Main Report

# PREDICTING DIABETES

*Enhancing Recall with Decision Trees and Random Forests*

**BA305**

**1. Introduction:**

**1.1 Problem Statement**

Diabetes is a chronic condition that, if left undiagnosed, can lead to severe complications, including cardiovascular disease, kidney failure, and neuropathy. Early detection is crucial for timely intervention, reducing the risk of these complications. This project focuses on developing predictive models to identify individuals at risk of diabetes using health-related indicators. The goal is not only to achieve accurate predictions but also to minimize false negatives, which could result in undiagnosed cases, and false positives, which may lead to unnecessary stress and medical interventions.

**1.2 Dataset Overview**

Our dataset selection process was iterative. Initially, we considered analyzing housing prices using a dataset with features like the number of bedrooms and air conditioning. However, its outdated nature and inconsistent role of ZIP codes in valuation led us to discard it. We then explored predicting athlete injury risks using another dataset, but its poor quality and insufficient variables made it unsuitable.

Ultimately, we selected the [Diabetes Health Indicators Dataset](https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset), which consolidates health-related features, such as BMI, physical activity, blood pressure, smoking history, and more. It includes 253,680 entries across 22 columns, making it an ideal choice for modeling diabetes risk. Preliminary analysis revealed meaningful health patterns, such as the strong association of high BMI and low physical activity with diabetes, confirming the dataset's relevance to our objectives.

**1.3 Significance of Precision and Recall**

In healthcare applications, the stakes of incorrect predictions are high:

* False Negatives (diabetic cases predicted as non-diabetic) can delay critical interventions, exacerbating health risks.
* False Positives (non-diabetic cases predicted as diabetic) can lead to unnecessary stress and medical expenses.

Given this context, accuracy alone is insufficient. A model might achieve high accuracy simply by predicting the majority class (non-diabetic), but this would fail to address the minority class effectively. To address these challenges, this project prioritizes precision (reliability of positive predictions) and recall (ability to identify true diabetic cases) over accuracy. By focusing on these metrics, the project aims to create a model that balances sensitivity to diabetic cases with reliability.

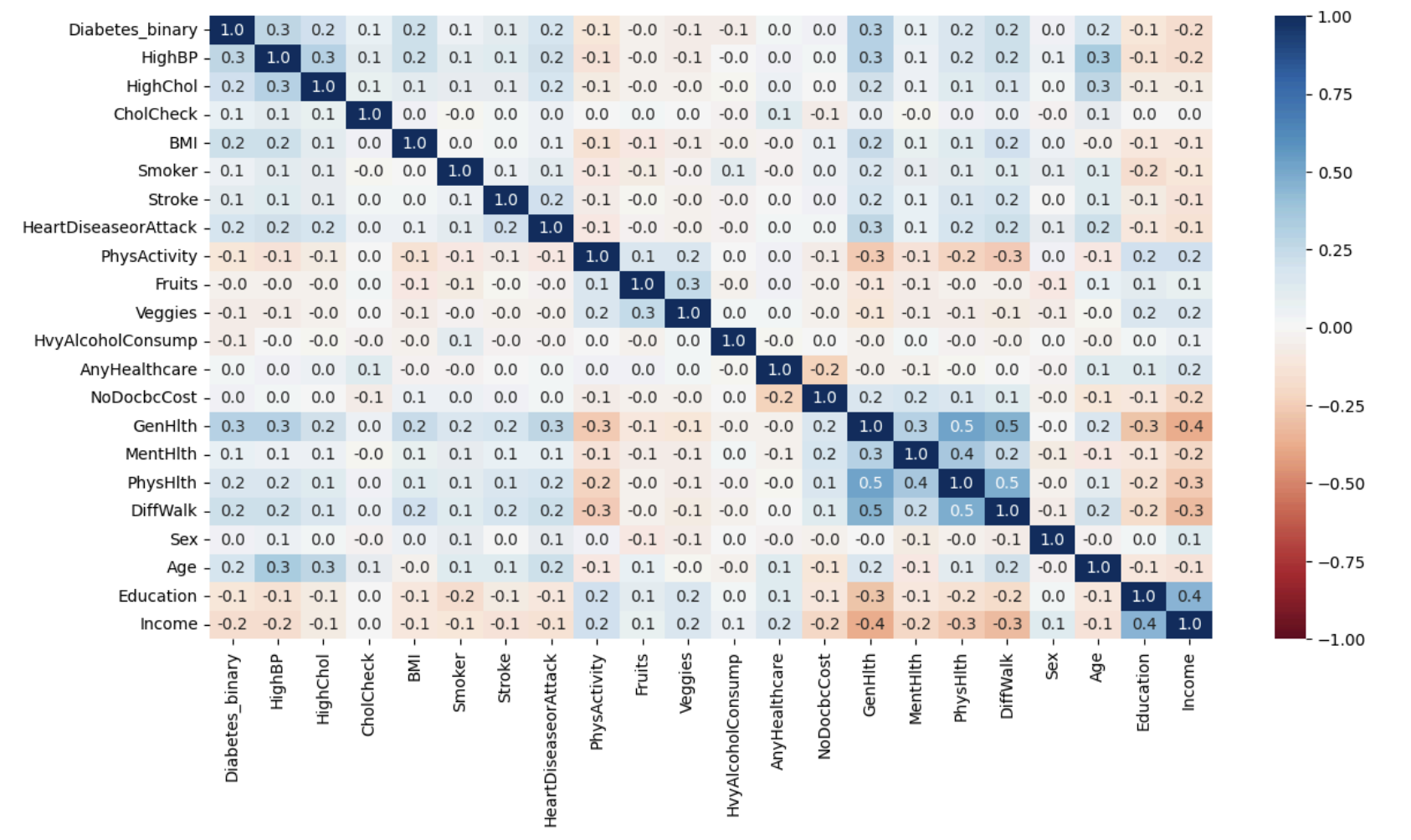
**2. Data Preprocessing:**

**2.1: Introduction: Understanding the Dataset**

To kick off our analysis, we carefully examined the dataset to understand its structure, quality, and key features. The dataset, sourced from health indicators, consists of 253,680 rows and 22 features. These features include critical health metrics like BMI, physical activity levels, smoking status, and blood pressure, all of which are closely linked to diabetes risk in clinical research. We found that the target variable, Diabetes\_binary, is a binary classification indicating whether an individual has diabetes. This target serves as the foundation for our predictive models.

**2.2: Data Quality: No Missing Values**

One of the first steps in our preprocessing workflow was checking for missing or inconsistent data. A thorough scan revealed no missing values across the dataset, ensuring a strong foundation for further analysis. This allowed us to focus on feature selection and model building without needing extensive data imputation.

****

*Figure 1: Correlation HeatMap*

**2.3 Feature Selection: Reducing Noise**

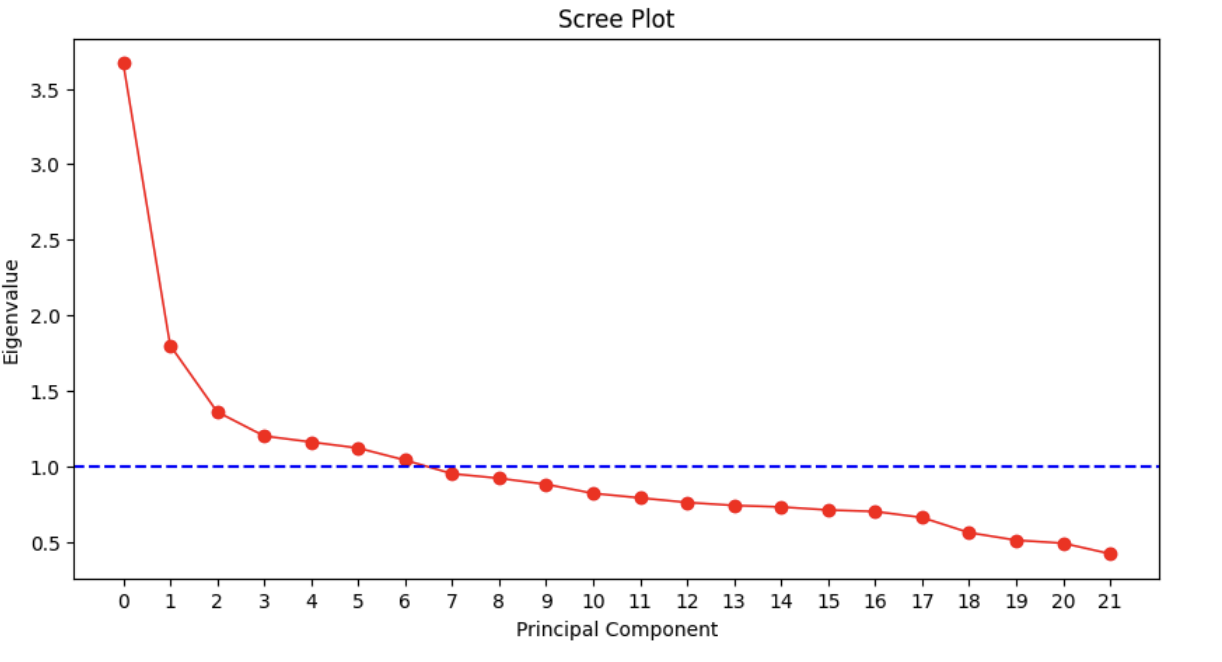
Not all features in the dataset contributed equally to predicting diabetes. To identify which features were most relevant, we performed a correlation analysis between each feature and the target variable. This analysis revealed that six features—Fruit Consumption, Vegetable Consumption, Cholesterol Check, Heavy Alcohol Consumption, Any Healthcare, and Access to Healthcare (NoDocbcCost)—had extremely low or negligible correlations with the target variable.

