Diabetes Data Analysis

Banan Bashar Aljarrah

2025-07-5

Goal:

The aim of this analysis is to explore how key demographic and health-related variables — specifically Body Mass Index (BMI) and Age — differ between individuals with and without diabetes. By comparing these average values across diabetes status groups, this exploratory analysis seeks to identify potential patterns or associations that may support early detection efforts or targeted interventions.

This analysis explores medical data from 768 female patients to identify key factors associated with diabetes.

Data Import and Cleaning

In medical datasets, values such as 0 for Glucose, Blood Pressure, or BMI are physiologically implausible and usually represent missing data. Replacing these zeros with NA ensures that these fields are correctly treated as incomplete, which improves the quality of subsequent analyses and prevents misleading averages.

Missing Value Imputation using KNN

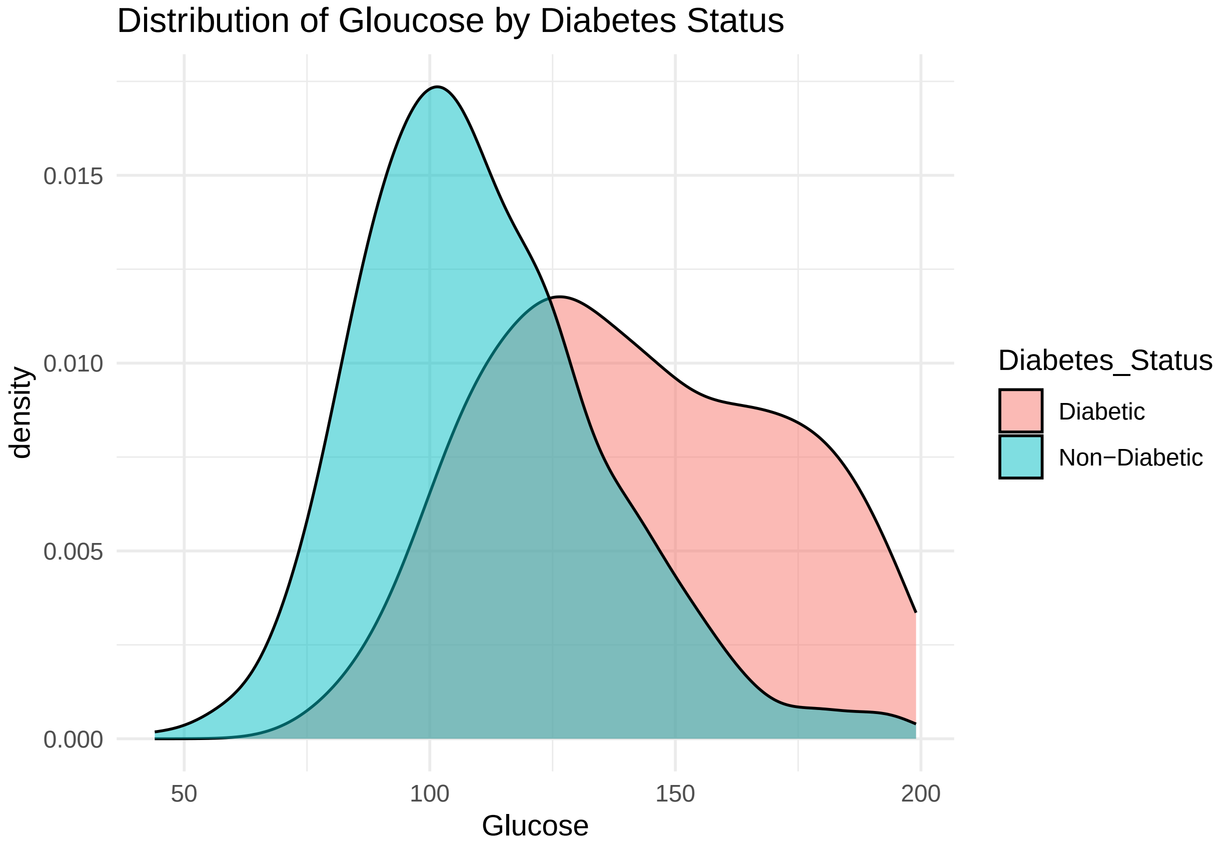
KNN (k-nearest neighbors) imputation replaces missing values using information from the most similar patients. This method preserves dataset size and leverages correlation structures in the data, making it especially suitable for clinical variables that vary in related ways (e.g., BMI and insulin)

