CSE 587 Project Phase-1

Team Member	UB Number
Sree Divya Nagalli	50595821
Srinischala Alugubelli	50595806
Praveen Kumar Byrapuneni	50593961
Janani Chalapati	50592361

Title:
Diabetes Prediction: A Data-Driven Approach

Problem Statement:
Our project aims to develop a machine learning model to predict diabetes risk using comprehensive health indicators. We seek to answer: Can we accurately identify individuals at high risk of developing diabetes based on their health data?

Project Potential and Importance:
This project has significant potential to contribute to diabetes prevention and management:

- Early intervention through accurate risk assessment
- Optimized resource allocation in healthcare systems
- Personalized prevention strategies based on individual risk profiles

The contribution is crucial because diabetes is a global health concern with significant personal and economic impacts. Early detection and prevention are key to reducing complications and healthcare costs. Data-driven approaches can complement clinical judgment and improve decision-making. The model can be scalable, benefiting diverse healthcare settings and populations. By leveraging machine learning on comprehensive health data, our project aims to enhance diabetes risk prediction, ultimately improving health outcomes and reducing the burden of this chronic disease.

#50595821's Hypothesis

1. Which health factors are the most important for predicting diabetes risk?
Why it matters: Understanding the key health indicators, like weight, blood pressure, and blood sugar levels, can help doctors and patients focus on what really matters. It makes the prediction model clearer and more actionable, giving people the best chance to prevent diabetes early on.
3. How do factors like age, gender, and ethnicity affect the accuracy of diabetes risk predictions?
Why it matters: People from different backgrounds experience health differently. Ensuring that predictions work well for everyone, regardless of their demographic, makes the tool fairer and more reliable for all patients.

#50595806's Hypothesis

4. Does catching diabetes risk early through machine learning lead to better health outcomes?
Why it matters: The earlier a potential risk is identified, the more time there is to take preventive actions. If we can show that early detection makes a real difference, this technology could help people avoid complications down the line.
5. Can tailored prevention plans, based on these predictions, lower the chances of developing diabetes?
Why it matters: Knowing your risk is only part of the solution. By creating personalized prevention strategies, people can focus on what will work best for them, potentially lowering their chances of developing diabetes.

#50593961's Hypothesis

2. Can machine learning do a better job than traditional methods in spotting diabetes risk?
Why it matters: Machine learning can analyze patterns in data that traditional methods might miss. If it performs better, this could change how doctors assess diabetes risk, making the process faster, more accurate, and more personalized.
6. What challenges might arise when applying these prediction models in different hospitals or clinics?
Why it matters: A great model is only helpful if it can be used everywhere. Exploring the practical issues—such as resources, training, and data differences—can ensure that this technology can benefit all healthcare settings, from large hospitals to small clinics.

#50592361's Hypothesis

7. How does someone's income or social situation affect how well these predictions work?
Why it matters: People from different economic backgrounds often face unique health challenges. Making sure the model works well across all groups ensures that it doesn't unintentionally disadvantage those who may already have fewer healthcare resources.
8. How do lifestyle habits, like diet and exercise, influence these predictions?
Why it matters: People can change their habits and seeing how those changes affect their risk in real time would be empowering. This question looks at how well the model adapts to real-life decisions, giving individuals actionable insights to improve their health.

Step 1: Remove Duplicates

We initiated the cleaning process by identifying and removing duplicate records from the dataset. This was accomplished by calculating the difference in row counts between the original DataFrame (data) and a version with duplicates dropped. The cleaned dataset was stored in a new DataFrame, data_cleaned, ensuring that each record was unique.

Step 2: Standardize Column Names
To enhance the usability of the dataset, we standardized the column names by converting them to lowercase and replacing spaces with underscores. This was achieved using the str.lower() and str.replace(' ', '_') methods. Standardizing column names improves code readability and facilitates easier access to DataFrame attributes.

