Report

Detailed project plan for analyzing the CDC Diabetes Health Indicators dataset

1. Data Understanding

- **Goal**: Gain familiarity with the data and understand the key features and their roles.
- **Actions**:
  - **Dataset Overview**: Load the dataset and inspect it. Review the number of records and features.

2. Data Preprocessing

- **Goal**: Clean and prepare the data for analysis and machine learning tasks.
- **Actions**:
  - **Validity Check**: Perform validity checks for all indicators except `BMI` and `ID` to ensure that their values fall within the expected ranges. This verification is essential for maintaining data integrity and ensuring that each indicator has appropriate, predefined values before proceeding with analysis.
    - The reason for not validating `BMI` could be that it is a continuous feature with a wide and natural range of values that can vary significantly depending on individual characteristics. Unlike categorical indicators with fixed valid values, `BMI` can naturally have outliers that represent valid and important data, especially in the context of health-related studies. Removing or altering these values might lead to loss of significant information, such as identifying individuals at extreme health risks, which is crucial for a diabetes-related analysis.
  - **Aggregation Index**:Indicators were summarized based on papers on domain knowledge
    - A factor combining
    `HighBP` and `HighChol` represent overall cardiovascular risk.
    - A factor combining
    `Fruits` and `Veggies` represent diet quality.

    - A factor combining `Smoker` and `HvyAlcoholConsump` represent unhealthy behavior.
    - A factor combining `AnyHealthcare` and `NoDocbcCost` represent healthcare accessibility.
  - **Data Transformation**: Scale continuous variables such as `Age`, `Income`, and `BMI` using Min-Max scaling.
  - **Class Imbalance**: Check for class imbalance in the target variable (`Diabetes_binary`). If needed, apply oversampling techniques such as SMOTE or undersampling to balance the dataset.

3. Exploratory Data Analysis (EDA)

- **Goal**: Gain insights into relationships between features and the target variable.
- **Actions**:
  - **Descriptive Statistics**: Compute summary statistics for each feature (mean, median, standard deviation, etc.) using Spark's `describe()` function.
  - **Correlation Analysis**: Use correlation heatmaps to visualize relationships between the numeric features and the target variable.
  - **Univariate Analysis**: Visualize the distribution of individual features like `Age`, `BMI`, `Income` to detect trends in healthy vs. diabetic populations.
  - **Bivariate Analysis**: Analyze relationships between pairs of variables, such as `BMI` and `Diabetes_binary`, `HighChol` and `Diabetes_binary`, using box plots or scatter plots.
  - **Chi-square Tests**: For categorical variables, run chi-square tests to check their independence with the target variable.

4. Feature Engineering and Dimensionality Reduction

- **Goal**: Reduce dimensionality and explore latent structures within the dataset while retaining important information.
- **Actions**:
  - **Principal Components Analysis (PCA)**:

        - Apply PCA to reduce the dimensionality of the dataset and identify the most important principal components.
        - Visualize the explained variance to determine how many components should be retained.
        - Analyze how each feature contributes to the principal components and whether patterns related to diabetes emerge.
    - **Factor Analysis**:
        - Perform factor analysis to identify underlying latent variables that could explain relationships between the observed variables.
        - Interpret the factors in relation to health indicators and lifestyle variables.

5. Machine Learning Algorithms

- **Goal**: Implement a broader set of machine learning algorithms to explore different approaches to classification.
- **Actions**:
    - **Logistic Regression**: Baseline for binary classification.
    - **Decision Trees**: Non-linear model to handle interactions between features.
    - **Random Forests**: Ensemble method to improve performance through multiple decision trees.
    - **Gradient-Boosted Tree (GBT)**:
        - Explore Gradient-Boosted Tree for high accuracy and efficient training.
    - **Support Vector Machines (SVM)**: Linear and non-linear kernels to handle complex decision boundaries.
    - **Naive Bayes**: Useful for handling categorical variables and class imbalance.
    - **Neural Networks**:
        - Implement a simple feed-forward neural network to explore deep learning approaches.

        - Use Multi-Layer Perceptron (MLP) from Spark's MLlib for classification.

6. Model Comparison and Tuning

- **Goal**: Thoroughly compare a larger variety of models to select the best performing one.
- **Actions**:
    - Use cross-validation to compare the performance of each algorithm based on:
        - **Accuracy, Precision, Recall, F1-Score**.
    - Use **Grid Search** or **Random Search** to optimize hyperparameters for each model.
    - Select the final model based on a combination of predictive power, interpretability, and computational efficiency.

7. Final Reporting and Documentation

- **Goal**: Summarize the findings and document the results.
- **Actions**:
    - **Summary of Findings**: Present key insights from the EDA and predictive modeling, such as which features are the strongest predictors of diabetes and their impact.
    - **Model Interpretation**: Explain how the model predicts diabetes using interpretable AI techniques. Utilize **SHAP** (SHapley Additive exPlanations) to understand the global feature importance and their effects on predictions.
    - **Recommendations**: Based on the analysis, provide actionable insights, such as lifestyle or health recommendations for diabetes prevention.

Tools and Libraries:

- **Apache Spark (PySpark)**: For distributed data processing and machine learning.
- **Matplotlib/Seaborn**: For creating visualizations (EDA, feature importance).

- **SHAP**: For model interpretation and understanding feature contributions to the predictions.