Descriptive Statistics

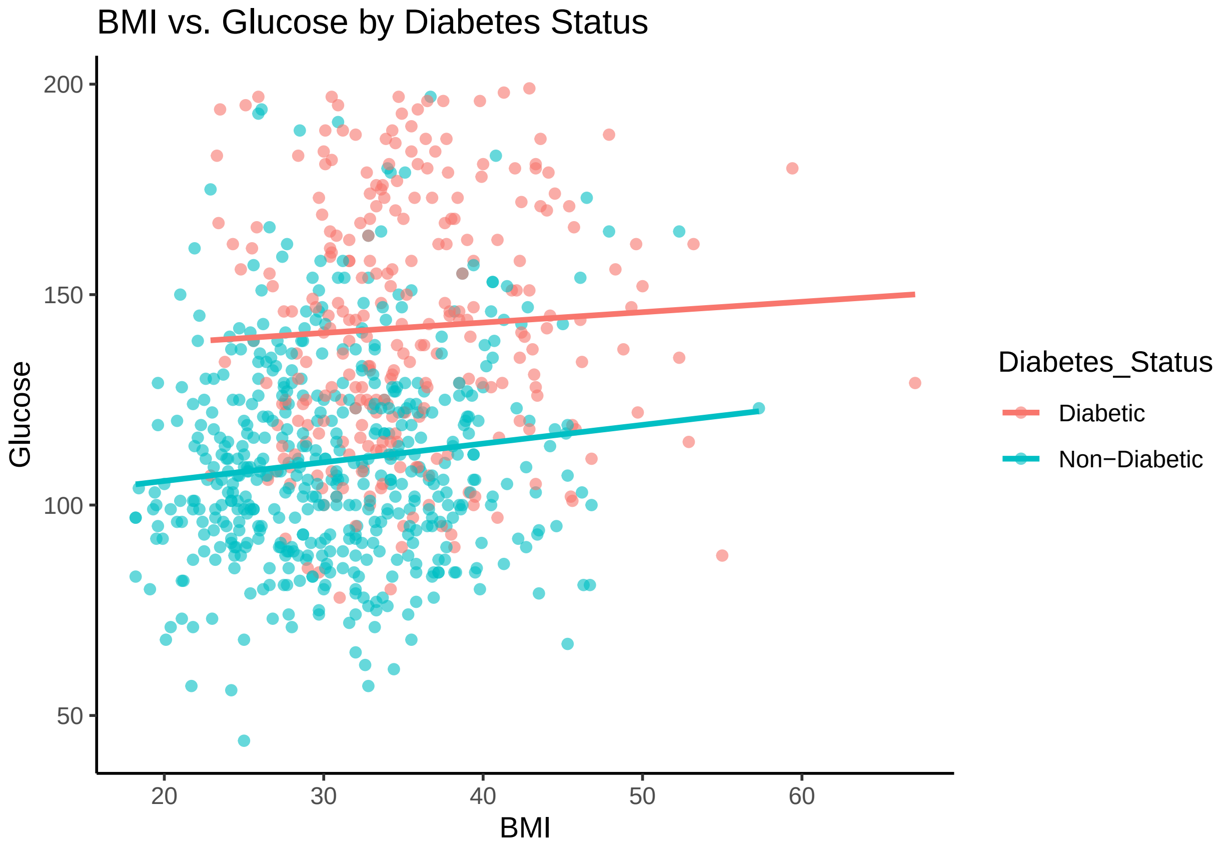
```
## # A tibble: 2 x 7
##   Diabetes_Status count mean_age mean_glucose mean_bmi mean_pressure
##   <fct>           <int>    <dbl>        <dbl>    <dbl>         <dbl>
## 1 Diabetic          268     37.1         142.     35.4          75.3
## 2 Non-Diabetic      500     31.2         111.     30.9          71
## # i 1 more variable: mean_insulin <dbl>
```

The table summarizes key metrics like age, glucose, and BMI by group allows for preliminary identification of patterns or group level differences, which can guide further analysis and hypothesis generation.