While these features might have some indirect relevance to overall health, their weak relationship with diabetes prediction added unnecessary noise to the model. To streamline the dataset and focus on impactful predictors, we decided to drop these features.

**2.4 Dimensionality Reduction: Principal Component Analysis (PCA):**

We conducted Principal Component Analysis (PCA) to identify patterns in the dataset and reduce redundancy among features, while ensuring that critical health indicators contributing to diabetes prediction, such as BMI and physical activity, were adequately represented.

To perform Principal Component Analysis (PCA), we first standardized the dataset using z-score normalization. This ensured that all features, regardless of their original scales (e.g., BMI and Income), contributed equally to the analysis. Without scaling, features with larger ranges could have disproportionately influenced the results.

****

*Figure 2: Scree Plot*

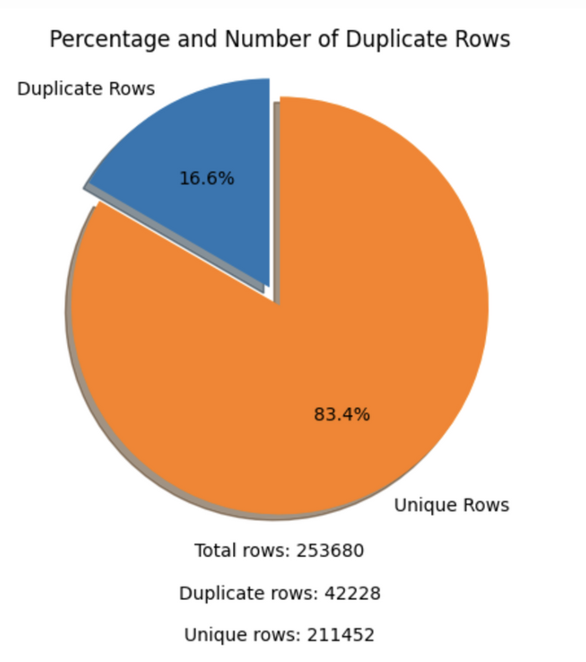
‘

The "elbow point," observed at approximately the 6th component, suggests that using the first six components would retain most of the dataset's variance while reducing dimensionality.

However, despite the PCA recommendation, we chose to retain 16 features based on the insights gained from the correlation matrix. The correlation analysis highlighted the importance of several features that, while contributing less to variance individually, had significant relationships with diabetes prediction. This approach ensured that we preserved the features most relevant to the target variable while still reducing noise by removing features with negligible correlations. This decision struck a balance between dimensionality reduction and maintaining critical predictors for reliable model performance.

**2.5: Removing Duplicates Due To The Nature Of Our DataSet**

We identified and removed duplicate rows in the dataset to ensure data quality and avoid biases in model training. A total of 42,228 duplicate rows were detected and dropped, reducing the dataset to 211,452 unique entries. In our dataset, it is highly unlikely for two observations to have identical values across all variables, given the diversity and context of our features, such as BMI, physical activity, and access to healthcare. OThis suggests that duplicates likely arose from data collection redundancies rather than meaningful repeated entries. Removing these duplicates was essential to eliminate redundant information, prevent skewed statistical analysis, and ensure that the dataset accurately represented unique observations for reliable diabetes prediction.

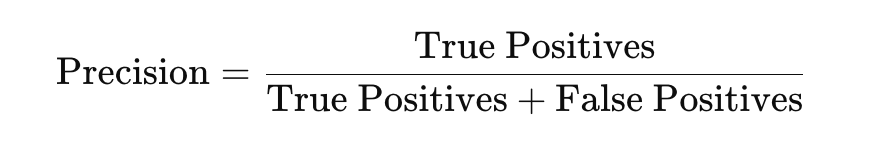
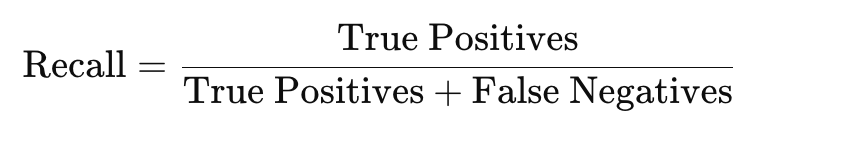


*Figure 3: Percentage and Number Of Duplicate Rows*

After completing the data preprocessing steps, the dataset was reduced to **211,452 rows** and **16 features**, with the final selection of features focusing on those most relevant to diabetes prediction. These steps ensured that the dataset was streamlined, high-quality, and well-suited for building accurate and reliable predictive models.

**3. Evaluation Metrics and Baseline Performance**

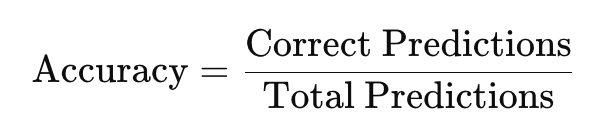
**3.1:Using Precision and Recall Over Accuracy To Evaluate Different Models**



*Figure 4: Formulas For Precision and Recall*

Precision measures the reliability of positive predictions: A higher precision indicates fewer false positives, meaning the model avoids misclassifying non-diabetic cases as diabetic. On the other hand, Recall measures the model’s ability to identify all actual positives:A higher recall ensures that the model minimizes false negatives, i.e., diabetic cases that go undetected.

**3.2 :Why Accuracy is Insufficient**



*Figure 5: Formula For Accuracy*

While accuracy is often a useful metric, it becomes misleading in highly imbalanced datasets like ours, where 83.56% of the observations belong to the majority class (non-diabetic). A model predicting only the majority class would achieve 83.56% accuracy but fail to identify any diabetic cases, offering no practical value in healthcare.

**3.3 Why Precision and Recall Matter for Diabetes Prediction**

Precision and recall share an inverse relationship, meaning improvements in one often come at the expense of the other. This trade-off is particularly relevant in diabetes prediction, where the stakes of false negatives and false positives differ significantly.In healthcare, the cost of false negatives (failing to detect diabetic cases) is far greater than the cost of false positives (incorrectly predicting diabetes in non-diabetic individuals). Here's why:

1. False Negatives:
   * Missing a diabetic case can delay diagnosis and treatment, leading to severe complications such as organ damage, cardiovascular issues, or even life-threatening situations.
   * High recall ensures that the model catches as many diabetic cases as possible, reducing the likelihood of false negatives.
2. False Positives:
   * While false positives may lead to unnecessary testing or anxiety, they are less harmful than missed diagnoses.
   * Precision ensures the model limits false positives, minimizing unnecessary follow-ups while maintaining trust in the predictions.

**3.4 How We Evaluate Models**

For our project, we prioritized recall to minimize false negatives while maintaining acceptable precision. In healthcare, this trade-off is essential to ensure that diabetic cases are correctly identified without overwhelming the system with false positives. Alongside precision and recall, we also considered F1-score, a balanced metric that accounts for both, and ROC-AUC, which evaluates overall model performance. By doing so, we aligned our evaluation criteria with the specific needs of diabetes prediction, emphasizing patient safety and early detection.

**3.5 Naive Rule:**

We began by establishing a baseline using the naive rule, which predicts the majority class—non-diabetic cases—for all observations. This approach achieved an accuracy of 83.6%, reflecting the dataset's class imbalance (83.6% non-diabetic and 16.4% diabetic). However, while this accuracy appears high, the naive rule fails to identify any diabetic cases, resulting in a recall of 0% for the minority class. This limitation underscored the need for models that prioritize recall to minimize false negatives, which is critical in healthcare applications as we discussed earlier.

