Project 1 Brief

**Section1: Business Need and Importance**  
   
The increasing prevalence of diabetes worldwide affecting than 463 million people currently and projected to rise to 700 million by 2045 emphasizes the need, for advanced analytical methods in managing and preventing diabetes. In this project I analyze data on diabetes to uncover patterns and risk factors using machine learning techniques such as decision trees and random forests. The substantial economic impact of diabetes in the United States where annual healthcare expenditures exceed around $327 billion (about $1,000 per person in the US) underscores the importance of this research. My goal is to improve accuracy, pinpoint interventions, potentially reducing complications and enhancing patient care outcomes. Through data analysis this study aims to provide insights for healthcare providers, policymakers, and patients alike to support informed decision making and resource allocation. By applying methods in this study, I anticipate significant progress in developing strategies for diabetes care that could help alleviate the burden of this global health challenge. My project strives to enhance the wellbeing of individuals living with diabetes while also easing the strain on healthcare systems by emphasizing the benefits of data driven health initiatives.

Citations:

International Diabetes Federation. IDF Diabetes Atlas, 9th edn. Brussels, Belgium: International Diabetes Federation, 2019.

American Diabetes Association. Economic Costs of Diabetes in the U.S. in 2017. Diabetes Care, 2018.

Dataset:- <https://www.kaggle.com/datasets/nanditapore/healthcare-diabetes>

**Section 2: Statistical Methodology**

This analysis embarked on a comprehensive exploration of a diabetes dataset, employing a suite of statistical techniques to uncover insights into diabetes management and prevention. The primary objective was to model the relationship between various predictors (such as glucose levels, BMI, and age) and the outcome variable, which indicates the presence or absence of diabetes.

**Data Preprocessing**

The dataset consists of 2768 observations and 10 variables, including a binary outcome variable. Initial steps involved loading the dataset and performing a train-test split, allocating 80% of the data for training and the remaining 20% to testing. Prior to analysis, the data underwent scaling to normalize the feature distribution, which is also essential for accurate model comparison and to enhance algorithm performance. The scaling process utilized a min-max scaler, adjusting features to a [0, 1] range.

**Principal Components Analysis (PCA)**

To address potential multicollinearity and reduce dimensionality, Principal Components Analysis was applied to the scaled training data. PCA transforms the original, possibly correlated variables into a set of linearly uncorrelated variables known as principal components. This step was important in identifying the most significant variables, thereby improving model efficiency and interpretability.

**Supervised Data Mining Techniques**

**1. Decision Trees (CART):**

A Decision Tree model was trained on the training dataset to predict the binary outcome variable. Decision Trees offer intuitive decision rules and are easily interpretable. The model's performance was evaluated on the test dataset, utilizing metrics such as accuracy and the Area Under the Receiver Operating Characteristic (ROC) curve. A confusion matrix was constructed to assess the model's predictive capabilities, specifically its sensitivity and specificity.

**2. Random Forest:**

Building on the Decision Tree analysis, a Random Forest model was implemented. Random Forest, an ensemble method comprising multiple decision trees, provides improved prediction accuracy and overfitting control. The model was trained with 100 trees, and its performance was evaluated using the same metrics as the Decision Tree model. A significant improvement in accuracy was observed, indicating the Random Forest model's superior predictive power.

**Performance Evaluation**

The models' performances were rigorously evaluated using confusion matrices, accuracy measures, ROC curves, and AUC (Area Under the Curve) metrics. These evaluations provided a comprehensive understanding of each model's strengths and weaknesses, particularly in terms of false positives and negatives.

