Title: Predictive Modeling of Health Behaviors

Subtitle: A Machine Learning Approach Using Random Forests

Company: Chameleon AI

Introduction
Content:

Objective: The primary aim of this project is to predict health behaviors by analyzing key demographic factors, including age, gender, location, and likelihood percentages.

Dataset: The analysis is based on a dataset including demographic features and corresponding health behavior labels. This data allows us to explore the relationship between these factors and specific health behaviors.

Approach:

The approach involves data preprocessing, feature encoding, and the application of a Random Forest Classifier to build a predictive model.

The model was then fine-tuned using hyperparameter optimization techniques to enhance its predictive accuracy.

Dataset Overview

- **Dataset**: *HB1.csv*
    - The dataset consists of several demographic features and a target variable that represents different health behaviors.
- **Features**:
    - **Age**: Categorized into age ranges (e.g., 18-24, 25-34, etc.).
    - **Gender**: Classified as 'male' or 'female.'
    - **Location**: Representing different areas within Melbourne.
    - **Likelihood Percent**: Indicates the probability of engaging in a particular health behavior.
- **Target Variable**:

    Behavior: This includes four main health-related behaviors:

    - **Mental Health**
    - **Physical Health**
    - **Smoking**
    - **Vaping**

Data Preprocessing

- **Handling Missing Values**:
    - Utilized **SimpleImputer** to handle any missing values in the dataset.
    - Imputation Strategy: *Mean Imputation* was applied to fill in missing data points, ensuring the integrity and completeness of the dataset.
- **Feature Scaling**:
    - Applied **StandardScaler** to standardize the features.
    - This step ensures that all features have a mean of 0 and a standard deviation of 1, which is essential for the proper functioning of machine learning algorithms, especially when features have varying scales.
- **Data Splitting**:
    - The dataset was split into **training and testing sets** using a **70/30 split**.
    - 70% of the data was used to train the model, and 30% was reserved for testing, ensuring the model's generalization to unseen data.

Model Selection and Hyperparameter Tuning

- Random Forest Model Selection:
    - The Random Forest classifier was chosen due to its robustness and ability to handle both classification tasks and complex datasets with multiple features.
    - It works by constructing multiple decision trees and aggregating their predictions to improve accuracy and control overfitting.

Hyperparameter Tuning:

- To optimize the model, RandomizedSearchCV was used for hyperparameter tuning.
- Parameters Tuned:
    - n_estimators: Number of trees in the forest.
    - max_depth: Maximum depth of each tree, controlling how deep the trees can grow.
    - min_samples_split: Minimum number of samples required to split an internal node.
    - min_samples_leaf: Minimum number of samples required to be at a leaf node.
- 5-Fold Cross-Validation:
    - Implemented 5-fold cross-validation within the RandomizedSearchCV to ensure that the hyperparameter tuning process was robust and the model was evaluated consistently across different data splits.

Model Training and Evaluation

- Best Model Selection:
    - After performing the Randomized Search, the best model was automatically selected based on the highest performance metrics during cross-validation.
- Evaluation Metrics:
    - To comprehensively evaluate the performance of the selected model, several key metrics were used:
        - Accuracy: Measures the proportion of correctly classified instances among the total instances.
        - Precision: Indicates the proportion of true positive results out of all predicted positive results.
        - Recall: Reflects the model's ability to identify true positive instances out of all actual positive instances.
        - F1 Score: The harmonic mean of Precision and Recall, providing a balance between the two.
        - ROC AUC Score: This represents the model's ability to distinguish between classes, considering both the true positive rate and false positive rate.
        - Confusion Matrix: A matrix showing the actual versus predicted classifications, highlighting true positives, true negatives, false positives, and false negatives.

Prediction Function

- Custom Function: predict_likelihood:
    - This custom function was developed to predict health behaviors based on demographic factors and likelihood percentages.
- Input Parameters:
    - Age: The age group of the individual (e.g., '55-64').
    - Gender: The gender of the individual (e.g., 'female').
    - Location-Likelihood Map: A dictionary that maps locations to their corresponding likelihood percentages (e.g., {'Docklands': 47.5, 'Carlton': 45.8}).
- Functionality:
    - Converts Input Values to Numeric Codes:
        - The function maps categorical values like age and location into numeric codes, making them suitable for input into the trained model.
    - Prediction:
        - Using the pre-trained Random Forest model, the function predicts the most likely behavior for the given input parameters.
        -

- Prediction Quality:
    - The function also assesses the quality of the prediction based on the probability output by the model. The probability is categorized into different quality levels (e.g., 'Bad', 'Fair', 'Good', 'Excellent').