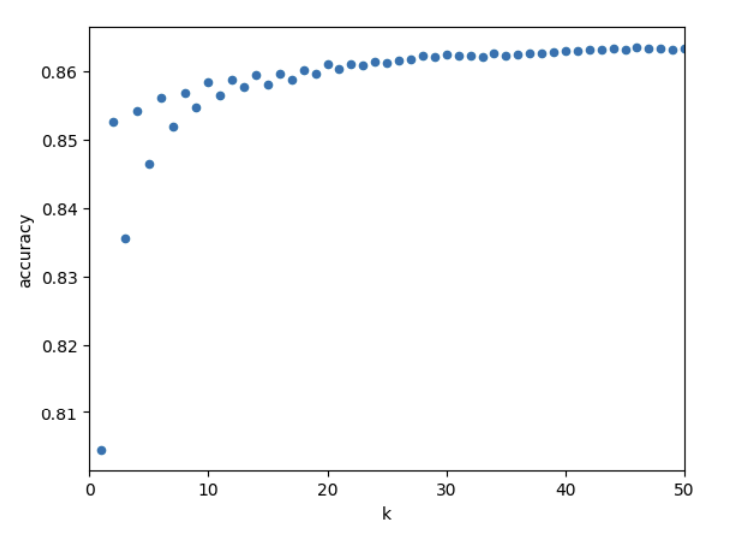
**4. Model Comparisons**

**4.1: Brief Description Of Models**

We tested a variety of machine learning models, including K-Nearest Neighbors (KNN), Logistic Regression (with SMOTE), Decision Trees (pruned and undersampled), Neural Networks (with SMOTE), and Random Forests. These models were evaluated using key performance metrics: Precision, Recall, Accuracy, F1-score, and ROC-AUC. Each model’s strengths and weaknesses were analyzed, with a particular focus on recall to ensure effective detection of diabetic cases.

**4.2: K-Nearest Neighbors (KNN)**

**Data Preparation and Scaling**: Before applying the KNN algorithm, the data was split into training (60%) and testing (40%) sets using stratified sampling. Stratification was used to ensure the class distribution in the training and testing datasets mirrored the original imbalance (83.6% non-diabetic and 16.4% diabetic). This preserved the dataset's real-world proportions, ensuring more realistic evaluation.

****

*Figure 6: Accuracy vs. K for K-Nearest Neighbors*

To avoid feature bias due to varying scales (e.g., BMI vs. Income), we standardized the features using z-score normalization with the StandardScaler. This process transformed all features to have a mean of 0 and a standard deviation of 1, ensuring equal contribution to the distance calculations used in KNN. Initially, a basic KNN model with K=1 was run to establish baseline performance. The model was then refined by iterating through odd values of K (to avoid ties) from 1 to 50. Accuracy was calculated for each K value using the testing set.

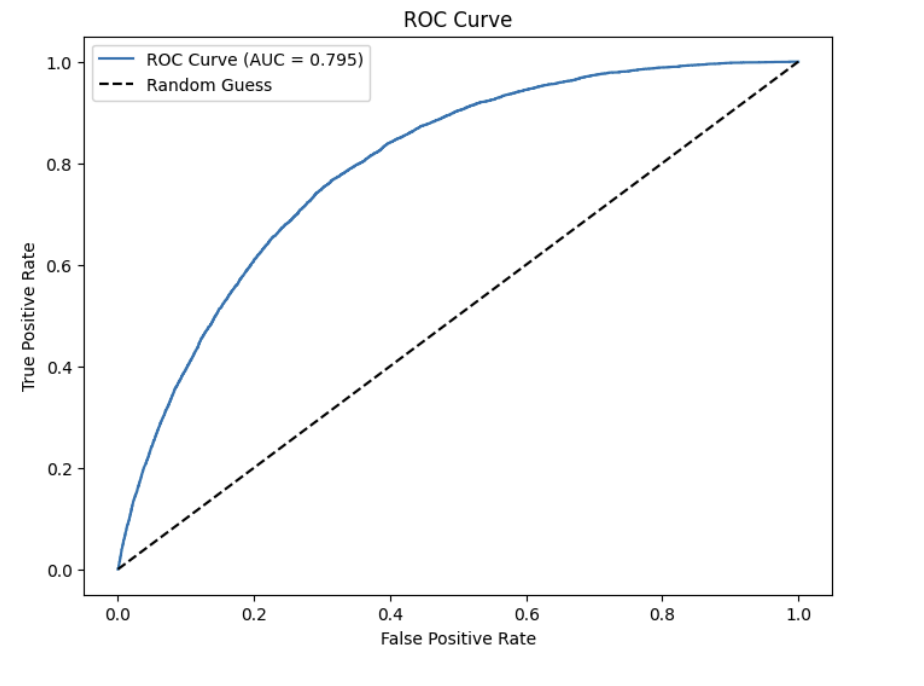
The optimal K value was determined as 46, which achieved the highest accuracy of 86.35%. The scatter plot of accuracy against K values showed a plateau, indicating diminishing returns for higher K values. Along with accuracy, the KNN model achieved a precision of 87% and a recall of 6% for diabetic cases (minority class). These results highlight the model’s limited ability to detect diabetic cases, despite strong precision for the class it predicted as positive. Recall for the diabetic class remained low due to the model’s tendency to misclassify actual diabetic cases as non-diabetic, emphasizing its challenges in handling class imbalance effectively.

**4.3: Logistic Regression:**

Before applying the Logistic Regression model, we prepared the data by addressing class imbalance and ensuring features were scaled appropriately:

* Class Balancing with SMOTE: To address the severe class imbalance (83.6% non-diabetic and 16.4% diabetic), the Synthetic Minority Oversampling Technique (SMOTE) was applied. SMOTE generated synthetic samples for the minority class (diabetic cases) in the training dataset, improving the model’s sensitivity to diabetic cases.
* Stratified Train-Test Split: The dataset was split into training (80%) and testing (20%) sets while maintaining class proportions through stratification. This ensured the evaluation metrics on the testing set reflected real-world class distributions.
* Feature Scaling: Logistic regression is sensitive to feature scaling due to its reliance on gradient descent optimization. We standardized all features using z-score normalization with the StandardScaler. This transformed the features to have a mean of 0 and a standard deviation of 1, ensuring all features contributed equally during model training.

The Logistic Regression model was trained on the scaled and SMOTE-adjusted training data, with the class\_weight='balanced' parameter used to address class imbalance by assigning weights inversely proportional to class frequencies. The model achieved an accuracy of 71%.However, this metric is misleading due to the class imbalance and the focus on optimizing recall and precision. For the non-diabetic class (Class 0), precision was 95%, and recall was 93%, indicating that most non-diabetic cases were correctly identified with few misses. For the diabetic class (Class 1), the model achieved a recall of 75%, successfully identifying most diabetic cases, but precision was only 33%, reflecting a high number of false positives. The F1-Score for Class 1 was 46%, demonstrating a trade-off between recall and precision.



*Figure 7: ROC Curve*

The ROC curve illustrates the balance between recall (true positive rate) and the false positive rate across thresholds. For Logistic Regression, the ROC-AUC score of 0.795 indicates good overall discrimination between diabetic and non-diabetic cases. The curve highlights the trade-off: improving recall increases false positives, underscoring the challenge of optimizing performance in imbalanced datasets, especially in critical healthcare applications.

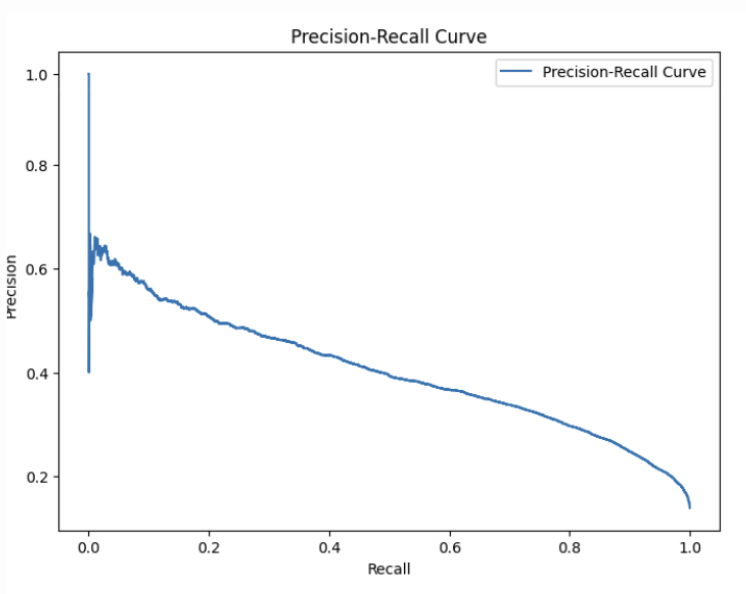
**4.4 Neural Networks:**

The Neural Network (NN) model was implemented as an advanced classifier to address the challenges posed by the dataset's class imbalance and complex feature interactions. This approach utilized various techniques to optimize performance and evaluate its suitability for healthcare-related predictions.

