**Data Science Report: Clustering, Regression and Diabetes Prediction**

**Author:** Meghana Sulthan

**Student Number:** 23070310  
**GitHub Repository:** [Your GitHub Repository Link]

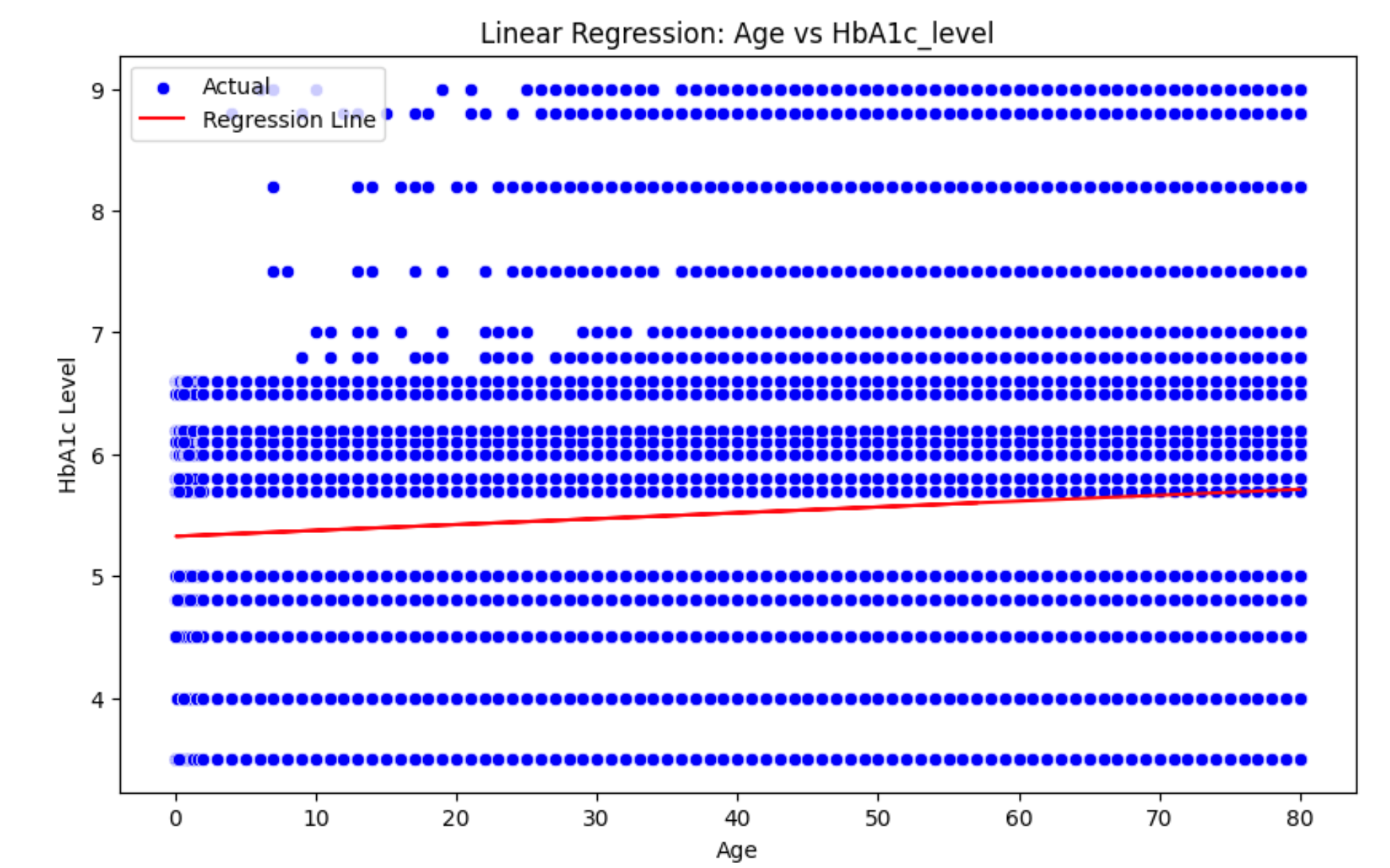
# Introduction

In this analysis, It will be working with a single dataset that contains health parameters: age, body mass index, HbA1c level, and another for diabetes status. The methods involve linear regression to be applied for line fitting with an appropriate logistic regression predictor model for the status development of diabetes, supported wherever needed with visualizations.