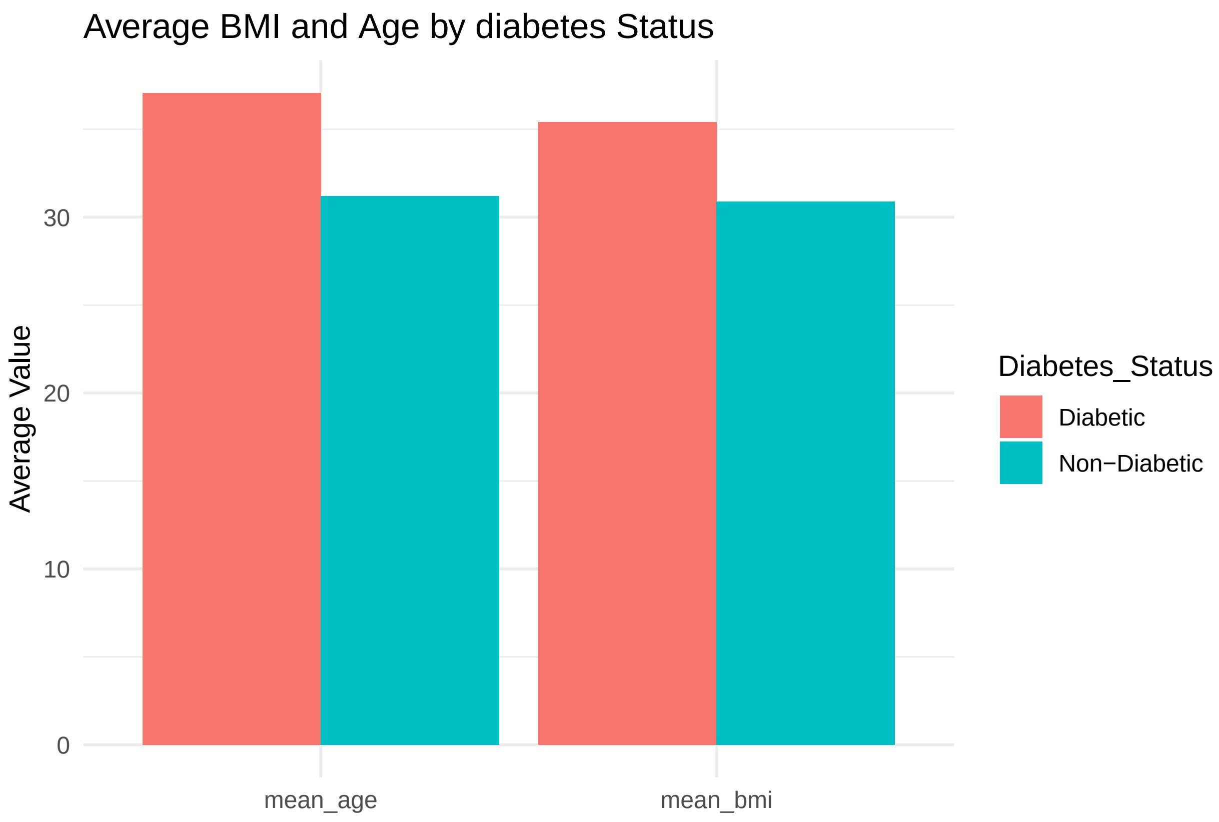
Data Visualization



This plot highlights how glucose levels differ between diabetic and non-diabetic individuals. A clear shift toward higher glucose among diabetic patients confirms its diagnostic relevance and validates the dataset.



Bivariate Relationship: BMI vs. Glucose Exploring the relationship between BMI and glucose helps assess whether body weight is associated with glycemic levels. Stratifying by diabetes status reveals differing trends between the groups.