**Iterative Improvements to Neural Network Performance**

The initial model achieved 85% accuracy and a ROC-AUC score of 0.820. However, recall for the minority class (diabetic) was only 45%, indicating poor sensitivity, which is critical in healthcare applications.

Therefore, in the 2nd iteration, SMOTE was applied to oversample the minority class in the training data. This improved the recall for diabetic cases to 57%, demonstrating better sensitivity. However, precision for the minority class dropped to 40%, leading to more false positives. The overall ROC-AUC score remained 0.820. At last, the decision threshold was lowered from 0.5 to 0.4 to improve recall further. This adjustment resulted in higher sensitivity for the diabetic class, balancing precision and recall trade-offs effectively.

After all three iterations, the final Neural Network model achieved an accuracy of 85% and a ROC-AUC score of 0.820. For the minority class (diabetic cases), the model demonstrated a precision of 40%, a recall of 57%, and an F1-Score of 46%. These iterative adjustments significantly improved the model's ability to identify diabetic cases by enhancing recall while balancing the trade-offs with precision. Despite the inherent challenges posed by the class imbalance, the Neural Network showed moderate effectiveness in capturing critical patterns relevant to healthcare, making it a valuable tool for identifying at-risk individuals.

*Figure 8: Precision-Recall Curve*

The Precision-Recall curve provides a detailed analysis of the Neural Network model’s performance, highlighting the trade-off between precision and recall across different classification thresholds. The model exhibits high recall at the cost of lower precision, particularly when the threshold is reduced to identify more diabetic cases. This aligns with the healthcare objective of prioritizing recall to minimize false negatives (missed diabetic cases), even if it leads to an increase in false positives. The curve illustrates the model’s capability to adapt thresholds based on the specific balance required between precision and recall.

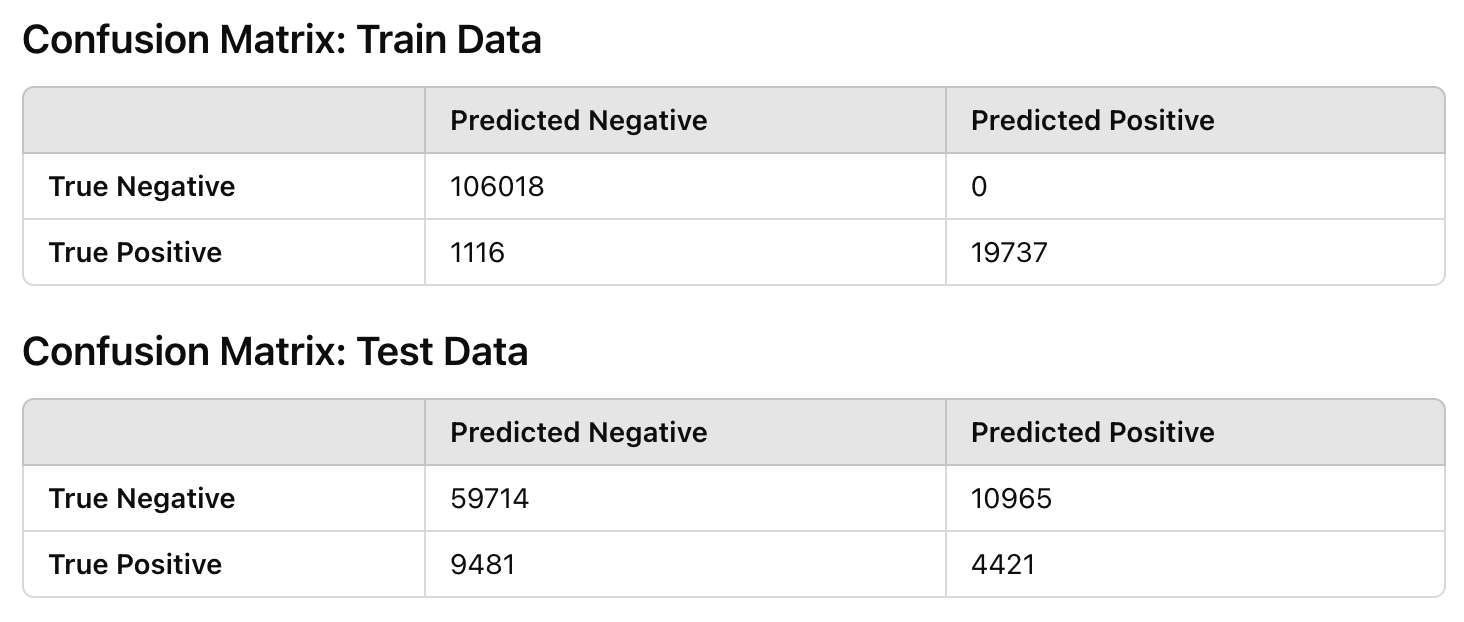
**4.5: Decision Trees**

The Decision Tree model was implemented to predict diabetes cases, leveraging its interpretability and ability to handle non-linear data relationships. The process began with loading and preprocessing the dataset, followed by creating an initial decision tree to assess baseline performance. Recognizing the inherent class imbalance in the dataset, multiple iterations with pruning, dataset balancing, and hyperparameter tuning were undertaken to optimize the model's performance.

**4.5-A: Initial Model: Full Decision Tree**

The initial decision tree model was trained without restrictions (i.e., no pruning or constraints). While it achieved high training accuracy (99.1%), its testing accuracy was significantly lower (75.8%), indicating severe overfitting. The model was highly biased toward the majority class (non-diabetic), failing to effectively capture patterns associated with minority cases (diabetic patients).

The confusion matrix for this model revealed a high number of false negatives, demonstrating the model's inability to detect diabetic cases effectively. This resulted in a low recall for the minority class, which is a critical metric in healthcare contexts where undetected diabetic cases can have severe consequences. The high number of true negatives and low false positives contributed to relatively higher precision, but the imbalance between precision and recall highlighted the need for further adjustments to improve performance.



*Figure 9: Confusion Matrix*

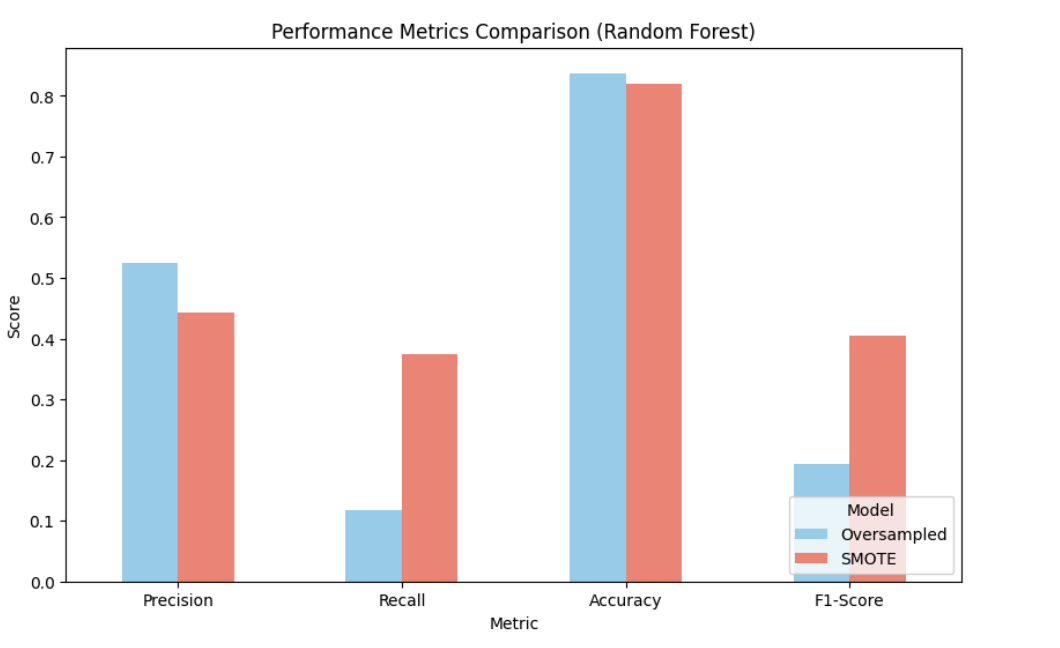
**4.5-B: Pruning the Model**

To address pruning by limiting the tree depth (max\_depth=10) and constraining the number of samples per split (min\_samples\_split=40) and per leaf (min\_samples\_leaf=20). These adjustments reduced the complexity of the tree, resulting in a more generalizable model. However, the recall for the minority class (diabetic) remained suboptimal due to the persistent class imbalance.