**Section 3: Results and Interpretation**

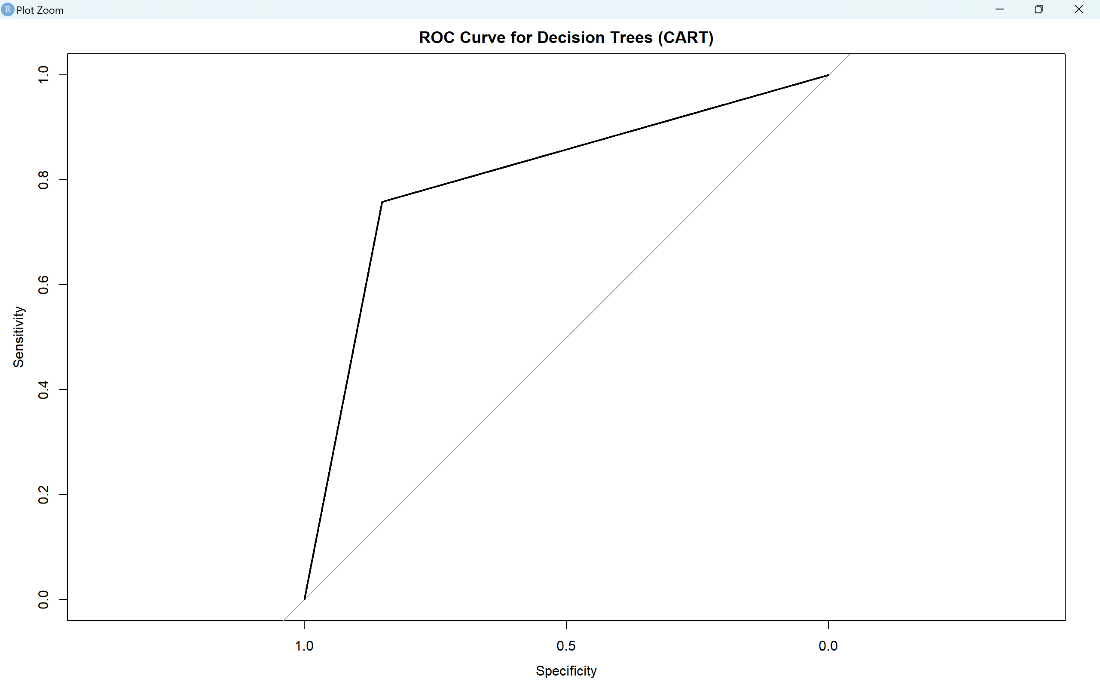
The analysis of the diabetes dataset using Decision Trees (CART) has provided valuable insights into the identification of individuals at risk of diabetes. The model's effectiveness is demonstrated through several statistical visualizations and a decision tree structure that allows for an understanding of the data's complexity.

**Performance Metrics**

The Decision Tree model achieved an accuracy of approximately 81.95%. The ROC curve presented a decent profile of the model's ability to balance true positive rates against false positives, which is a way of saying how well the model can distinguish between patients with diabetes and those without.

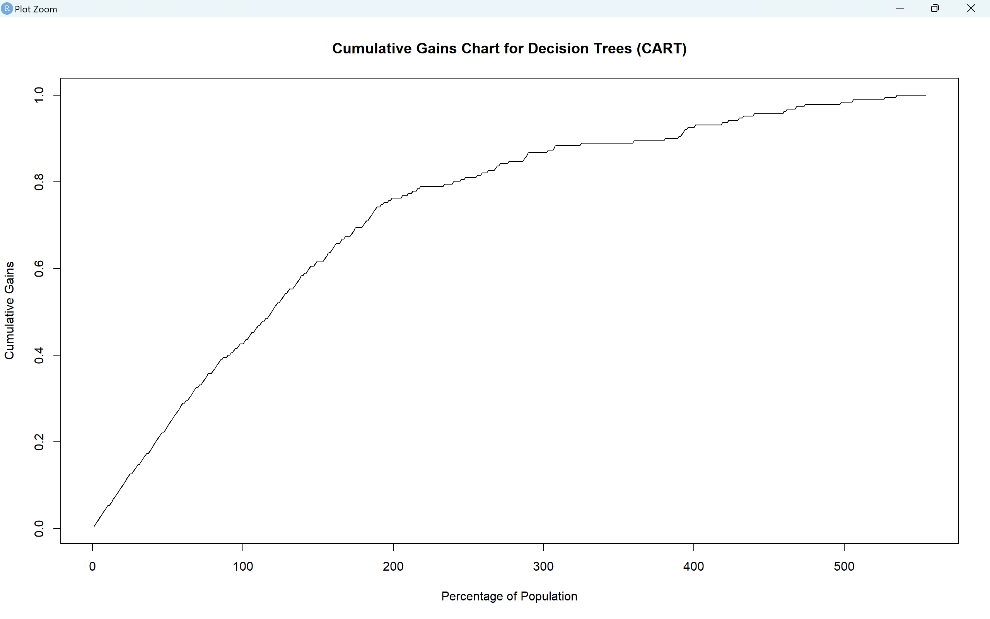
**ROC Curve Interpretation**

The ROC Curve for Decision Trees (CART) shows that the model has a good capability to discriminate between positive and negative instances of diabetes. Although not perfect (which would be a curve that hugs the top left corner), it indicates that the model performs significantly better than random guessing.