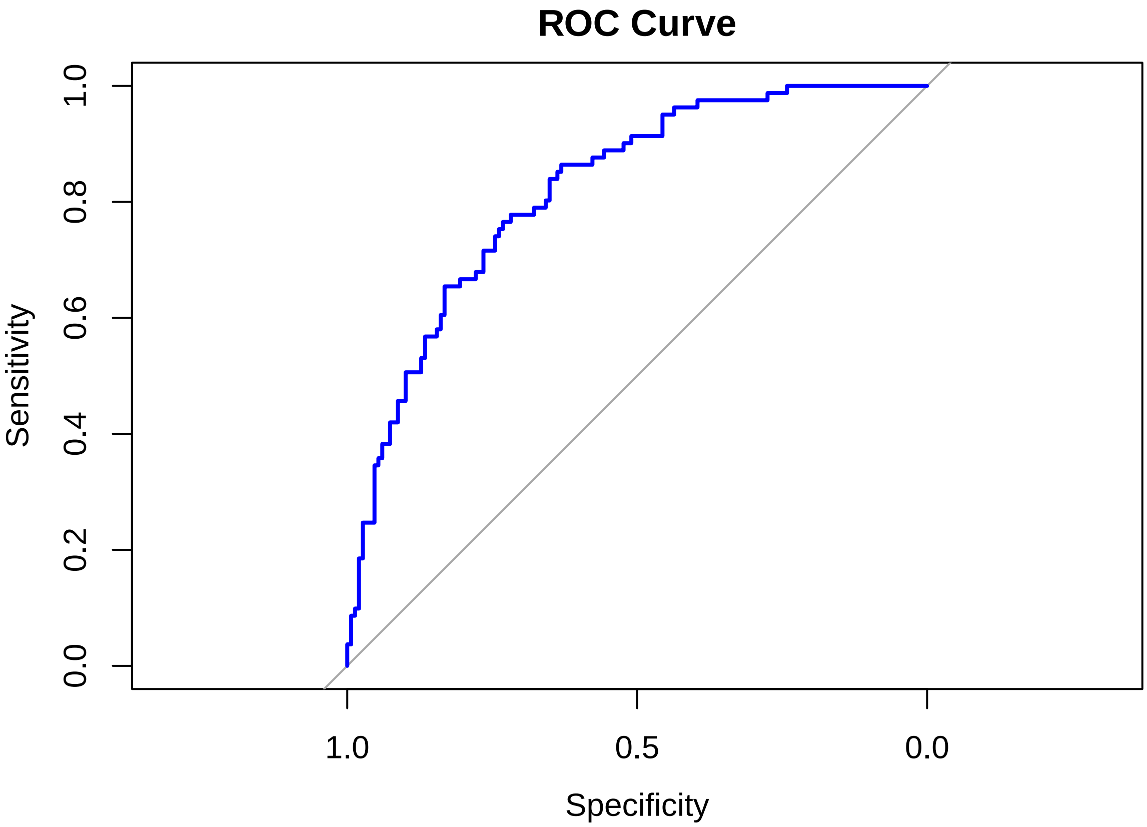
Group-wise Comparison of Key Predictors Bar plots comparing average BMI and age across diabetes status groups highlight key demographic and health differences. These insights can support targeted intervention strategies.

Predictive Modeling

```
##
## Call:
## glm(formula = Outcome ~ Age + BMI + Glucose + BloodPressure +
##     Insulin, family = "binomial", data = train_data)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -9.008550   0.957237  -9.411  < 2e-16 ***
## Age            0.030065   0.009747   3.085  0.00204 **
## BMI            0.095324   0.018839   5.060 4.19e-07 ***
## Glucose        0.034833   0.005038   6.915 4.69e-12 ***
## BloodPressure -0.005416   0.010028  -0.540  0.58914
## Insulin        0.001312   0.001518   0.864  0.38749
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 695.03  on 537  degrees of freedom
## Residual deviance: 505.83  on 532  degrees of freedom
## AIC: 517.83
##
## Number of Fisher Scoring iterations: 5
```

A multivariable logistic regression estimates the odds of diabetes as a function of multiple clinical predictors. This model quantifies the strength of association for each factor and provides a basis for individual-level risk estimation.

ROC Curve and AUC Evaluation



```
## Area under the curve: 0.8203
```

The ROC curve visualizes the tradeoff between sensitivity and specificity across prediction thresholds. The AUC (Area Under Curve) quantifies model performance: values closer to 1 indicate better discrimination between diabetic and non-diabetic cases.