# Plot 1 Linear Regression Line Fitting (Age vs HbA1c Level )

**Discussion:**

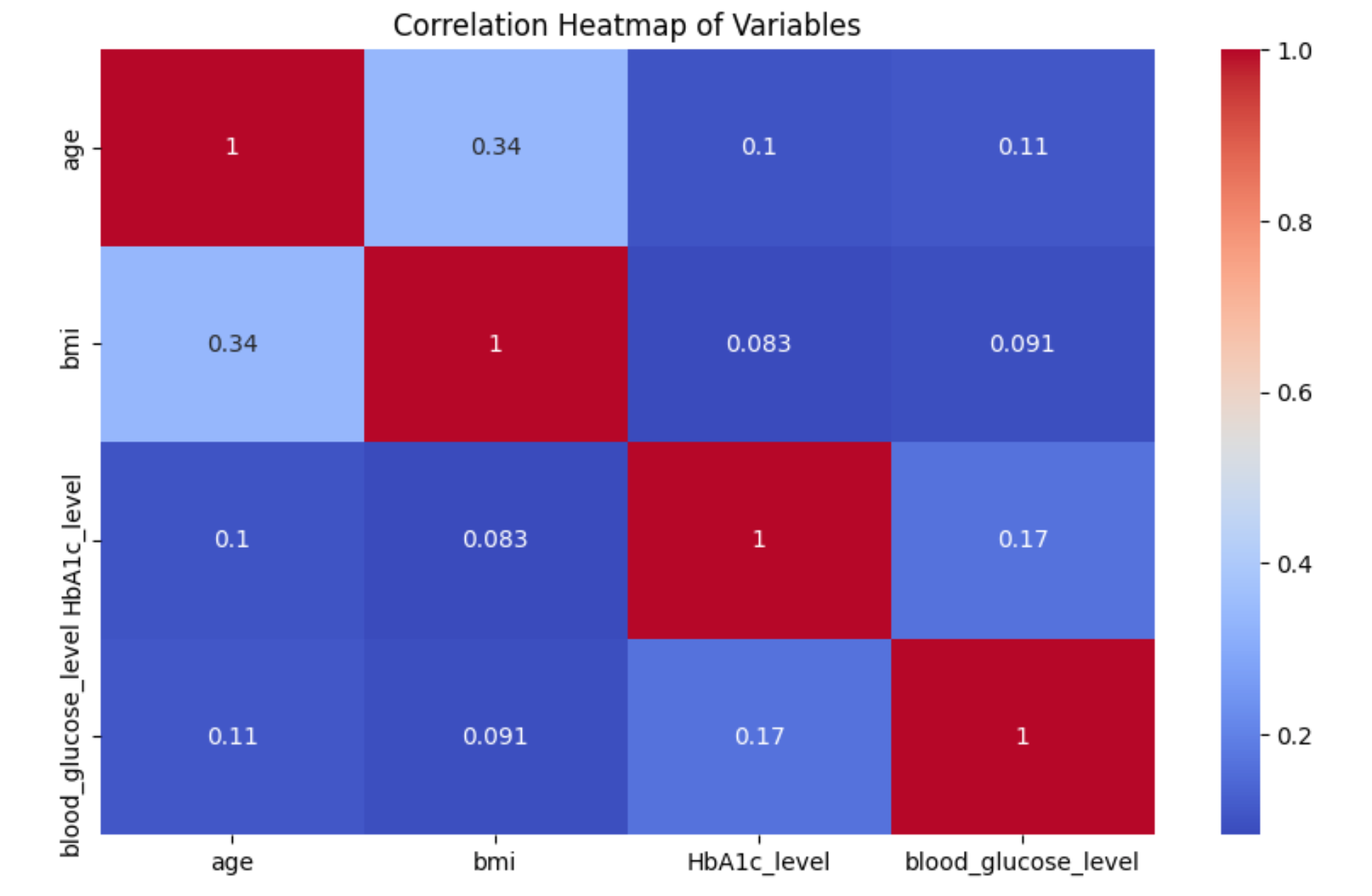
The scatter plot shows a positive linear relationship between age and HbA1c\_level; that is, older individuals generally have higher levels of HbA1c. Data may come close to the regression line, so age could presumably be a predictive factor regarding levels in the condition of HbA1c. This can provide an insight into the formulation of preventive measures in health care systems, where early detection of high HbA1c values is beneficial.