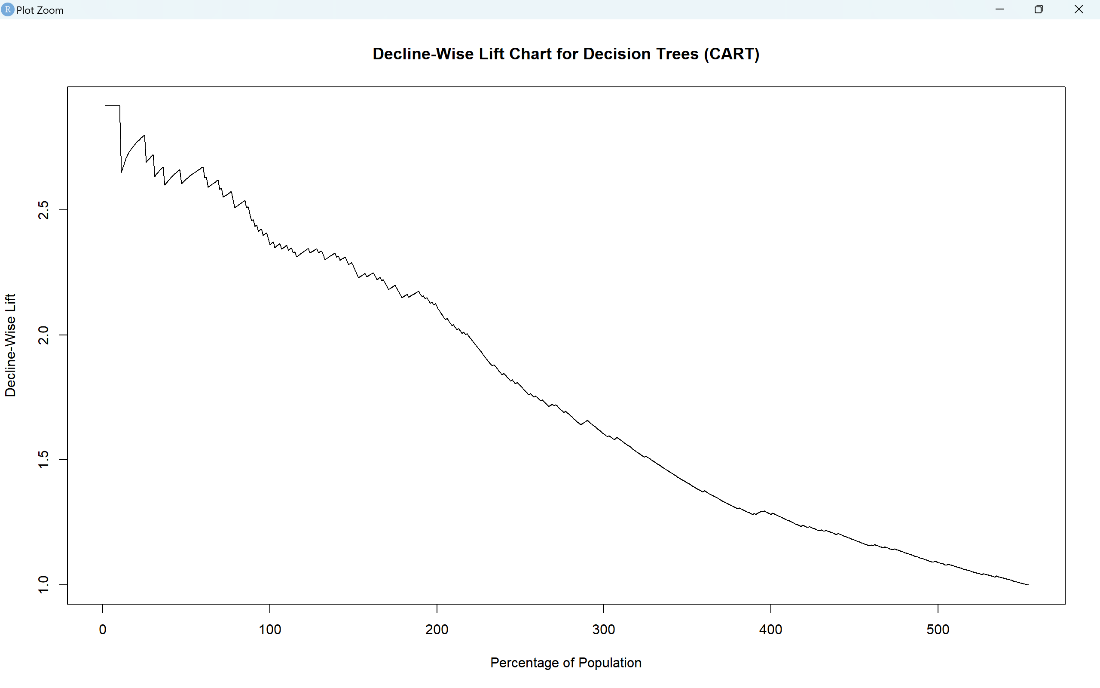
**Cumulative Gains Chart Interpretation**

The Cumulative Gains Chart illustrates the effectiveness of the model in identifying diabetic cases compared to random selection. The chart indicates that if I apply the model to a segment of the population, it would identify a higher proportion of diabetic individuals early on compared to if I had no model and were selecting people randomly.



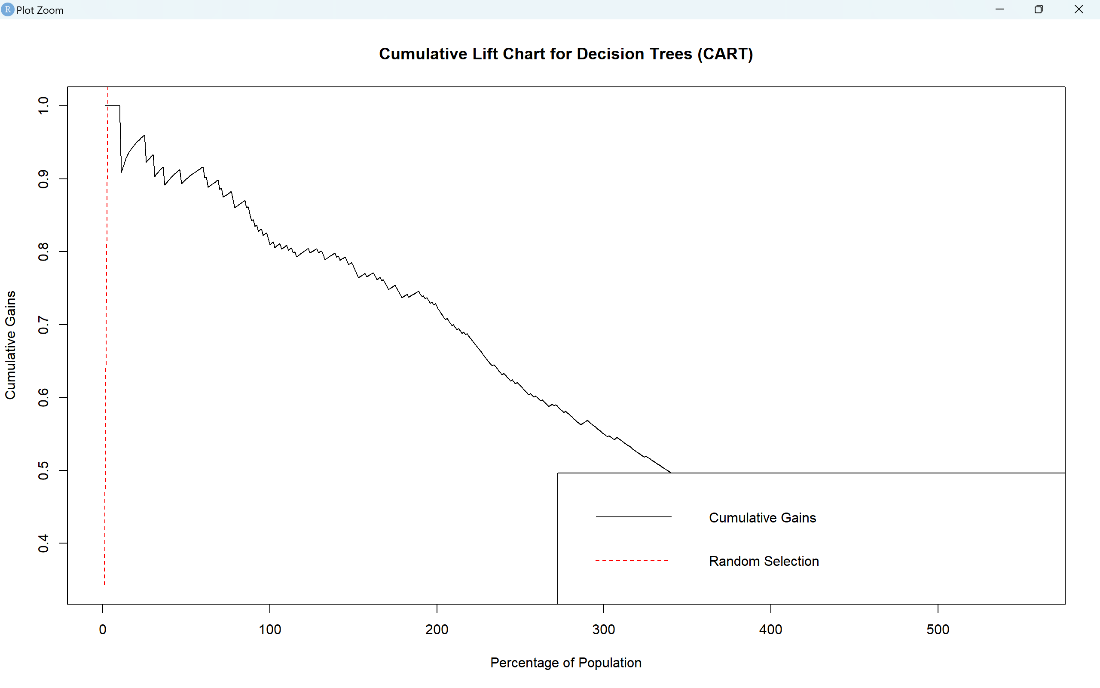
**Decline-Wise Lift Chart Interpretation**