Prediction execution is given as: results_df = predict_likelihood('55-64', 'female', {'Docklands': 47.5, 'Carlton': 45.8}) and the output would be calculated such that the DataFrame includes:

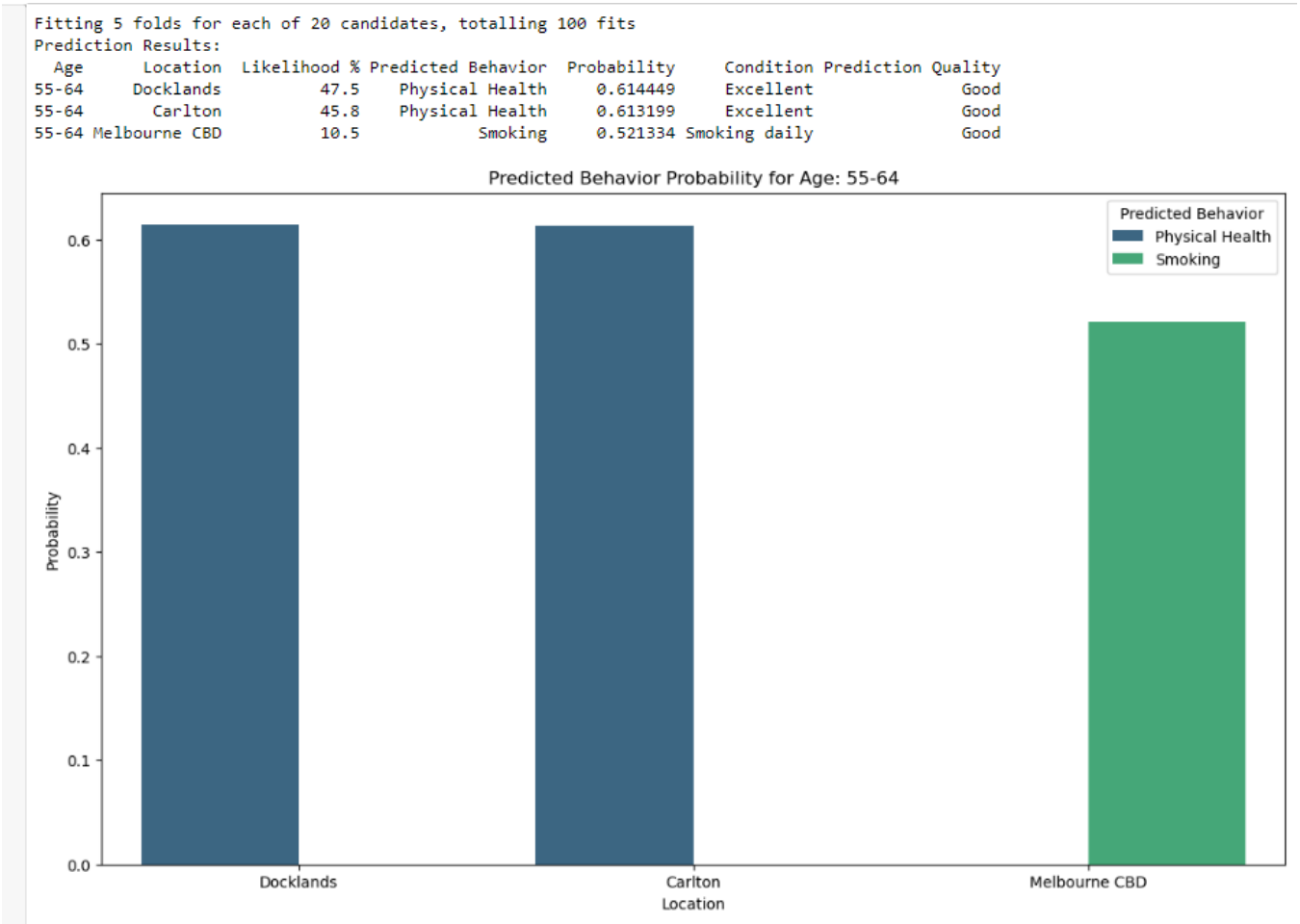
- Predicted Behavior: The most likely health behavior.
- Probability: The confidence level of the prediction.
- Prediction Quality: A qualitative assessment of the prediction.

