**Diabetes Prediction Project Report**

**1. Data Exploration and Cleaning**

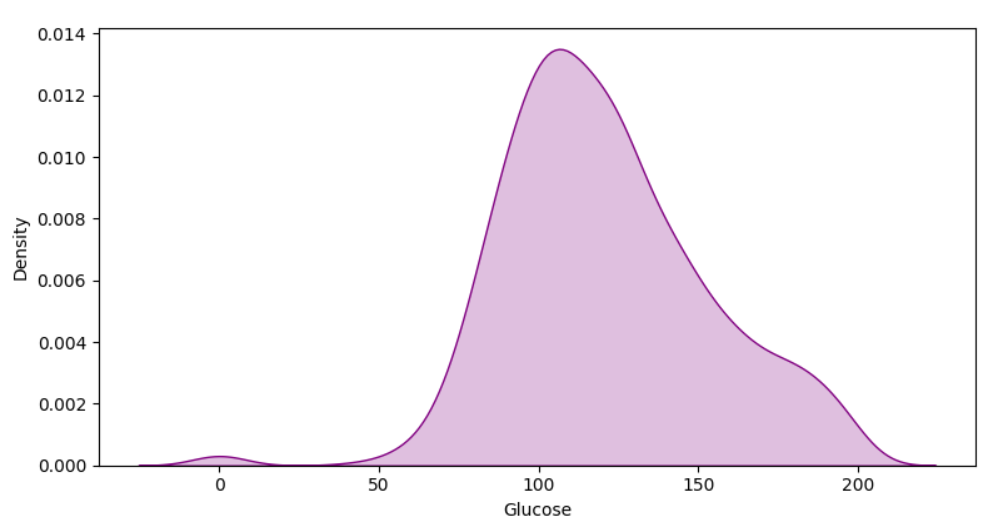
**Summary Statistics and Data Visualizations**

The dataset consists of multiple health-related attributes used to predict diabetes outcomes. Below are key statistics:

* The average glucose level is high, indicating potential diabetes risk.
* BMI values show a wide distribution, with some outliers.
* Blood pressure values are mostly within a normal range but it also have some values as 0 so some missing data is there.
* A screenshot of a computer screen

  Description automatically generatedInsulin levels exhibit significant variation.

**Key Insights from Visualizations:**

* The distribution of glucose, BMI, and age is right-skewed.

A purple line graph with numbers

Description automatically generatedA purple and white graph

Description automatically generated

**Data Cleaning Process**

**Handling Outliers:**

* The IQR method was used to detect and remove extreme outliers in insulin and skin thickness.

**Handling Missing Values:**

* Missing values were found in insulin, skin thickness.
* A screenshot of a computer

  Description automatically generatedWe applied median imputation for these missing values to retain as much data as possible.

**2. Feature Engineering**

**New Features Created and Justification**

To enhance model performance, we created the following features:

* **Glucose-BMI Index:** Combining glucose and BMI to better capture metabolic risk.
* **Age-BMI Index:** Combining Age and BMI.

**Feature Selection Process**

Feature selection was conducted using:

* **Correlation Analysis:** Dropped highly correlated features to avoid redundancy.

**3. Model Building and Evaluation**

**Models Built and Reasoning Behind Selection**

We experimented with multiple models:

1. **Logistic Regression**
2. **Random Forest Classifier**
3. **Support Vector Machine (SVM**
4. **Decision Tree**
5. **Gradient Boosting (XGBoost**

**Performance Evaluation**

We evaluated each model using:

* **Accuracy, Precision, Recall, F1-score, and AUC-ROC.**
* Cross-validation to ensure robustness.

**Final Chosen Model**

* A graph of a normalized curve

  Description automatically generated with medium confidenceA diagram of a normal value

  Description automatically generated with medium confidenceA screenshot of a computer program

  Description automatically generated**Random Forest was selected as the best model** due to its superior AUC-ROC and Recall which is important for outcomes like in this project.

**4. Code**

The entire process, including data preprocessing, feature engineering, and model building, is documented in a Jupyter Notebook.

* The code follows a modular structure for readability and reproducibility.
* Proper comments are included for clarity.

**5. Project Summary & Key Insights**

**Key Takeaways:**

1. **Glucose and BMI are strong predictors of diabetes.**
2. **Feature interaction improved model performance, particularly with new composite metrics.**
3. **Random Forest outperformed other models, achieving the highest Recall.**
4. **Addressing missing values and outliers significantly enhanced model reliability.**

**Future Improvements:**

* Collect additional data to improve generalizability.
* Test deep learning models for enhanced performance.
* Deploy the model as a web app for real-world usability.

**End of Report**