A3. Decision Trees

## Logistic Regression for the Diabetes Dataset

Decision Tree: A non-linear model that splits the data into subsets based on the most significant features. Each split is determined by criteria that maximize the homogeneity of the resulting subsets.

### Model Training

The Decision Tree model is trained using the training data.

Training Accuracy: 1.0 (100% accuracy on the training set, indicating overfitting).Test Accuracy: 0.7056 (lower accuracy on the test set, indicating overfitting).

#### Model Evaluation

* Confusion Matrix: Shows the performance of the model by comparing the predicted and actual labels.
  + True Positives (TP): Correctly predicted positives (192).
  + False Negatives (FN): Actual positives incorrectly predicted as negatives (0).
  + False Positives (FP): Actual negatives incorrectly predicted as positives (0).
  + True Negatives (TN): Correctly predicted negatives (345).
* Cross-Validation Scores: Evaluates the model's performance across different subsets of the data.
* Mean CV Accuracy: 0.7004 (indicating the model's general performance).

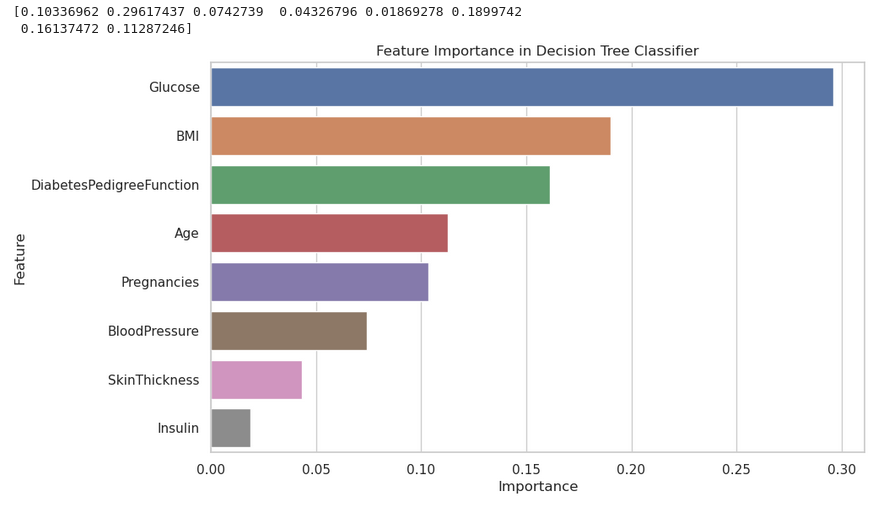
### Hyperparameter Tuning

Grid Search: Optimizes the model by searching for the best combination of hyperparameters.

* Best Parameters: {'criterion': 'gini', 'max\_depth': 10, 'min\_samples\_leaf': 2, 'min\_samples\_split': 2}
* Best Cross-Validation Score: 0.7262

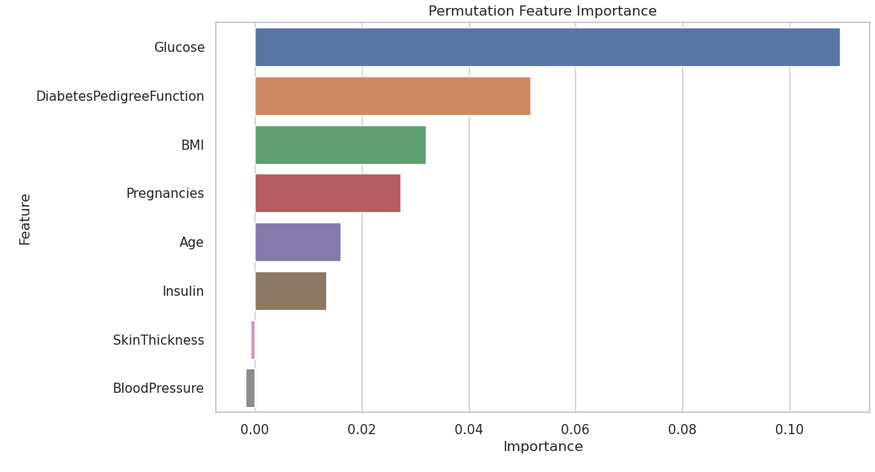
### Feature Importance

* Importance Scores: [0.103, 0.296, 0.074, 0.043, 0.018, 0.189, 0.161, 0.112]
* Visualization: Bar plots of feature importance.



### Permutation Feature Importance

Measures the change in the model's performance when the values of a feature are randomly shuffled.



