OPTIMIZING MACHINE LEARNING MODELS FOR CANCER DIAGNOSIS USING THE BREAST CANCER WISCONSIN DATASET

Report for COEN 432 Assignment 2 - Machine Learning for Cancer Diagnosis

By: Qian Yi Wang (ID: 40211303) - Philip Carlsson-Coulombe (ID: 40208572)

A) Problem Description

In this assignment, we aim to build and optimize machine learning models to diagnose breast cancer using the Breast Cancer Wisconsin dataset. This dataset contains 569 instances, each with 30 attributes related to various cell characteristics such as radius, texture, smoothness, and compactness. The target variable indicates whether a tumor is malignant (M) or benign (B). The challenge lies in building a model that can accurately classify new data and handle the complexity of large feature spaces and the relationships between features and the target variable.

We focus on using two machine learning models: Decision Tree (DT) and k-Nearest Neighbors (k-NN), with a goal to optimize these models for the best performance in terms of both accuracy and computational efficiency. Accuracy is evaluated based on the percentage of correctly predicted test instances, and we also focus on computational efficiency by measuring the runtime for testing.

B) Methods Description

Model Building & Training Process

- **Dataset Split**: The dataset will be split into training and testing sets, ensuring that no data instance used for training is included in the test set. The split ratio is 4:1 (training to testing).
- **Model Selection**: We will use two machine learning models:
    - **Decision Tree (DT)**: A decision tree will be used as one of the models due to its interpretability and ability to handle complex relationships in the data.
    - **k-Nearest Neighbors (kNN)**: This model is chosen because it is simple and effective for classification problems with a large number of features.
- **Model Training**: Both models will be trained on the selected training set using standard training algorithms.
- **Parameter Optimization**:
    - **Grid Search**: Grid search will be employed to find the optimal hyperparameters for both models, such as the maximum depth for the decision tree and the number of neighbors for kNN.

Model Evaluation & Validation

- **Performance Metrics**: After training, the models will be evaluated using the following metrics:
    - **Accuracy**: Percentage of correctly predicted instances.
    - **Precision**: Proportion of true positives among predicted positives.
    - **Recall**: Proportion of true positives among actual positives.
    - **F1 Score**: Harmonic mean of precision and recall.

C) Results & Conclusions

Model Performance

After implementing the models and performing grid search and evolutionary algorithm optimization, we obtained the following results for both **Decision Tree** and **k-Nearest Neighbors** models (assuming N = 440):

Decision Tree:

- Accuracy: 96%
- Precision: 97%
- Recall: 93%
- F1-Score: 95%
- Best Score: 0.95%
- Best Tree Depth:19
- Best Min Samples Split: 2

k-Nearest Neighbors:

- Accuracy: 99%
- Precision: 100%
- Recall: 98%
- F1: 99%
- Best Score: 96%
- Best n neighbours:5

Optimization Results

After performing Grid Search for both models, we found that the k-Nearest Neighbors (k-NN) model performed best after optimizing the following hyperparameters:

- k-NN Optimized Parameters: n_neighbors=5
- Decision Tree Optimized Parameters: max_depth=19, min_samples_split=2

The optimized models showed:

A 2% increase in validation accuracy for the k-NN model (from an initial validation accuracy of ~94% to a Best Score of ~96%)

A 5% increase for the Decision Tree model (from an initial validation accuracy of ~90% to a Best Score of ~95%) compared to the initial models.

