trees for the prediction of diabetes mellitus among patients in Saudi Arabia. *Journal of Infection and Public Health*, 10(6), pp.711–715. https://doi.org/10.1016/j.jiph.2017.08.013 [Accessed 14 May 2025].
5. Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), pp.861–874. https://doi.org/10.1016/j.patrec.2005.10.010 [Accessed 14 May 2025].
6. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I. and Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. *Computational and Structural Biotechnology Journal*, 15, pp.104–116. https://doi.org/10.1016/j.csbj.2016.12.005 [Accessed 14 May 2025].
7. ResearchGate. (n.d.). *Description of PIMA Indian diabetes Dataset*. Available at: <https://www.researchgate.net/figure/Description-of-PIMA-Indian-diabetes-Dataset_fig1_342443526> [Accessed 14 May 2025].
8. Sharma, S. (2020). Exploratory Data Analysis(EDA) and Classification on PIMA Indian Diabetes Dataset. *Medium*. Available at: https://medium.com/@ayushijaiswal2000/eda-and-classification-on-pima-indians-diabetes-dataset-15598c6e5d1f [Accessed 14 May 2025].
9. Manlangit, E., Tang, S. and Hobbs, R. (2024). An effective correlation-based data modelling framework for automatic diabetes prediction using recurrent neural networks. *PubMed Central (PMC)*. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10796578/> [Accessed 14 May 2025].
10. Islam, S.M.S., Hasan, M.K., Wang, H., Germack, H.D. and Noor-E-Alam, M. (2019). A systematic review on healthcare analytics: application and theoretical perspective of data mining. *Healthcare*, 7(1), p.20. https://doi.org/10.3390/healthcare7010020 [Accessed 14 May 2025].