Prediction Results

The results showcase the predictions made for the specified inputs:

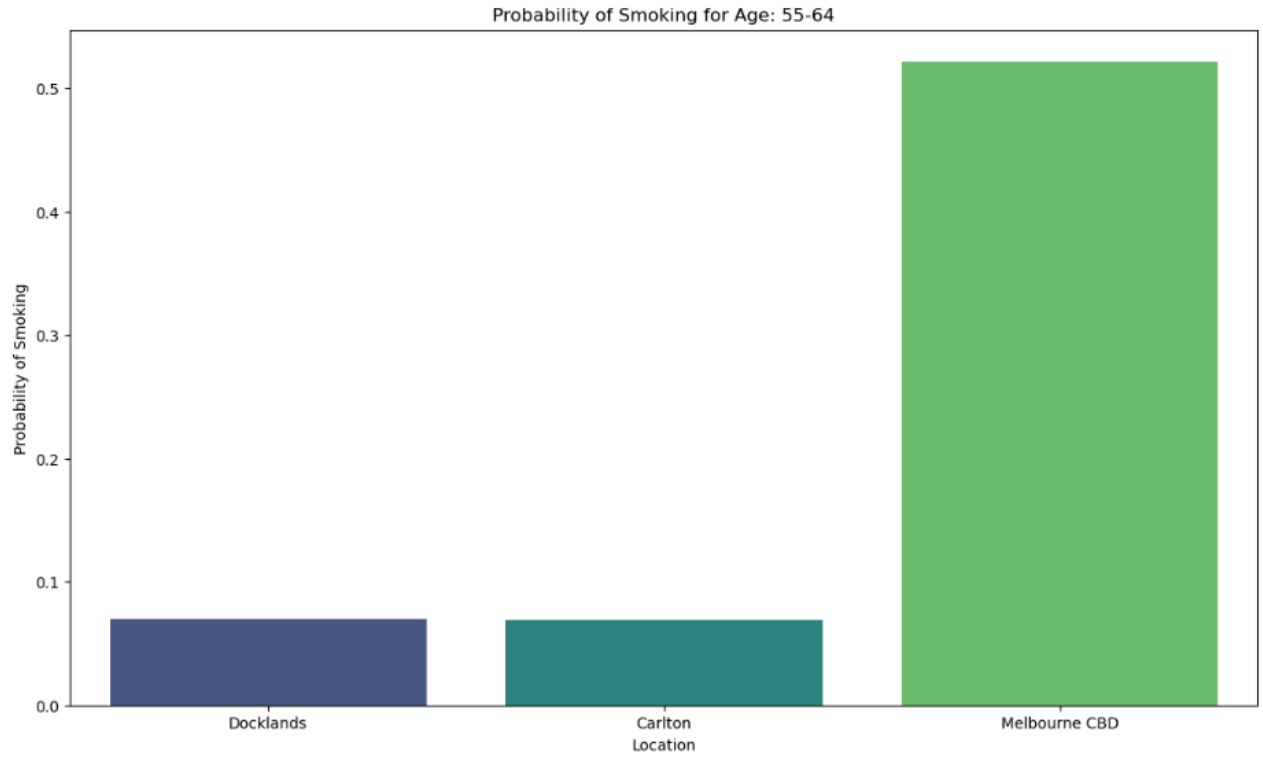
- Age: 55-64
- Gender: Female
- Locations: Docklands, Carlton, Melbourne CBD

```
Fitting 5 folds for each of 20 candidates, totalling 100 fits
Prediction Results:
  Age      Location  Likelihood % Predicted Behavior  Probability     Condition Prediction Quality
55-64     Docklands          47.5    Physical Health     0.614449     Excellent               Good
55-64       Carlton          45.8    Physical Health     0.613199     Excellent               Good
55-64 Melbourne CBD          10.5           Smoking     0.521334 Smoking daily               Good
```