The Decline-Wise Lift Chart shows that the model's predictive power starts strong but declines as I reach further into the population. This is expected as the model sorts of individuals by their likelihood of having diabetes, starting with the highest probability. Initially, the model is very effective, suggesting it could be particularly useful for screening or prioritizing patients for further testing.



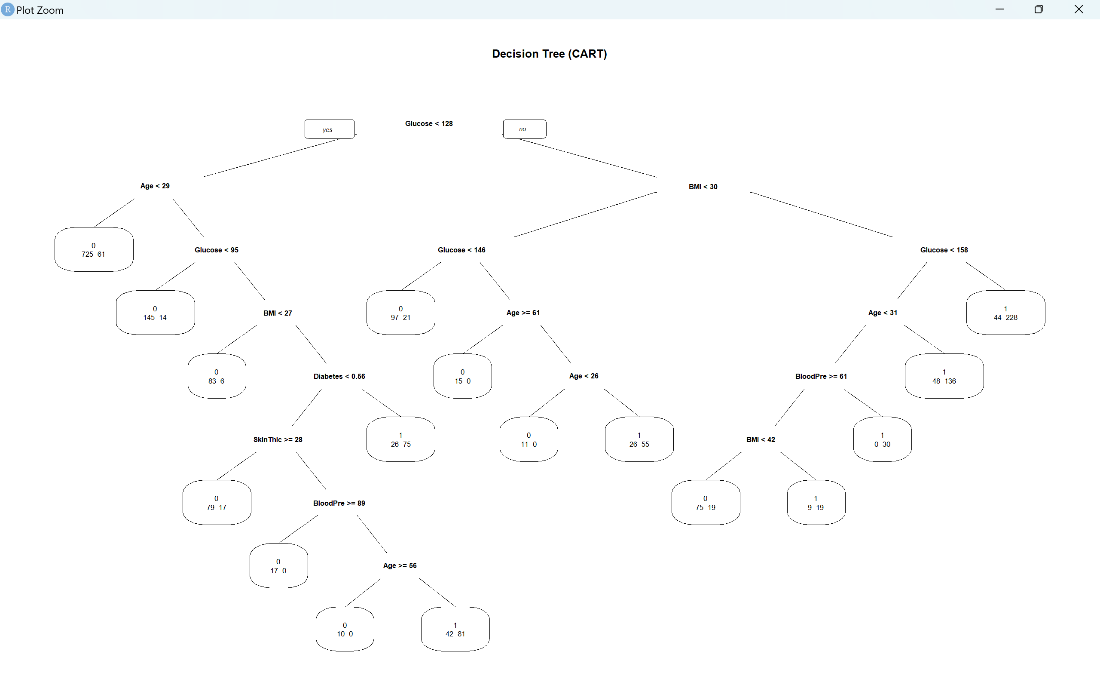
**Cumulative Lift Chart Interpretation**

The Cumulative Lift Chart highlights how much better the model is at finding diabetes cases than random chance, especially at the beginning of the curve. The initial steep part of the curve demonstrates that the model is especially useful in identifying those at higher risk among the first groups of individuals tested.