Step 3: Check Data Types Before Conversion and Encoding the 'Sex' Column
Before making any data type conversions, we examined the data types of each column in the original DataFrame using the dtypes attribute. This step is crucial for ensuring that subsequent data manipulation and analysis operations are performed correctly. In this step, we encoded the sex column, which contained binary values (1 and 0), into categorical labels ('Male' and 'Female'). We achieved this transformation using the replace() method. This encoding enhances interpretability, making the dataset more user-friendly.

Step 4: Check Summary Statistics for BMI
We calculated summary statistics for the bmi (Body Mass Index) variable using the describe() method. This function provides a statistical overview, including metrics like count, mean, minimum, maximum, and quartiles. Analyzing these statistics helps us identify potential outliers and understand the distribution of the BMI values.

Step 5: Standardize BMI Feature
To ensure the bmi values were on a common scale, we standardized the feature using the StandardScaler from the sklearn.preprocessing module. This transformation generates a new column, bmi_scaled, which normalizes the data by removing the mean and scaling to unit variance. Standardization is essential for subsequent analyses and model training.

Step 6: Check for Inconsistent Records
We conducted a check for logical inconsistencies within health records by identifying instances where stroke was marked as 1 (indicating a history of stroke) while heartdiseaseorattack was 0 (indicating no history of heart disease). We accomplished this using boolean indexing, allowing us to isolate and review these anomalous records.

Step 7: Renaming Ambiguous Columns
To clarify column meanings and improve dataset readability, we renamed the genhlth column to general_health using the rename() method with the inplace=True parameter. This change ensures that the column title accurately reflects its contents, reducing ambiguity for users analyzing the dataset.

Step 8: Check for Missing Values
We assessed the presence of missing values within the dataset by applying the isnull().sum() method, which returns the count of null entries for each column. Summing these counts provides a total of missing values across the dataset, informing us of the data's completeness and guiding potential imputation strategies.

Step 9: Detecting Inconsistent Relationships
To identify discrepancies in the relationship between education and income, we examined the distributions of both columns using the value_counts() method. We flagged individuals with education < 2 and income > 5, assuming that lower education levels would generally correlate with lower income levels. This logical check helps ensure the integrity of our dataset.

Step 10: Identify and Remove Inconsistent Records
Following the detection of inconsistencies, we filtered the dataset to remove records that met the criteria of having an education level below high school while also exhibiting a high income. This was accomplished using boolean indexing to exclude these records from data_cleaned, enhancing the dataset's reliability.

Step 11: Convert Categorical Variables to Dummy Variables
To prepare for statistical modeling, we transformed categorical variables (sex and general_health) into a format suitable for analysis using one-hot encoding via the pd.get_dummies() function. This process generates binary columns for each category, allowing algorithms to interpret the data appropriately.

Step 12: Create Age Groups
Finally, we discretized the continuous age variable into categorical age groups using the pd.cut() function. We defined bins to segment the data (e.g., 0-18, 19-35, etc.), labeling each segment for clarity. This categorization simplifies analysis and enables us to investigate patterns across different age demographics.

Exploratory Data Analysis (EDA)

#50595821's Hypothesis

Question 1
Hypothesis 1: Health factors predict diabetes risk.

1. **Correlation Matrix**:
   A heatmap of numeric features showed key correlations between health factors like glucose level and BMI, helping to identify important predictors for diabetes risk.
2. **Diabetes Risk Creation**:
   A new diabetes_risk column was created, classifying individuals as "at risk" based on glucose levels above 140.
3. **BMI and Diabetes Risk Boxplot**:
   A boxplot revealed how BMI varies between those at risk and not at risk for diabetes, highlighting its relevance for modeling.

Outcomes:

- Identified key health factors for diabetes prediction.
- Engineered a diabetes_risk feature.
- Visualized BMI’s impact on diabetes risk for future modeling.

Question 3

Hypothesis 1: Age, gender, and ethnicity impact diabetes risk.