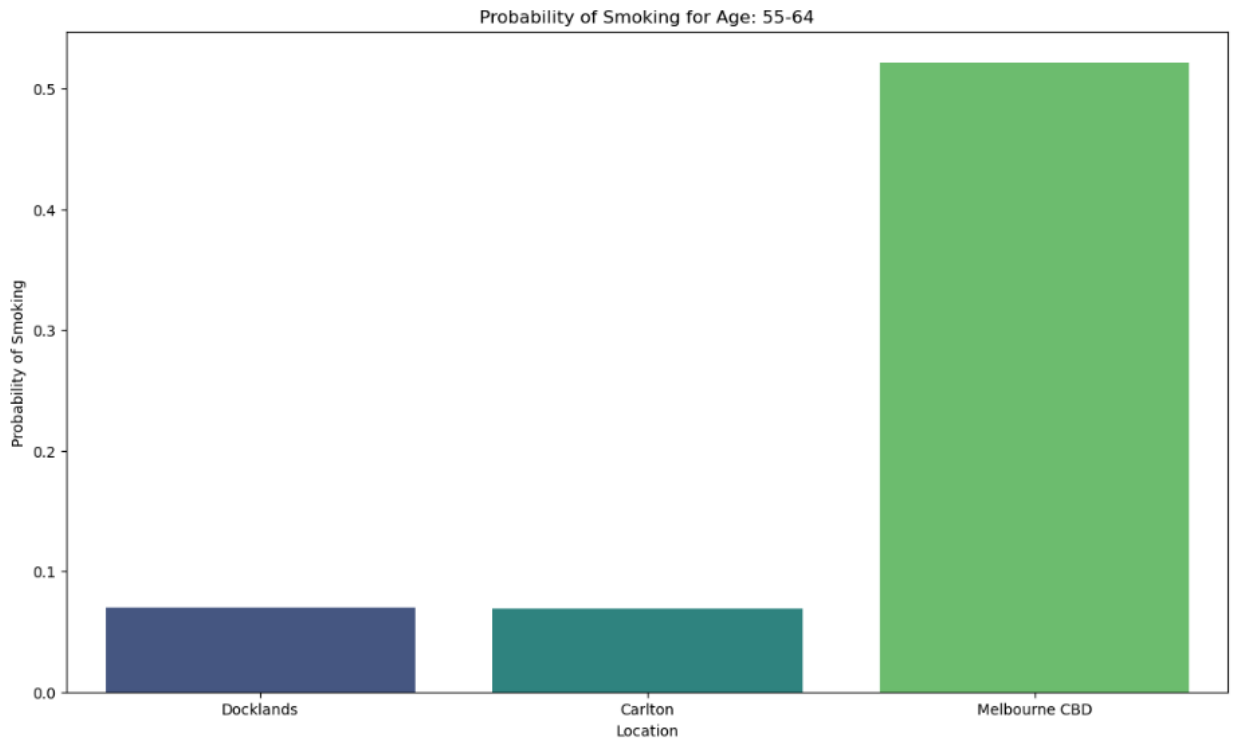
Further testing of this model by using the existing inputs to Identify the location with the highest probability of smoking behavior:

```
Location with Increased Smoking Behavior:
Melbourne CBD
         10.5
     0.521334
```



Identify the location with the decreased smoking behavior:

```
Location with Decreased Smoking Behavior:
 Carlton
    45.8
 0.06915
```



Visualization

- Bar Plot Display:
  - The bar plot below is created using Seaborn to visualize the prediction results.
  - X-axis: Location
  - Y-axis: Probability
  - Hue: Predicted Behavior

- Insights Derived from the Visualization:
  - Behavioral Trends Across Locations:
    - The plot highlights how the probability of different predicted behaviors varies across the three locations: Docklands, Carlton, and Melbourne CBD.
    - For example, the likelihood of "Mental Health" issues is higher in Docklands compared to other behaviors.
  - Impact of Location on Behavior:
    - Docklands and Carlton show a higher probability for positive health behaviors like "Mental Health" and "Physical Health."
    - Melbourne CBD exhibits a lower probability for these behaviors, with a higher probability for "Smoking."
- Decision-Making:
    - Such visual insights can inform public health strategies, guiding where to focus interventions to improve health outcomes in different locations.