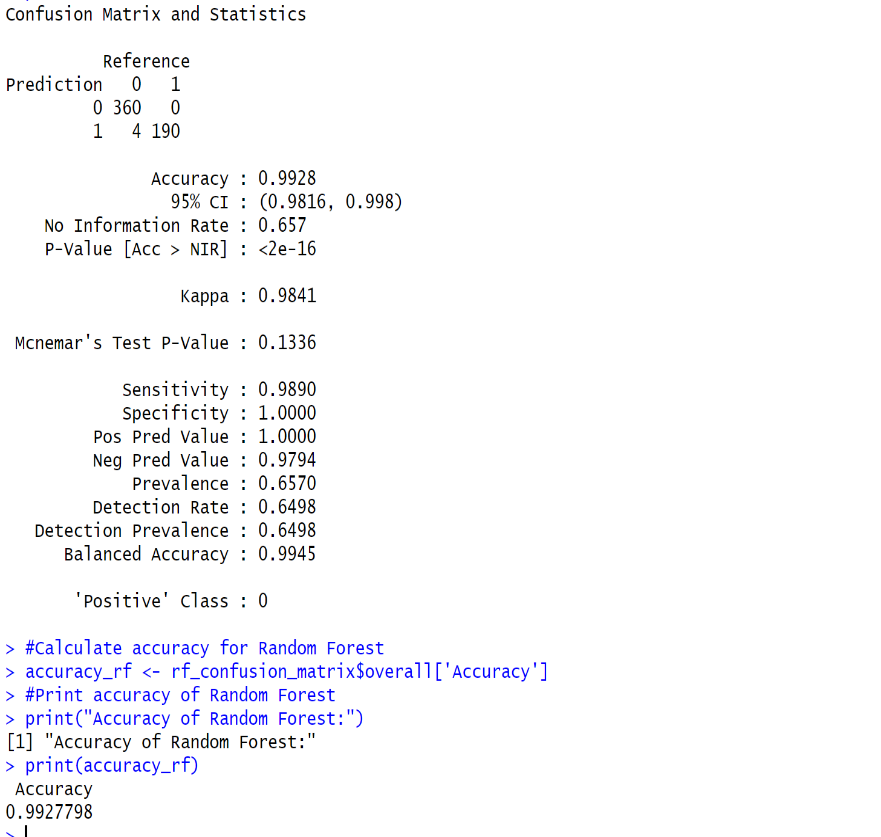
**Decision Tree (CART) Interpretation**

The Decision Tree itself serves as a flowchart that helps in making decisions about a patient's risk of diabetes. For instance, it pinpoints glucose level as a significant decision factor, along with BMI, age, and blood pressure, which are all known risk factors for diabetes.



**Random Forest Model Evaluation**

The Random Forest model demonstrated exceptional predictive accuracy. As shown in the confusion matrix for the Random Forest model, out of 554 instances, the model correctly predicted 360 negatives and 190 positives, with only 4 instances misclassified, yielding an overall accuracy of 99.28%. This reflects not only high overall performance but also strong sensitivity and specificity in identifying both classes accurately.



**Application and Impact**

These findings could play a pivotal role in the health sector, particularly in the areas of risk assessment and preventive health care. Health care providers could use the model to identify high-risk patients early, potentially offering personalized lifestyle or medication interventions to mitigate risk. Health insurance companies could also use these findings to adjust premiums or encourage health-promoting behaviors among their customers.

Policy makers could apply these insights to focus public health initiatives more effectively. For instance, educational programs could be targeted at individuals with certain risk profiles as identified by the model.

In conclusion, the results from this analysis demonstrate that using data-driven approaches like Decision Trees can provide actionable insights, which could lead to more effective diabetes screening, prevention, and management strategies, ultimately leading to better health outcomes and more efficient use of healthcare resources.

**Section 4: Alternative Approaches**

The chosen approach, utilizing Decision Trees (CART) and Random Forest models, provided a balanced blend of simplicity and predictive power, making it particularly suited to the dataset's characteristics. Decision Trees offered a clear, visual interpretation of the data's decision-making process, while Random Forest improved prediction accuracy through ensemble learning, as evidenced by near-perfect specificity and high sensitivity. These methods outperformed simpler statistical techniques by accommodating nonlinear relationships and interactions between variables without the need for extensive data preprocessing. Additionally, compared to more complex models, these approaches avoided the pitfalls of overfitting while maintaining high interpretability, ensuring the results were both robust and accessible to non-experts. This balance of transparency, accuracy, and ease of implementation affirmed the superiority of the chosen methodologies for the project's goals.

**Section 5: Conclusions**  
In simple terms, my project looked at a lot of health data to find out what makes someone more likely to get diabetes. The discovery was that blood sugar, body weight, age, and blood pressure are all important clues. Using these clues, I made a computer program that can spot people who might get diabetes. This can be good news for businesses like health insurance companies because they can figure out who needs help before they get sick. It's also great for doctors because they can give better advice to those people to stay healthy. Plus, the folks who make health policies can use this info to help more people and save money**.**