OVERVIEW OF THE MODELING PROCESS

The goal of this analysis was to build a classification model to predict breast cancer diagnosis (Malignant or Benign) based on ten real-valued features from the Breast Cancer Wisconsin dataset. The dataset was loaded and cleaned. The target variable ('Diagnosis') was converted to binary values:

* + **1 = Malignant (M)**
  + **0 = Benign (B)**

Feature Scaling was applied using Min-Max Normalization to ensure compatibility with distance-based models like K-Nearest Neighbors (KNN).

The evaluate\_model function is designed to assess the performance of a classification model by training it on the given training dataset and making predictions on the test dataset. It computes key evaluation metrics, including accuracy, precision, recall, F1-score, and ROC AUC, while also providing a classification report for detailed insights. If cross-validation (CV) is enabled, the function performs k-fold CV, reporting the average scores across multiple splits to ensure robustness. Additionally, it generates a learning curve, visualizing how the model's accuracy varies with different training sizes, helping to diagnose underfitting or overfitting. This function provides a comprehensive performance analysis, making it useful for comparing multiple models and tuning hyperparameters efficiently.

Three classification techniques were tested:

1. Decision Tree Classifier
2. Logistic Regression
3. K-Nearest Neighbors (KNN)

During the model creation process every model was trained and tested using three approaches to find the best methods:

**APPROACH 1**: To fit the model without normalizing or using other methods like Cross Validation or Grid Search.

**APPROACH 2**: To fit the model with normalized data

**APPROACH 3**: To fit the model with normalized data (except for decision tree classifier), K Fold Cross Validation, and selection of best pairs of hyperparameters using Grids Search CV method. This method turns out to be the most optimal way of training a model to get its best metric values (Accuracy, ROC-AUC, and Recall [since recall is the metric of interest for this use case])

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Approach** | **Best Parameters** | **Test Accuracy (%)** | **Test Recall (%)** | **Test Precision (%)** | **Test F1 (%)** | **Test ROC AUC (%)** | **Learning Curve Result** |
| **Decision Tree** | **Approach 1** | Default | 94 | 91 | 93 | 92 | 93.2 | \_ |
| **Decision Tree** | **Approach 2** | Default | 75 | 49 | 75 | 59 | 69.4 | Underfitting |
| **Decision Tree** | **Approach 3** | criterion='entropy', max\_depth=5, min\_samples\_split=2, min\_samples\_leaf=1 | 94 | 98 | 88 | 92 | 94.5 | Overfitting |
| **Logistic Regression** | **Approach 2** | Default | 95 | 98 | 89 | 93 | 99.8 | \_ |
| **Logistic Regression** | **Approach 3** | **C=1.0, penalty='l1', solver='liblinear'** | **96** | **100** | **90** | **95** | **99.7** | **Well Fitted** |
| **K-Nearest Neighbors** | **Approach 1** | K = 5 | 75 | 53 | 74 | 62 | 81 | Underfitting |
| **K-Nearest Neighbors** | **Approach 2** | K = 5 | 96 | 93 | 95 | 94 | 99.3 | \_ |
| **K-Nearest Neighbors** | **Approach 3** | n\_neighbors=9, metric='manhattan', weights='uniform' | 96 | 98 | 93 | 95 | 99.8 | Well Fitted |

After fitting all the models in three different approaches, I found that **“LOGISTIC REGRESSION MODEL** TRAINED WITH **K FOLD CROSS VALIDATION**, WITH **NORMALIZED DATA**, AND HYPER PARAMETER COMBINATION OF **PENALTY {L1}**, **C {1}**, **SOLVER {LIBLINEAR}”** is the best fit model (trained using recall as target) for predicting breast cancer diagnosis (Malignant or Benign) for the provided dataset, since it has the highest **RECALL** and also the gap between the testing accuracy and training accuracy in the **LEARNING CURVE** is minimum compared to other models and approaches.

NOTE:

1. Overfitting and underfitting can be identified using a learning curve, which is a plot of a model's performance metric (e.g., accuracy, loss, or error) against the number of training samples. It helps assess whether a model generalizes well to unseen data.
2. In this breast cancer classification use case, recall is the key metric because identifying malignant cases correctly is more important than overall accuracy. A high recall ensures that most malignant cases are detected, minimizing false negatives, which is critical in medical diagnosis to avoid missing cancerous cases.