PROJECT WRITEUP: “HealthCare: Diabetic or Non-Diabetic”

DESCRIPTION:

Objective:

The objective of the project is to build a predictive model that accurately determines whether or not a patient has diabetes based on a set of medical predictor variables. The target variable, called "Outcome," is binary and represents whether a patient has diabetes or not, with 1 indicating the presence of diabetes and 0 indicating the absence of diabetes.

Problem Statement:

NIDDK (National Institute of Diabetes and Digestive and Kidney Diseases) research creates knowledge about and treatments for the most chronic, costly, and consequential diseases.

- The dataset used in this project is originally from NIDDK. The objective is to predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset.
- Build a model to accurately predict whether the patients in the dataset have diabetes or not.

Domain: As the ML Developer assigned to the NIDDK, you have been asked to create ML predictive model.

Content: Dataset: (“health care diabetes.csv” : Click here to download dataset)

Fields in the data –

Column Name	Description
Pregnancies	Number of times pregnant
Glucose	Plasma glucose concentration in an oral glucose tolerance test
BloodPressure	Diastolic blood pressure (mm Hg)
SkinThickness	Triceps skinfold thickness (mm)
Insulin	Two hour serum insulin
BMI	Body Mass Index
DiabetesPedigreeFunction	Diabetes pedigree function
Age	Age in years
Outcome	Class variable (either 0 or 1). 268 of 768 values are 1, and the others are 0.

Project Task: Week 1

Data Exploration:

1. Perform descriptive analysis. Understand the variables and their corresponding values. On the columns below, a value of zero does not make sense and thus indicates missing value:
    - Glucose
    - BloodPressure
    - SkinThickness
    - Insulin
    - BMI

2. Visually explore these variables using histograms. Treat the missing values accordingly.

3. There are integer and float data type variables in this dataset. Create a count (frequency) plot describing the data types and the count of variables.

4. Check the balance of the data by plotting the count of outcomes by their value. Describe your findings and plan future course of action.

5. Create scatter charts between the pair of variables to understand the relationships. Describe your findings.

6. Perform correlation analysis. Visually explore it using a heat map.

Project Task: Week 2

Data Modeling:

1. Devise strategies for model building. It is important to decide the right validation framework. Express your thought process.
2. Apply an appropriate classification algorithm to build a model.
3. Compare various models with the results from KNN algorithm.
4. Create a classification report by analyzing sensitivity, specificity, AUC (ROC curve), etc.

Data Reporting:

Create a dashboard in tableau by choosing appropriate chart types and metrics useful for the business. The dashboard must entail the following:

- Pie chart to describe the diabetic or non-diabetic population.
- Scatter charts between relevant variables to analyze the relationships
- Histogram or frequency charts to analyze the distribution of the data.
- Heatmap of correlation analysis among the relevant variables
- Create bins of these age values: 20-25, 25-30, 30-35, etc. Analyze different variables for these age brackets using a bubble chart.

Challenges and Limitations:

1. **Data Quality and Missing Values:**
   - **Challenge:** The presence of missing values in the dataset could impact the quality of the predictions. Imputing these missing values is essential for meaningful analysis.
   - **Limitation:** Imputing missing values might introduce bias if not handled carefully.

2. **Class Imbalance:**
   - **Challenge:** Assessing the balance of the Outcome variable (diabetic or non-diabetic) to avoid biased model performance.
   - **Limitation:** Imbalanced datasets can lead to skewed model predictions, especially if not addressed during model training.

3. **Model Selection and Evaluation Metrics:**
   - **Challenge:** Choosing the right machine learning models and evaluation metrics is critical for accurate predictions.
   - **Limitation:** Different algorithms may perform differently, and the choice might impact the model's accuracy.

4. **Validation Framework:**
   - **Challenge:** Selecting the right validation framework for model evaluation.
   - **Limitation:** Improper validation methods may lead to overfitting or underfitting, affecting the model's generalizability.

5. **Interpretability:**
   - **Challenge:** Ensuring the interpretability of the selected model, especially if the application is in a medical setting.
   - **Limitation:** Complex models may provide accurate predictions but lack interpretability, making it challenging for healthcare professionals to trust and understand the results.

6. **Dashboard Design:**
   - **Challenge:** Creating an effective and informative dashboard in Tableau.
   - **Limitation:** The effectiveness of the dashboard depends on the clarity of visualization and the relevance of chosen metrics; a poorly designed dashboard may hinder decision-making.

7. **Age Binning:**
   - **Challenge:** Binning age values and analyzing variables for specific age brackets.
   - **Limitation:** The choice of age brackets may impact the granularity of insights, and defining them arbitrarily might miss important trends.

8. **Business Relevance:**
   - **Challenge:** Ensuring that the chosen metrics and visualizations in the dashboard align with the business needs.
   - **Limitation:** If the dashboard lacks relevance to key business questions, it may not provide actionable insights.

Findings:

- **Correlation Analysis:**
  - The correlation matrix shows the relationships between variables. Notably, the "Glucose" variable has the highest positive correlation with the "Outcome" variable, followed by "Age" and "BMI." This suggests that these variables may be important in predicting diabetes.

- **Model Performance:**
  - The AUC (Area Under the ROC Curve) and accuracy scores are used to evaluate the performance of different machine learning models. The Random Forest Classifier has the highest AUC score (0.991) and accuracy (0.870), indicating it performs exceptionally well on this dataset.

  - The Decision Tree Classifier also shows strong performance with an AUC score of 0.967 and an accuracy of 0.865.

- **F1 Score and Precision-Recall Curves:**
  - F1 scores and precision-recall curves are additional metrics that consider the balance between precision and recall. Random Forest and Decision Tree classifiers have high F1 scores, indicating a good balance between precision and recall. This means these models can effectively classify both true positive and true negative cases.

Recommendations:

- **Model Selection:**
  - The Random Forest Classifier and Decision Tree Classifier appear to be the best-performing models in terms of AUC, accuracy, and F1 score. These models provide an excellent balance between precision and recall, making them strong candidates for predicting diabetes**.**

- **Further Analysis:**
  - Conduct feature importance analysis for Random Forest and Decision Tree models to understand which variables contribute most to the predictions. This can provide valuable insights for healthcare professionals.

- **Hyperparameter Tuning:**
  - Consider optimizing the hyperparameters of the selected models, especially the Random Forest and Decision Tree, to potentially improve their performance further.

- **Ensemble Models:**
  - You might also explore ensemble models, such as bagging and boosting, to see if combining multiple models can enhance prediction accuracy.

- **Communication:**
  - Clearly communicate the strengths and limitations of the selected models to stakeholders.
  - Highlight the importance of continuous monitoring and updating of the model as more data becomes available.

Thank you!