Performance Comparison of SVM, Decision Tree, and Random Forest for Breast Cancer Classification

By: Deepak Vishwakarma (Student ID: 23035559)

I. Introduction

Breast cancer diagnosis is considered very important in effective treatment; hence, the need for developing accurate prediction models becomes very indispensable. The study employs comparing performance efficiency among three machine learning algorithms-Support Vector Machine (SVM), Decision Tree, and Random Forest-on a dataset pertaining to breast cancer. The dataset used in this study is fetched from the UCI Machine Learning Repository. The analysis tries to assess the effectiveness and efficiency of these models in differentiating benign and malignant tumors.

II. Data Overview and Preprocessing:

This dataset contains 699 samples and 11 attributes, including tumor characteristics and the corresponding target class:

Missing Value Treatment: Only one column named bare_nuclei had missing values, which were managed using the median method.

Feature Selection: The Sample_code_number column was deleted since it added no classification task value. The target Class variable was converted into binary values with 0 representing benign and 1 representing malignant. The dataset was then split into 80% and 20% training and testing subsets, respectively and scaled using StandardScaler to improve the model's performance.

III. Classification Models and Their Evaluation:

1. **Support Vector Machine:** SVM is most commonly used for classification task, here the model is trained using the RBF kernel function as its captures the non-linearity really well.
   The model is trained using all the features of the dataset while two features are used for the decision boundary for visualization.
   Evaluation: The models accuracy turned out to be 98% which is extremely well and a more clear decision boundary can be visualized below,

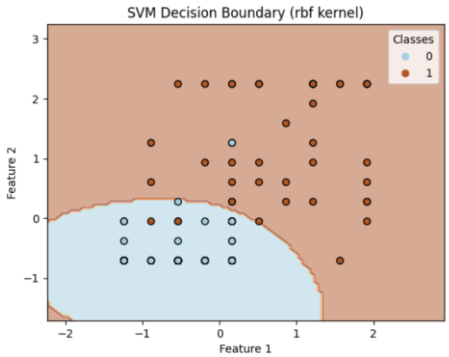

Fig 1. Decision Boundary for SVM

2. **Decision Tree and Random Forests:** Both the models her used GridSearchCV to optimize the models and determining the best parameters for the models.
   - For decision tree GridSearchCV was used to optimize max_depth, min_samples_split, and criterion. The best parameters came out as: max_depth: 6, min_samples_split: 2 and criterion: Gini

- For random forests GridSearchCV was used to optimize n_estimators, max_depth, min_samples_split, and criterion. The best parameters came out as: n_estimators: 100, max_depth: 6 and, criterion: Entropy.

Random Forrest performs better than Decision tree with an accuracy of 0.97 over 0.94. While analysing the feature importance of both we get the following result,

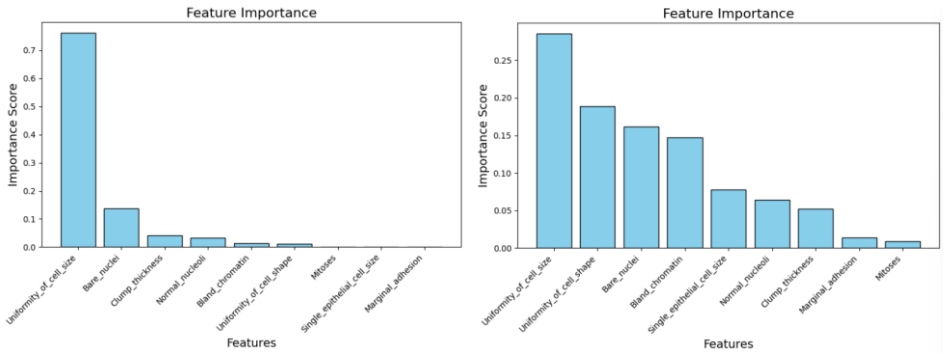

Fig 2 & 3. Feature Importance of Decision Tree (Left) and Random Forests (Right)

IV. Analysis of the Result:

- SVM has the highest accuracy of 98%, showing its capacity of finding complex relationships within the data, be it linear or non-linear, especially with the Radial Basis Function (RBF) kernel. This kernel allows SVM to define flexible decision boundaries, which are mainly effective for datasets that contain complex internal structures, such as tumor classification. Furthermore, the margin-maximization approach provides more generalization and minor overfitting.
- As far as accuracy is concerned, the Decision Tree model scores only a very low 94%, less than both SVM and Random Forest. The skewed feature importance is one of the major limitations. A Decision Tree usually puts consider attention to only a few features that yield the first substantial splits in the tree, leading to a very skewed distribution of importances. Due to this skewness, the model tends to over-rely on a particular few main features, which may harm other dependencies, especially in the presence of noisy features or features which are not generalizable.
- A score of 97% in accuracy was reflected in the Random Forest, which was close to SVM's performance. This form of ensemble learning, which combines many decision trees that will be trained on random parts of the data and features, affects variance by reducing the overfitting process. Besides, unlike Decision Trees, which give biased importance to only a few salient features, Random Forest balances the importance across all features.

V. Conclusion:

SVM, Decision Tree, and Random Forest were compared with respect to the cancers' classification. Among them, SVM was found to achieve a higher accuracy score, which is evidence of its aptness for modeling complicated and non-linear relationships as it creates flexible decision boundaries. Random Forest, the robust ensemble of many classifiers, offers an overall robust performance with its more uniform distribution of the importance of features, therefore having lower overfitting problems than Decision Trees. Decision Trees are simple and efficient to compute but are limited in accuracy due to skewed feature importance; hence, these methods are less robust.

References:

1. Khan, K., Rehman, S.U., Aziz, K., Fong, S. and Sarasvady, S., 2014, February. DBSCAN: Past, present and future. In *The fifth international conference on the applications of digital information and web technologies (ICADIWT 2014)* (pp. 232-238). IEEE.
2. Deng, D., 2020, September. DBSCAN clustering algorithm based on density. In *2020 7th international forum on electrical engineering and automation (IFEEA)* (pp. 949-953). IEEE.
3. Burkardt, J., 2009. K-means clustering. *Virginia Tech, Advanced Research Computing, Interdisciplinary Center for Applied Mathematics*.

GitHub Link: https://github.com/Deepakvishwakarma1/Data-Mining-Assignment-.git