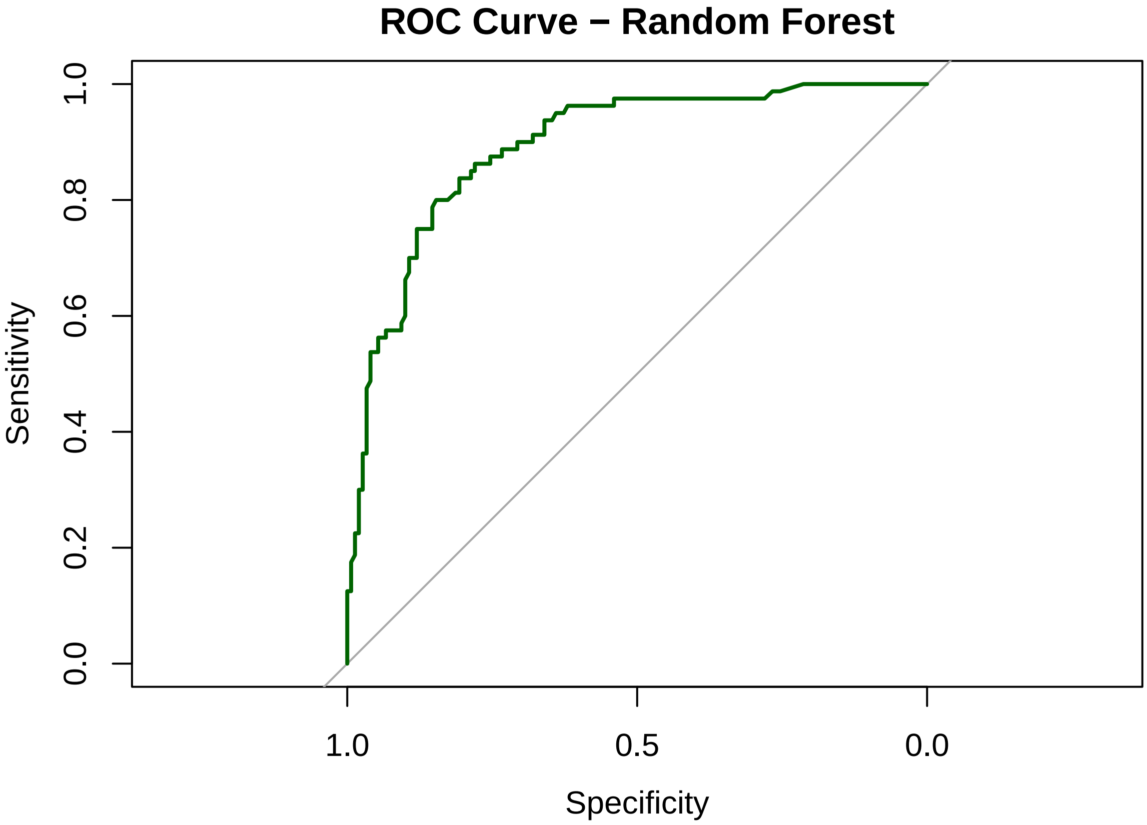
Summary Table of Key Metrics

Table 1: Average Values by Diabetes Status

Diabetes_Status	mean_age	mean_glucose	mean_bmi
Diabetic	37.06716	142.2276	35.39216
Non-Diabetic	31.19000	110.5580	30.88480

Random Forest Classifier

Random Forest is a non-linear ensemble model that captures complex interactions between variables without assuming linearity. It can enhance predictive performance, especially in datasets with intricate relationships like those in clinical settings. The ROC curve and AUC demonstrate its ability to discriminate diabetic cases.



```
## Area under the curve: 0.8929
```

To assess the tradeoff between interpretability and performance, both logistic regression and random forest models were developed. While logistic regression provides insight into the individual effect of each variable, the random forest model captured non-linear relationships and improved predictive accuracy.

Table 2: Comparison Between Logistic Regression and Random Forest

Model	AUC_Score	Interpretability	Notes
Logistic Regression	0.8203	High	Simple and explainable
Random Forest	0.8929	Medium	Better predictive performance

Conclusion

This analysis provided valuable insights into the clinical characteristics associated with diabetes among female patients. Key variables such as glucose level, BMI, and age showed clear differences between diabetic and non-diabetic individuals. Missing data was appropriately handled using KNN imputation to ensure data quality and preserve sample size.

Two predictive models were developed: logistic regression and random forest. While the logistic model offered high interpretability and identified significant predictors of diabetes, the random forest model demonstrated superior predictive performance by capturing non-linear relationships in the data.

Overall, the findings support the use of BMI and glucose as early indicators of diabetes risk.