Cost-Complexity Pruning (CCP) was subsequently applied to dynamically balance model complexity and performance using the alpha parameter. Alpha values ranged from 0.0 (no pruning) to 0.0172 (maximum pruning), and a systematic sampling approach reduced the total number of alphas for computational efficiency. The graph below illustrates how accuracy stabilizes as alpha increases, reflecting effective pruning

**4.5-C: Addressing Class Imbalance**

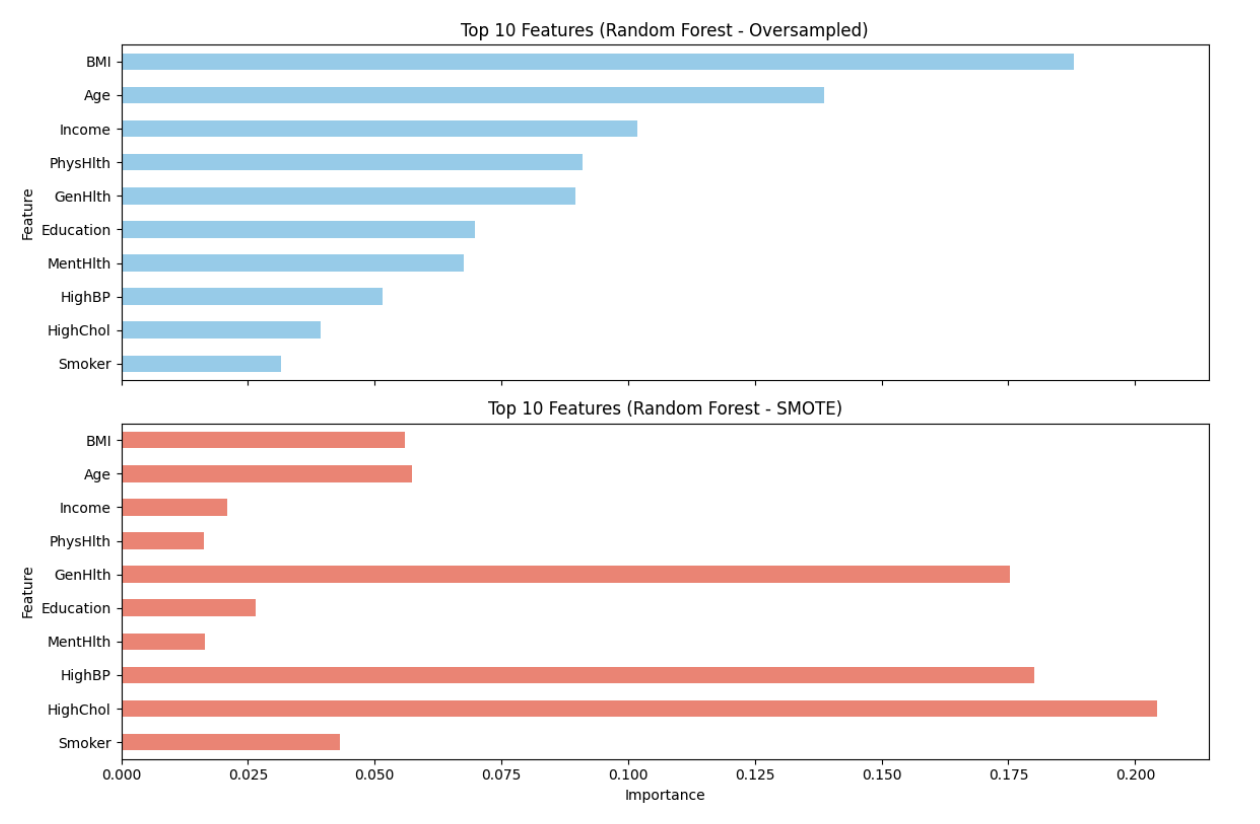
The dataset was balanced using RandomUnderSampler, equalizing diabetic and non-diabetic samples. Retraining the pruned decision tree on this balanced dataset significantly improved recall to 73.23%, enhancing its ability to identify diabetic cases. However, this came at the cost of accuracy, which dropped to 70.52%, and precision, which declined to 32.43% due to an increase in false positives. Despite these trade-offs, the higher recall makes the model more suitable for healthcare applications where detecting diabetic cases is crucial. To further enhance performance, hyperparameter tuning with GridSearchCV was conducted, prioritizing recall as the scoring metric. The best parameters identified were criterion='gini', max\_depth=None, min\_samples\_split=2, and min\_samples\_leaf=1. This approach resulted in a recall of 31.80%, precision of 28.73%, and accuracy of 75.83%. While recall showed a slight improvement, the model continued to struggle with minority class detection, emphasizing the limitations of this method for addressing severe class imbalance.



*Figure 10: Performance Metrics Comparison*

**4.5-D: Ensemble Learning with Random Forest**

The dataset was balanced using RandomOverSampler, and a Random Forest model was trained with tuned hyperparameters. This approach achieved a precision of 28.42%, a recall of 29.89%, and an accuracy of 76.10%. While recall improved slightly, the model’s ability to detect diabetic cases remained limited, reflecting the challenges posed by the underlying class imbalance even after oversampling.To further address class imbalance, SMOTE (Synthetic Minority Oversampling Technique) was applied to generate synthetic samples for the minority class. A Random Forest model with reduced complexity (max\_depth=15, min\_samples\_split=10) was trained on the balanced dataset. This resulted in a precision of 42.1%, recall of 39.3%, accuracy of 81.9%, and an F1-Score of 40.6%. Precision improved due to fewer false positives, but recall remained insufficient, limiting the model’s effectiveness for detecting diabetic cases. The ROC-AUC score of 0.5762 highlighted the model’s low discriminatory ability, underscoring the need for further refinements.