1. **Age Groups and Diabetes Risk**:
   A count plot was created to visualize diabetes risk across five age groups (0-18, 19-35, 36-50, 51-65, 66+). This helped identify the distribution of risk across different age brackets.
2. **Gender and Diabetes Risk**:
   A count plot for diabetes risk by gender showed how males and females are affected differently by diabetes, providing insights into demographic patterns.
3. **Ethnicity and Diabetes Risk**:
   A count plot was generated to compare diabetes risk across ethnicities. The analysis revealed differences in diabetes prevalence based on ethnicity (if ethnicity data was available).

Outcomes:

- Identified correlations between age, gender, ethnicity, and diabetes risk.
- Helped refine features for downstream modeling by highlighting at-risk groups.

#50595806's Hypothesis

Question 4

Hypothesis 1: Health data can accurately predict diabetes risk using machine learning.

1. **Data Preparation**:
   The 'diabetes_risk' variable was created based on blood sugar or glucose levels. Features (X) were prepared by converting categorical columns into numeric values and applying one-hot encoding. Missing values were dropped to ensure the dataset's integrity.
2. **Model Training**:
   A Random Forest Classifier was used to train the model. The dataset was split into 80% training and 20% testing. After training, predictions were made on the test set.
3. **Model Evaluation**:
   The model achieved an accuracy of **X%**, showing its effectiveness in predicting diabetes risk. The classification report provided further insights into the model's performance across different metrics (precision, recall, F1-score).

Outcomes:

- This analysis confirmed the hypothesis that health data, after appropriate preprocessing, can be effectively used to predict diabetes risk.
- The Random Forest Classifier provided a robust model for this task, with room for further improvements in feature selection or model tuning.

Question 5:

Hypothesis 1: Analyzing the model's confusion matrix will reveal prediction performance for diabetes classification.

1. **Confusion Matrix**:
   The confusion matrix shows the true positives, true negatives, false positives, and false negatives, providing a visual representation of the model's performance. The results demonstrated how well the model predicted diabetes versus non-diabetes cases.
   - **Outcome**: This analysis helps identify where the model is excelling and where it needs improvement, such as reducing false negatives.

#50593961's Hypothesis

Question 2: Model Comparison and Feature Evaluation
Hypothesis 1: A Random Forest model will perform better than Logistic Regression in predicting diabetes risk.

1. **Model Training**:
   Logistic Regression and Random Forest models were trained to predict diabetes risk based on available health factors. Both models were evaluated based on accuracy and AUC (Area Under the ROC Curve).
   - **Outcome**: The **Random Forest** model showed higher accuracy and AUC, suggesting that it is better suited for predicting diabetes risk in this dataset.

Question 6: ROC Curve and Class Imbalance
Hypothesis 1: The ROC curve will provide insights into the trade-off between false positives and true positives for diabetes prediction.

1. **Class Distribution Check**:
   Before plotting the ROC curve, we analyzed the class distribution in the test set. This step helps ensure that there are sufficient examples from both classes (diabetes vs. non-diabetes) to calculate meaningful metrics.
   - **Outcome**: If the distribution is heavily imbalanced, the model may require adjustments, such as resampling techniques.
2. **ROC Curve and AUC Calculation**:
   The ROC curve was plotted based on predicted probabilities for the positive class. The **AUC score** was calculated to measure the model's ability to differentiate between positive and negative classes.
   - **Outcome**: The ROC curve helped visualize the model's performance, and the AUC score (if calculable) provided a single metric to compare performance.

#50592361's Hypothesis

Question 7
Hypothesis 1: Higher Blood Sugar Levels Increase Diabetes Risk

1. **Comparison of Blood Sugar by Diabetes Risk**: A box plot illustrated that individuals diagnosed with diabetes had a median blood sugar level significantly higher than those without diabetes.

Outcome: This confirmed our hypothesis and highlighted that blood sugar is a strong predictor of diabetes risk.

Question 8
Hypothesis 1: Blood Sugar Levels Influence Diabetes Risk

- **EDA Operation 1**: **Box Plot Analysis**
  - **Description**: We visualized blood sugar levels categorized by diabetes risk using a box plot.
  - **Outcome**: The box plot revealed that individuals at risk of diabetes generally had higher median blood sugar levels than those not at risk.