Introduction

The project aimed to support doctors in identifying which factors are most predictive of developing disease and discovering patients at high risk, enabling earlier intervention. The analysis involved data cleaning, exploration, and developing a predictive model using key clinical variables.

Analysis Summary

Data cleaning

The data was cleaned analysis. Column names were corrected for consistency. Nineteen fully duplicated rows were removed. Multiple entries for the same patient ID were found where all columns matched except for 'Blood_Chemistry_III'; these were retained, as they may reflect repeated measurements important to the analysis.
Missing values in clinical features were replaced with median values.

Exploratory analysis

The outcome was imbalanced, with more patients without disease than with disease. Bivariate analysis (comparing each feature to outcome) and correlation analysis revealed that Blood_Chemistry_I, BMI and Pregnancies were the strongest predictors of disease. The exact nature of the variable Blood_Chemistry_I is not specified in the dataset, but it may represent a key laboratory marker such as blood glucose, enzyme, or other routine blood test. Further clarification of this variable's clinical meaning would be valuable for medical interpretation. These features were used for model building, with the others considered supportive.

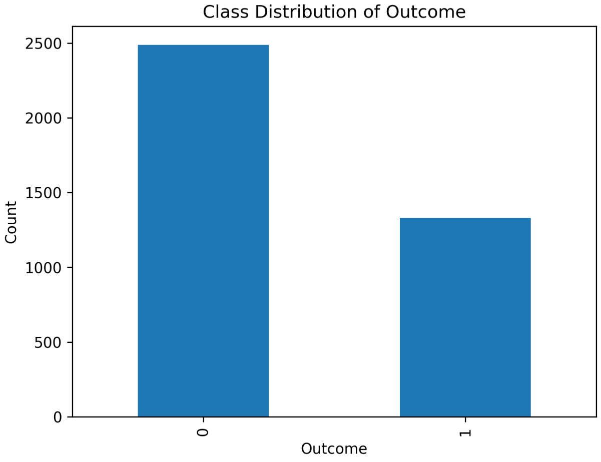

Figure 1

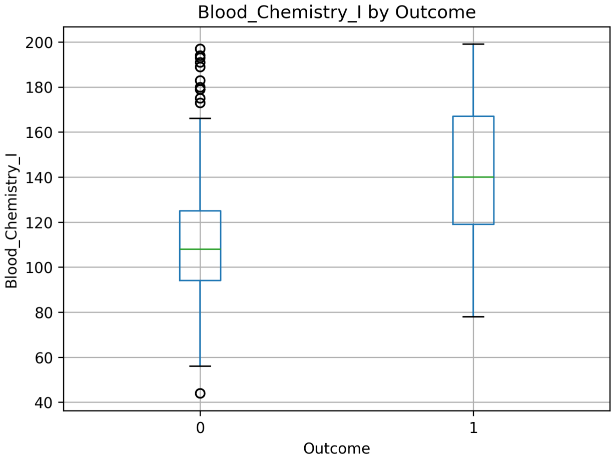

Figure2

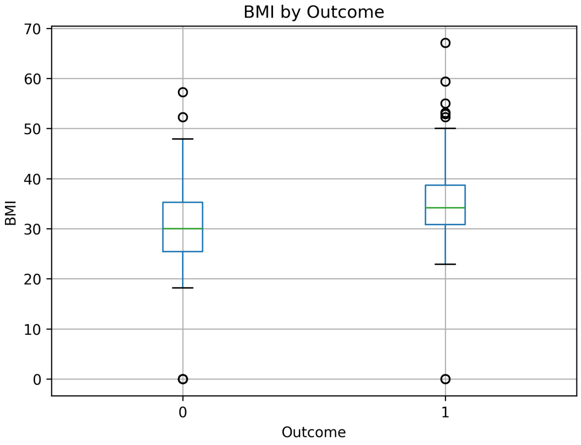

Figure 3

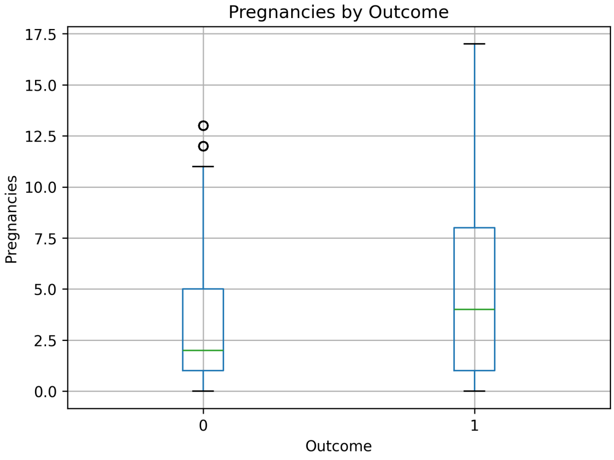

Figure 4

Figure 1, This bar chart shows that the dataset is imbalanced with more patients in the “no disease” group (0) than in the “disease” group (1). **Figure 2, p**atients with disease (Outcome = 1) tend to have higher Blood_Chemistry_I values than those without disease. This feature appears to be a strong predictor of disease incidence. **Figure 3,** BMI values are generally higher in patients with disease compared to those without. This suggests BMI is an important risk factor for disease in this dataset. **Figure 4,** the number of pregnancies is higher, on average, in the disease group (1), indicating that pregnancies may also be associated with increased disease risk.

Modelling

A supervised learning approach using a Random Forest classifier was employed. The model used only the most predictive clinical features and was evaluated on a holdout test set using accuracy, recall and ROC-AUC.

The model achieved high performance with an accuracy of 99.9%, recall of 100% and ROC-AUC of 0.999.

For best practice, the workflow was checked for data leakage to ensure only clinical features were included.

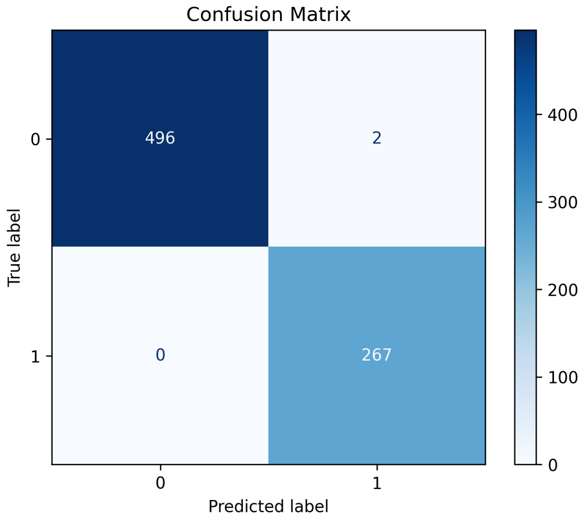

Figure 5, The confusion matrix shows the performance of the Random Forest model on the test data. The model correctly identified 496 patients without disease (true negatives) and 267 patients with disease (true positives). Only 2 patients without disease were incorrectly predicted as having the disease (false positives), and no patients with disease were missed (false negatives = 0). This indicates extremely high accuracy and recall, with the model making almost no errors in classifying patients.

Recommendation

Based on the analysis, Blood_Chemistry_I is the most predictive marker for disease risk in your patients. However, BMI and Pregnancies also contribute to disease prediction. These factors can be used in early screening and further testing to support early intervention and improve patient care.