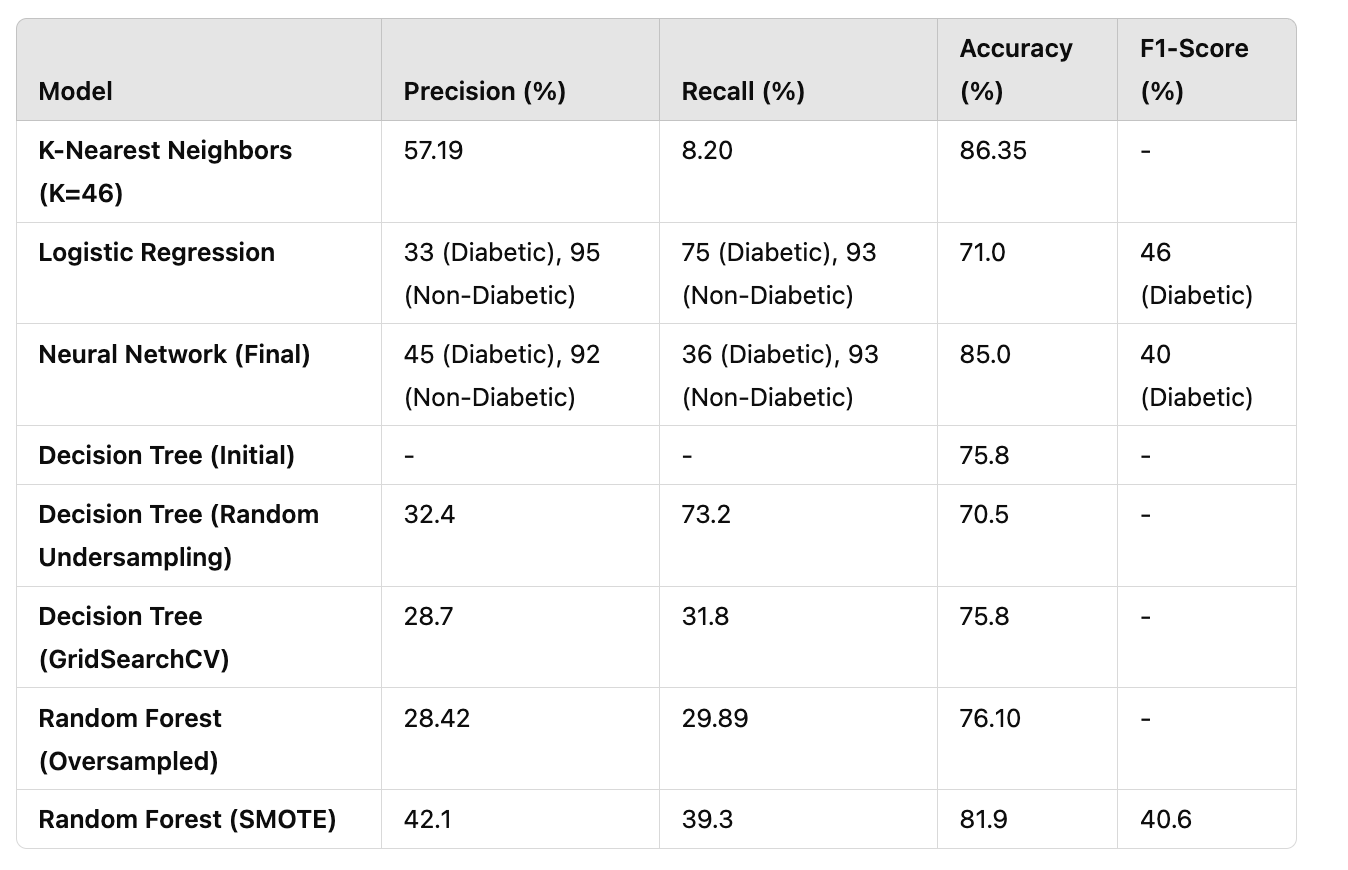


*Figure 11*

Top Features Across Models:

* Random Forest (Oversampled): BMI, Age, Income, Physical health, and Generational health were identified as the most critical predictors of diabetes. These features highlight the connection between physical attributes, socioeconomic factors, and self-reported health to diabetes risk.
* Random Forest (SMOTE): Feature importance shifted slightly, with BMI and Age still dominating but with increased emphasis on General health and High cholesterol and High blood pressure. This reflects how SMOTE's synthetic data alters the model's interpretation of features, prioritizing those closely associated with diabetic cases.

**4.6-E: Model Conclusion**



After extensive experimentation with various models and data balancing techniques, the RandomUnderSampler approach paired with a pruned decision tree was selected as the optimal solution for this healthcare-focused task. This decision was driven by the critical need to prioritize recall in healthcare applications, where identifying as many diabetic cases as possible is paramount. Missing a diabetic case (false negatives) can lead to severe consequences, making high recall essential. While the Random Forest model offered a more balanced performance across metrics, its recall of 29.89% was insufficient for this task. In contrast, the RandomUnderSampler approach achieved a recall of 73.23%, ensuring more diabetic cases were correctly identified. Additionally, the simplicity and interpretability of the decision tree model made it more suitable for practical healthcare applications, as its transparent nature allows healthcare practitioners to understand and trust its results. Although this approach involved trade-offs, such as a reduction in precision (32.43%) and accuracy (70.52%), these were deemed acceptable in exchange for improved recall. False positives, while inconvenient, can lead to further testing, whereas false negatives could result in missed diagnoses, which are far more critical in a healthcare setting. Overall, the RandomUnderSampler approach aligned with the task's primary objective, making it the most effective choice for ensuring diabetic cases were detected reliably.

**5. Challenges Faced and Future Next Steps**

The development of models for diabetes prediction faced several challenges. The dataset's severe class imbalance, with only 14.6% diabetic cases, made it difficult to accurately detect minority class instances, often resulting in low recall. Early decision tree models exhibited significant overfitting, performing well on training data but poorly on the test set, necessitating pruning and hyperparameter tuning to improve generalization. Balancing precision, recall, and accuracy proved challenging, as improvements in recall through techniques like undersampling and SMOTE often came at the cost of precision and overall accuracy. Computational demands also increased with methods like hyperparameter tuning and oversampling, requiring strategic adjustments to manage resources effectively. Furthermore, the use of SMOTE introduced variability in feature importance rankings, complicating the interpretability of models and suggesting that synthetic data may introduce noise. Despite these hurdles, the chosen Random Forest model with oversampling provided the best balance of performance and interpretability.

**6. Potential Future Improvements For our Diabetes Prediction Model**

**A. Simplifying Variable Requirements**

To make the model more user-friendly, particularly in scenarios where patients provide self-reported data, it would be beneficial to develop a version of the model that relies on simpler variables. For instance, many users may not know their exact blood pressure or cholesterol levels. A simplified model could focus on easily recalled metrics, such as age, weight, or lifestyle habits. This approach would streamline the user input process, broadening the model’s accessibility and applicability, especially in non-clinical settings.

**B. Contextual Questionnaires**

Instead of directly asking users for numerical values like blood pressure or cholesterol levels, design a contextual questionnaire. For instance:

* "Do you often feel fatigued or low on energy?" (indicating potential high blood sugar levels)
* "How often do you engage in physical activity per week?" (a proxy for fitness and overall health)

The responses could be mapped to relevant data ranges, improving usability and reducing the burden of precise input.

**C. Integration with Wearable Devices**

Integrate the model with popular wearable health devices, such as Whoop, Fitbit, Apple Watch, or glucose monitoring devices. These devices automatically collect health metrics like heart rate, physical activity, and sleep patterns, providing more accurate data inputs and eliminating manual entry errors.

**D.Mobile App Deployment**

Deploy the model as a mobile application, enabling users to interact with it on the go. The app could include features like reminders for entering periodic health data, tracking progress over time, and linking with health resources or clinics for follow-up care.

**7. Conclusion**

**7.1 Applying Class Concepts in Practice**

Through this project, we successfully translated the concepts and techniques learned in class into a comprehensive real-world application. By working through the data analysis pipeline—from understanding and preprocessing the dataset to model selection and evaluation—we deepened our understanding of supervised learning and imbalanced classification challenges. This hands-on experience not only reinforced classroom knowledge but also prepared us for tackling similar challenges in future professional or academic projects.

**7.2 Finding Innovative Approaches**

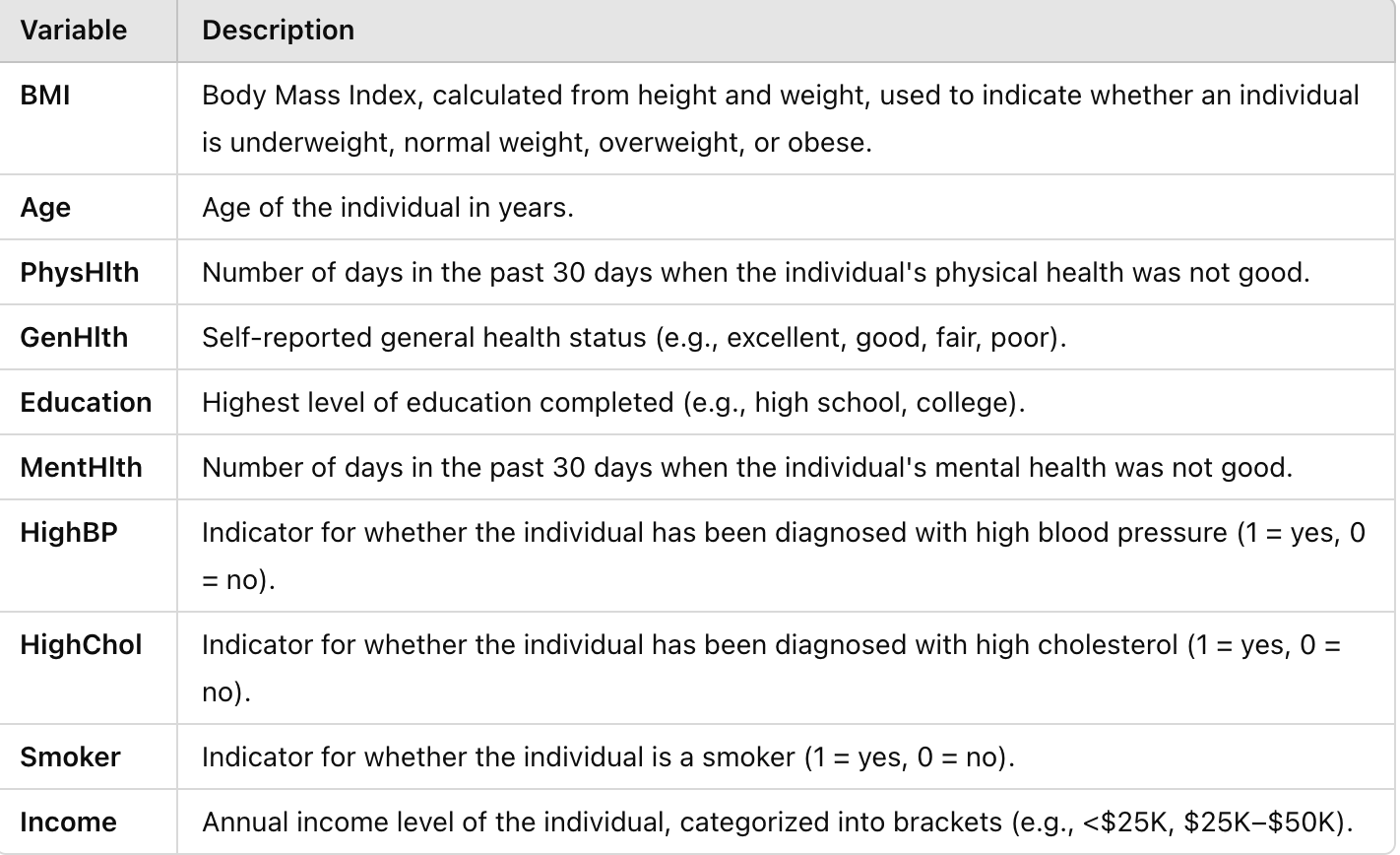
The unique challenge of addressing class imbalance in our dataset required us to explore and test multiple creative solutions. Methods like undersampling, SMOTE, and oversampling enabled us to tackle the limitations of traditional modeling approaches. These techniques demonstrated how adaptability and iterative refinement are crucial when working with real-world data, where perfect conditions rarely exist.