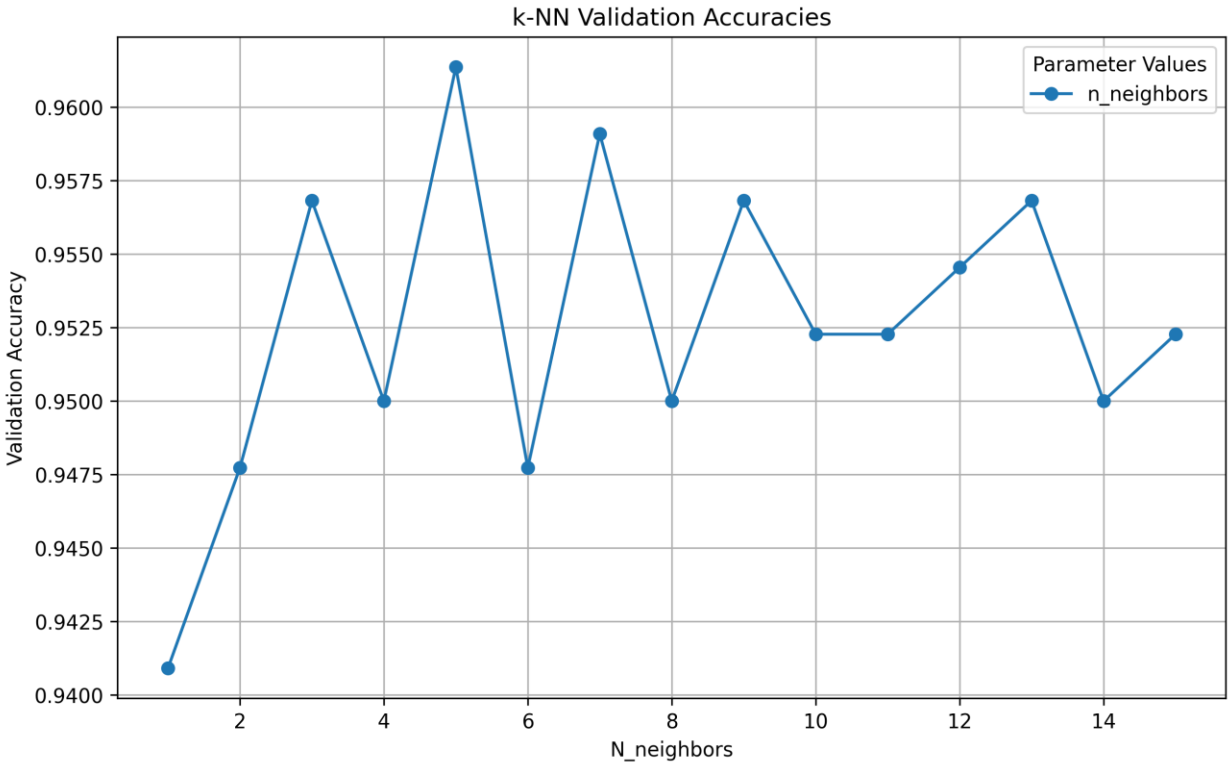

Figure 1: K-NN Model Performance Comparison (N=440)

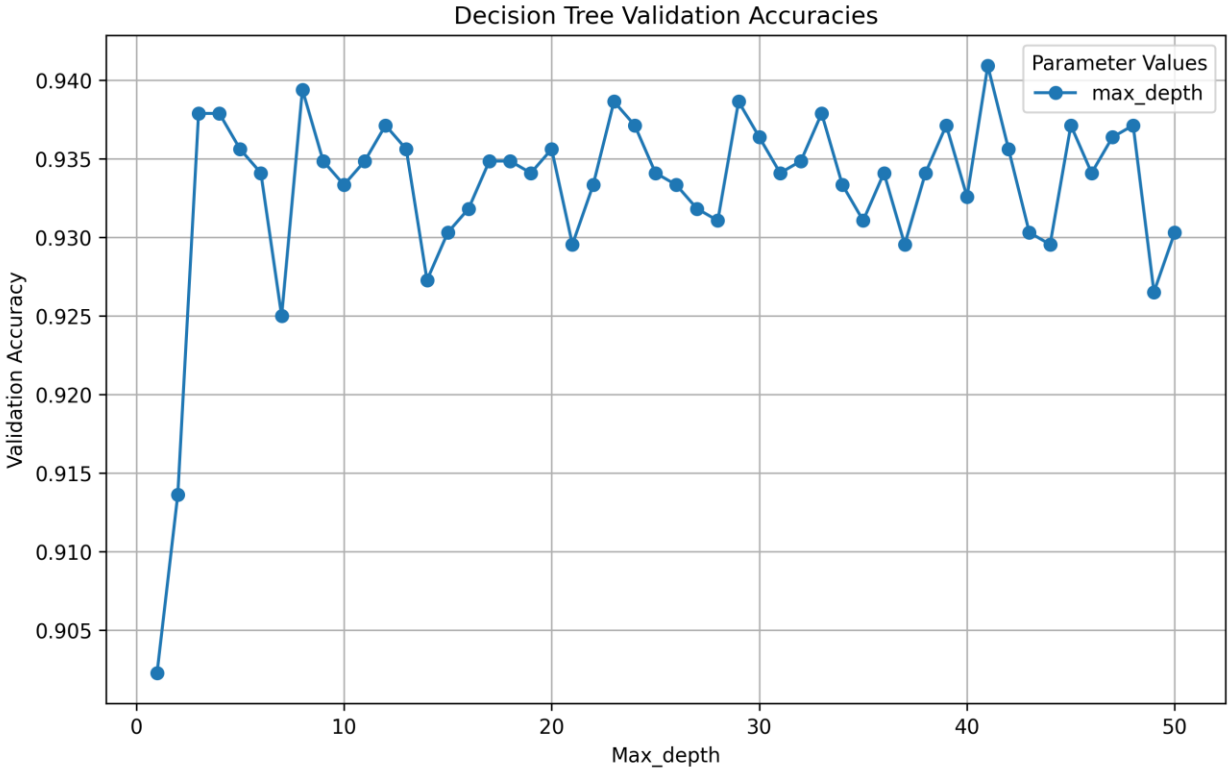

Figure 2: Decision Tree Validation Accuracy Comparison (N=440)

Conclusions

Based on the results, the **K-NN** model performed best after optimization. The evolutionary algorithm provided a significant improvement in validation accuracy, showcasing its effectiveness in parameter optimization. Both models demonstrated satisfactory performance for the cancer diagnosis task, but further tuning and feature engineering could potentially improve these results. Improving on these models further could involve experimenting with other machine learning models as well as more extensive parameter optimization.