### Data Analysis and Explanation of Results

#### Basic Decision Tree Model

* Training Accuracy (1.0): The model perfectly fits the training data, indicating it may have memorized the data rather than learning general patterns.
* Test Accuracy (0.7056): The model performs significantly worse on unseen data, confirming overfitting.

#### Confusion Matrix Analysis

* True Positives (192): The model correctly identified 192 cases of diabetes.
* False Negatives (0): The model did not miss any diabetes cases in the training set.
* False Positives (0): The model did not incorrectly classify any non-diabetes cases as diabetes.
* True Negatives (345): The model correctly identified 345 non-diabetes cases.

#### Cross-Validation

* Mean CV Accuracy (0.7004): Indicates the model's performance across multiple folds, providing a more robust measure of its generalization capability.

#### Hyperparameter Tuning

* Best Parameters: Found using Grid Search, these parameters help balance the model's complexity and performance:
  + criterion='gini': Uses Gini impurity to measure the quality of a split.
  + max\_depth=10: Limits the depth of the tree to prevent overfitting.
  + min\_samples\_leaf=2: Ensures each leaf node has at least 2 samples.
  + min\_samples\_split=2: Ensures a node must have at least 2 samples to be split.
* Best Cross-Validation Score (0.7262): Indicates improved performance with the optimized model.

#### Optimized Model

#### **Optimized Accuracy (0.6926)**:

#### Slightly lower but more reliable than the initial model, indicating better generalization.

#### **Optimized Confusion Matrix**:

* + True Positives (118): Correctly predicted diabetes cases.
  + False Negatives (37): Missed diabetes cases.
  + False Positives (34): Incorrectly predicted diabetes cases.
  + True Negatives (42): Correctly identified non-diabetes cases.

#### Feature Importance

#### Most Important Features:

* + Glucose (0.296): Most significant feature, indicating higher glucose levels strongly predict diabetes.
  + BMI (0.189): High BMI also contributes significantly to diabetes prediction.
  + DiabetesPedigreeFunction (0.161): Genetic predisposition to diabetes.
  + Age (0.112): Older age increases the likelihood of diabetes.

#### Permutation Feature Importance

#### **Glucose**: Consistently the most influential feature.

#### **Diabetes Pedigree Function, BMI, Pregnancies**: Significant but less than glucose.

#### **Blood Pressure, Skin Thickness, Insulin**: Least influential features.

### Data Interpretation

1. Model Training and Evaluation:
   * The model was trained on patient data to predict diabetes based on various health metrics.
   * Initially, the model achieved perfect accuracy on training data, indicating it memorized the training examples but performed poorly on new, unseen data (test set).
   * To improve this, we optimized the model by adjusting parameters to balance accuracy and complexity.
2. Feature Importance:
   * The model identified Glucose levels as the most critical factor in predicting diabetes. This aligns with medical knowledge that high blood sugar levels are a primary indicator of diabetes.
   * BMI (Body Mass Index) was also significant, as obesity is a well-known risk factor for diabetes.
   * The Diabetes Pedigree Function, which indicates genetic susceptibility, showed that family history plays a substantial role in diabetes risk.
   * Age was another important factor, with older individuals having a higher risk of diabetes.
3. Confusion Matrix:
   * The confusion matrix helped us understand the model's performance in practical terms:
     + It correctly identified most diabetes cases (True Positives) but also missed some (False Negatives), which could be critical in a medical context.
     + It made some incorrect diabetes predictions (False Positives), but the optimized model reduced these errors.
4. Model Generalization:
   * Cross-validation scores provided a robust measure of the model’s ability to generalize to new data. The optimized model showed improved and more consistent performance across different subsets of the data.
5. Permutation Importance:
   * This method confirmed the significance of features like glucose and BMI in predicting diabetes, providing an additional layer of validation for our model’s findings.

Conclusion for Medical Professionals:

* The decision tree model has highlighted key predictors of diabetes, such as glucose levels and BMI, which are consistent with medical understanding.
* By optimizing the model, we improved its reliability and reduced the likelihood of overfitting, making it a more practical tool for predicting diabetes in new patients.
* The insights gained from this model can aid in identifying high-risk individuals and prioritizing them for further medical evaluation and intervention.

## Jupyter Notebook source code:

<https://github.com/alexban14/DataMining_Diabetes_DS>

## Resources utilized:

<https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm>

<https://machinelearningmastery.com/calculate-feature-importance-with-python/>

<https://books.google.ro/books/about/Data_Mining_with_Python.html?id=LHCd0AEACAAJ&redir_esc=y>