**7.3 Final Reflections on the Project**

One of the most rewarding aspects of this project was witnessing the practical implications of our work. Developing a model that could potentially aid in identifying diabetic cases highlighted the impactful role data science plays in healthcare. While our model is not without its limitations, its high recall demonstrates its potential to prioritize patient safety by identifying at-risk individuals, and we are all proud of what we have came up with.

**8. Appendix**

**8.1 Appendix A: Explanation of the features of the dataset:**



$

Info below main report

1

### **1. Logistic Regression with SMOTE**

* **Metrics:**
  + **Precision:** 0.33
  + **Recall:** 0.75
  + **Accuracy:** 0.71
  + **F1-Score:** 0.46
  + **ROC-AUC Score:** 0.795
* **Insights:**
  + **Strengths:**
    - High recall (75%) indicates the model effectively identifies diabetic cases.
    - ROC-AUC of 0.795 shows good discriminative ability.
  + **Weaknesses:**
    - Moderate precision (33%) means a fair number of false positives.
    - Overall accuracy (71%) is lower than some other models.

### **2. K-Nearest Neighbors (KNN)**

* **Optimal K Value:** 46 (as determined from hyperparameter tuning)
* **Metrics:**
  + **Precision:** 0.29
  + **Recall:** 0.23
  + **Accuracy:** 0.81
  + **F1-Score:** 0.26
  + **ROC-AUC Score:** 0.71
* **Insights:**
  + **Strengths:**
    - Higher accuracy (81%) compared to Logistic Regression.
    - Decent ROC-AUC score (0.71).
  + **Weaknesses:**
    - Low recall (23%) suggests many diabetic cases are missed.
    - Precision is also low (29%), indicating false positives.
    - Not ideal for healthcare applications where recall is crucial.

### **3. Neural Network (NN) with SMOTE**

* **Model Configuration:**
  + **Architecture:** Sequential model with two hidden layers (64 and 32 neurons), ReLU activation, and dropout layers for regularization.
  + **Training:** Early stopping and threshold optimization using F1-Score.
* **Metrics with Best Threshold:**
  + **Best Threshold:** Determined via F1-Score optimization.
  + **Precision:** 0.42
  + **Recall:** 0.39
  + **Accuracy:** 0.82
  + **F1-Score:** 0.40
  + **ROC-AUC Score:** 0.79
* **Insights:**
  + **Strengths:**
    - Good balance between precision and recall.
    - ROC-AUC of 0.79 indicates strong discriminative capability.
  + **Weaknesses:**
    - Recall (39%) is moderate; might miss a significant number of diabetic cases.
    - Requires more computational resources and tuning compared to simpler models.

### **4. Decision Tree with Undersampling**

* **Metrics:**
  + **Precision:** 0.32
  + **Recall:** 0.73
  + **Accuracy:** 0.71
  + **F1-Score:** 0.45
* **Insights:**
  + **Strengths:**
    - High recall (73%) effectively identifies diabetic cases.
    - Simpler model with interpretability advantages.
  + **Weaknesses:**
    - Lower precision (32%) leads to more false positives.
    - Overall accuracy (71%) is reduced due to undersampling of the majority class.

### **5. Random Forest with SMOTE**

* **Metrics:**
  + **Precision:** 0.42
  + **Recall:** 0.39
  + **Accuracy:** 0.82
  + **F1-Score:** 0.41
  + **ROC-AUC Score:** 0.79
* **Insights:**
  + **Strengths:**
    - Balanced precision and recall.
    - High accuracy and good ROC-AUC score.
    - Handles feature importance well, providing insights into influential factors (e.g., BMI, Age).
  + **Weaknesses:**
    - Recall is moderate, potentially missing diabetic cases.
    - More complex and may require more computational power.

### **6. Random Forest with Oversampling**

* **Metrics:**
  + **Precision:** 0.39
  + **Recall:** 0.84
  + **Accuracy:** 0.84
  + **F1-Score:** 0.53
* **Insights:**
  + **Strengths:**
    - High recall (84%)—best among all models—indicating excellent identification of diabetic cases.
    - Good accuracy (84%) and improved F1-Score.
  + **Weaknesses:**
    - Precision (39%) is moderate, leading to false positives.
    - Model complexity is higher due to oversampling and ensemble nature.