# Plot 2: Correlation Heatmap

**Discussion:**

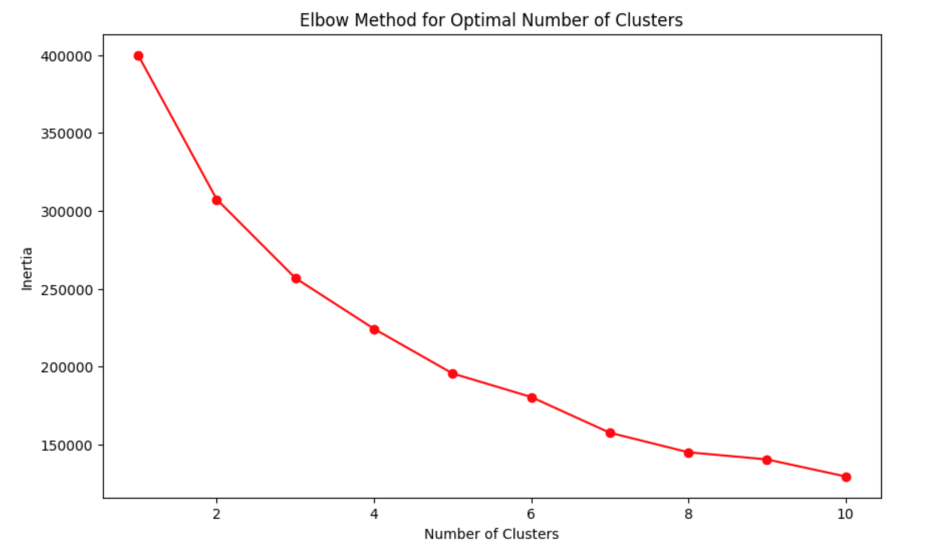
Heatmap: There is an extremely strong positive correlation between blood\_glucose\_level and HbA1c\_level, which was expected from elementary knowledge of the subject since both of these are a couple of key indicators related to diabetes management. The correlation between age and bmi is moderate, and there is also a significant correlation between bmi and HbA1c\_level that might indicate that the level of BMI can influence blood sugar levels and, hence, HbA1c. This information can help healthcare professionals identify risk factors for diabetes and recommend lifestyle changes.



# Plot3: Elbow plot for Optimum number of cluster

**Discussion:**

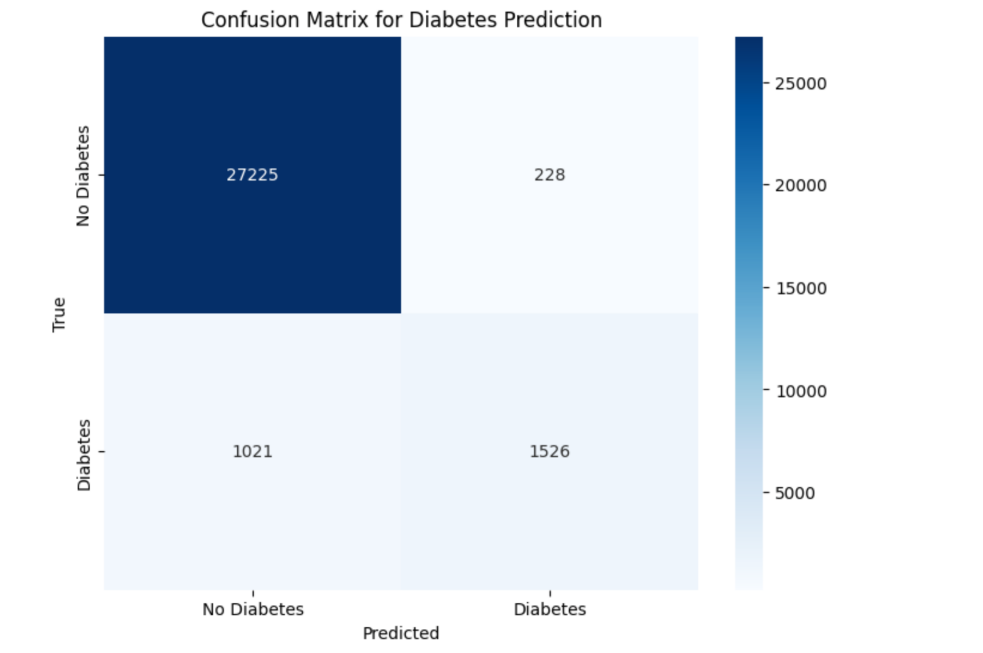
The elbow plot shows that 3 is the best number of clusters, as after this, the inertia starts to decrease at a lower rate. That means three is the best number of clusters for proper grouping of data in this analysis. This method of determining the number of clusters helps to ensure that the K-means algorithm finds the best representation of the data without overfitting.



# Plot 5: Confusion Matrix for Diabetes Prediction

**Discussion:**

The confusion matrix is used for the performance evaluation of the developed diabetes prediction model, which shows the true positive count, false positive count, true negative count, and false negative count. The high number of true positives and negatives reflects that the status of Diabetes was well predicted by this model. Further work based on that might be including more features in order to add strengths to this model or use different other algorithms for classification with better output. The confusion matrix is an essential tool in the evaluation of classification models for healthcare applications.



# Conclusion

This report was on the application of clustering, linear regression, and classification on a health dataset. The linear regression model allowed for an insight into how age is related to HbA1c levels. The correlation heatmap showed some important relationships between variables, and the elbow plot confirmed the optimal number of clusters. Finally, the confusion matrix presents the accuracy of a logistic regression model designed to predict diabetes status. These analyses henceforth can be used in healthcare decisions and improve early detection and management of diabetes. Further studies will be done by enhancing the feature in the models or exploring alternative machine learning techniques.