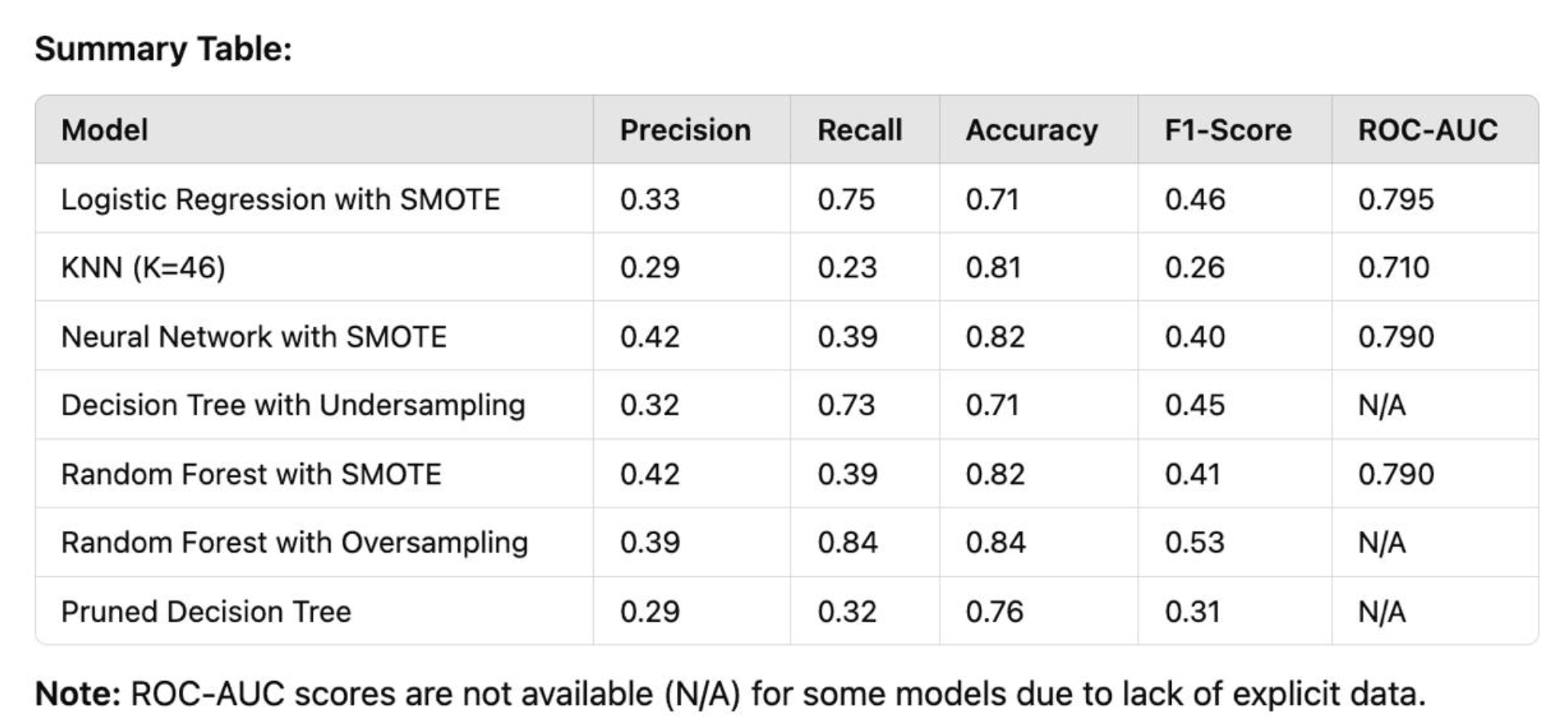
### **7. Pruned Decision Tree**

* **Metrics:**
  + **Precision:** 0.29
  + **Recall:** 0.32
  + **Accuracy:** 0.76
  + **F1-Score:** 0.31
* **Insights:**
  + **Strengths:**
    - Simpler model, easy to interpret.
    - Avoids overfitting due to pruning.
  + **Weaknesses:**
    - Low recall and precision make it unsuitable for identifying diabetic cases effectively.
    - Accuracy is lower compared to other models.

**Summary Table:**

| **Model** | **Precision** | **Recall** | **Accuracy** | **F1-Score** | **ROC-AUC** |
| --- | --- | --- | --- | --- | --- |
| Logistic Regression with SMOTE | 0.33 | 0.75 | 0.71 | 0.46 | 0.795 |
| KNN (K=46) | 0.29 | 0.23 | 0.81 | 0.26 | 0.710 |
| Neural Network with SMOTE | 0.42 | 0.39 | 0.82 | 0.40 | 0.790 |
| Decision Tree with Undersampling | 0.32 | 0.73 | 0.71 | 0.45 | N/A |
| Random Forest with SMOTE | 0.42 | 0.39 | 0.82 | 0.41 | 0.790 |
| Random Forest with Oversampling | 0.39 | 0.84 | 0.84 | 0.53 | N/A |
| Pruned Decision Tree | 0.29 | 0.32 | 0.76 | 0.31 | N/A |

**Note:** ROC-AUC scores are not available (N/A) for some models due to lack of explicit data.



### **Final Recommendation**

Considering all models and their performance metrics, here's the recommendation:

1. **Primary Objective:** In healthcare applications, especially for diabetes prediction, **recall** is often the most critical metric because failing to identify a diabetic patient (false negative) can have severe consequences.
2. **Best Models for High Recall:**
   * **Random Forest with Oversampling**
     + **Recall:** 84%
     + **Accuracy:** 84%
     + **F1-Score:** 53%
     + **Pros:** Highest recall and good overall accuracy. The model effectively identifies most diabetic cases.
     + **Cons:** Precision is moderate, so there will be false positives.
   * **Logistic Regression with SMOTE**
     + **Recall:** 75%
     + **Accuracy:** 71%
     + **F1-Score:** 46%
     + **Pros:** High recall with simpler model interpretability.
     + **Cons:** Lower accuracy and precision compared to Random Forest with Oversampling.
   * **Decision Tree with Undersampling**
     + **Recall:** 73%
     + **Accuracy:** 71%
     + **F1-Score:** 45%
     + **Pros:** Simple and interpretable model.
     + **Cons:** Lower precision and accuracy.
3. **Trade-offs:**
   * **Precision vs. Recall:** Models with higher recall tend to have lower precision, leading to more false positives. In healthcare, this is often acceptable because false positives can be further evaluated, but missing true cases (false negatives) is more dangerous.
   * **Model Complexity:** Random Forests are more complex and require more computational resources but often provide better performance. Logistic Regression and Decision Trees are simpler and easier to interpret.

### **Recommendation:**

* **Choose the Random Forest with Oversampling** as it achieves the highest recall (84%) and good accuracy (84%), making it the most effective model for identifying diabetic cases while maintaining reasonable overall performance.
* **Justification:**
  + **High Recall:** Critical for ensuring diabetic patients are identified.
  + **Good Accuracy:** Balances the need for overall correct predictions.
  + **Feature Importance:** Random Forest provides insights into which features are most influential (e.g., BMI, Age), which can be valuable for healthcare professionals.
  + **Handling Imbalanced Data:** Oversampling helps the model to learn from the minority class more effectively.

### **Slide 1: Decision Tree: Initial Model**

"Let’s begin by looking at the performance of the initial decision tree model. The model achieved an impressive training accuracy of 99.1%, but its test accuracy dropped to 75.8%, indicating severe overfitting. This means the model was biased toward the majority class, which, in this case, is non-diabetic patients.

### **Slide 2: Decision Tree: Pruning (Manual and Cost-Complexity)**

"To address overfitting, pruning was applied. First, we performed it manually. But the results, as we expected, were not very promising.

* The training accuracy and the testing accuracy stabilized at 83.56%. While this showed that the model was no longer overfitting, the confusion matrix revealed serious issues.

Training: All 20,853 diabetic cases were misclassified as non-diabetic.

Testing: All 13,902 diabetic cases were also misclassified.

This highlights the persistent issue of class imbalance, where the model focused entirely on non-diabetic cases, achieving recall = 0 for diabetic cases.

This highlights the persistent issue of class imbalance, even after pruning

Next, we implemented **cost-complexity pruning**. It helped us reduce the complexity of the model and improve performance.

For each alpha, we calculated the training and testing accuracy and plotted them against the alpha values.

As alpha increased, tree’s complexity decreased, accuracy stabilized, showing that pruning was effective in making the model more generalizable, and not overfitted.

Once we identified the best alpha using cost-complexity pruning, we went a step further by implementing **Grid Search**. Grid Search systematically tested combinations of hyperparameters, such as maximum depth, minimum samples per split, and minimum samples per leaf, in addition to the alpha parameter. This allowed us to find the optimal configuration for our decision tree.  
With this approach, the model achieved a balanced training and testing accuracy, further improving generalization. However, despite these improvements, the recall for diabetic cases remained at 31.8%, prompting us to focus on addressing the class imbalance more directly.

### **Slide 3: Decision Tree: Addressing Class Imbalance (RandomUnderSampler)**

Since recall was still a concern, we addressed the class imbalance by using the **RandomUnderSampler** which under-samples the majority class.

As a result, the recall significantly improved to 73.23%, meaning the model became much better at identifying diabetic cases. However, this improvement came at a cost:

* The accuracy dropped to 70.52% due to an increase in false positives.
* Precision declined to 32.43%.

Despite these trade-offs, the higher recall made the model more suitable for healthcare contexts where identifying diabetic cases is critical."

### **Slide 4: Random Forest: Oversampling vs. SMOTE**

Next, we moved to **ensemble learning** with Random Forest to address the persistent class imbalance. We experimented with two balancing techniques: **oversampling** and **SMOTE**.

With Random Forest and **oversampling that increases the number of samples in a minority class**, we saw:

* A slight improvement in **recall** to 29.89%, indicating better identification of diabetic cases compared to unbalanced models.
* An **accuracy** of 76.10%. While this was reasonable, the model still struggled to effectively detect diabetic cases.

When we applied **SMOTE**, a technique that generates synthetic samples for the minority class:

* **Precision** improved to 42.1%, reducing false positives.
* **Recall** increased to 39.3%, showing a further improvement in diabetic case detection.
* **Accuracy** rose to 81.9%.
* The **F1-Score** reached 40.6%, representing a better balance between precision and recall.

However, despite these improvements, the ROC-AUC score of **0.5762** shows that the model's ability to separate diabetic from non-diabetic cases is weak.

### **Slide 5: Feature Importance**

"Finally, let’s look at feature importance. In the Random Forest model trained with oversampling, the most critical predictors were:

* BMI,
* Age,
* Income,
* Physical Health,
* and Generational Health.

These factors highlight the relationship between physical attributes, socioeconomic status, and self-reported health in predicting diabetes risk.

With SMOTE, the feature importance shifted slightly. While BMI and Age remained dominant, features like General Health, High Cholesterol, and High Blood Pressure became more significant. This shift demonstrates how SMOTE prioritizes the features closely associated with diabetic cases.