Cancer Treatment Outcomes Prediction with Machine Learning: A Model Comparison

Yashvi Singh
2023MT60677

Project Overview :

Cancer treatment planning remains a complex challenge, where the selection of the most effective treatment regimen is critical to patient's outcomes. This project aims to predict the effectiveness of cancer treatments by evaluating multiple classification models. The model includes - Logistic Regression, Naive Bayes, Support Vector Classifier(SVC) and Random Forest. The goal is to identify the best-performing model for predicting breast cancer tumor malignancy with machine learning.

Objectives :

The primary objectives of this project are as follows:

1. To preprocess and prepare the dataset for machine learning, ensuring data quality and model performance.
2. To evaluate the performance of four distinct machine learning models—Logistic Regression, Naive Bayes, Support Vector Classifier (SVC), and Random Forest—on a cancer treatment dataset.
3. To evaluate each model's performance in predicting tumor classification (malignant or benign) based on tumor-related features, including radius, texture, perimeter, area, and other tumor properties.
4. To identify the model that delivers the most accurate and reliable predictions for treatment outcomes.

DataSet Description :

The dataset used in this project is derived from a breast cancer diagnosis dataset, containing both numerical and categorical features related to tumor's characteristics, and patient diagnosis.

- Tumor features: Radius, texture, perimeter area, smoothness, compactness, concavity, symmetry and fractal dimension.
- Diagnosis: Target variable indicates whether the tumor is malignant/cancerous(M) or benign/non-cancerous(B)

Approach and Methodology :

1. Data Preprocessing:
    a. Handling missing values by using mean imputation for numerical variables and mode imputation for categorical variables were used to address missing data.
    b. Categorical Encoding: Using LabelEncoder the column 'diagnosis' was converted from string to binary format.

    c. Normalization and Scaling: Features were standardized using StandardScaler to ensure model stability during training.
2. Model Development: Four classification models were applied to predict tumor classification (Benign or Malignant)
    a. Logistic Regression: A linear model for predicting the probability of malignancy based on feature relationships.
    b. Naive Bayes: A probabilistic model assuming feature independence to predict tumor classification.
    c. Support Vector Classifier (SVC): A robust model with both linear and non-linear kernels, designed to maximize the margin between classes.
    d. Random Forest: An ensemble method combining decision trees to improve accuracy and reduce overfitting, ideal for capturing complex feature interactions.
3. Model Evaluation:
    a. Data Split: The dataset is split into 25% testing sets and 75% training sets for performance evaluation.
    b. Performance Metrics: Models were evaluated based on accuracy, precision, recall, F1 score, and ROC-AUC.
    c. Hyperparameter Tuning: Grid search was used to optimize hyperparameters for each model.

Results:
Upon evaluation following insights were gained:

1. Logistic Regression: While effective for binary classification, it showed limitations in capturing complex, non-linear relationships in the data.
2. Naive Bayes: Performed well in terms of computational efficiency but lagged behind in accuracy and recall.
3. Support Vector Classifier (SVC): Delivered good results, especially with non-linear kernels, but required careful tuning of hyperparameters.
4. Random Forest: Outperformed all other models in terms of accuracy, precision, recall, and F1 score. It was the most reliable in predicting treatment outcomes, benefiting from its ensemble nature and ability to model complex interactions.

Conclusion: The machine learning models tested in this project provided valuable insights into cancer treatment prediction. **Random Forest** emerged as the most effective model, offering the highest performance in terms of predictive accuracy and handling complex relationships between features. This model's robustness, flexibility, and ability to manage overfitting make it the ideal choice for predicting treatment outcomes in cancer patients.
This project demonstrates that machine learning can be a powerful tool for personalizing cancer treatment, allowing healthcare professionals to make more informed decisions. The successful

application of Random Forest can lead to better treatment planning, ultimately improving patient outcomes.

Future Work:

- Model Deployment: The next step would be to deploy the Random Forest model into a clinical environment to assist oncologists in real-time treatment decision-making.
- Feature Expansion: Incorporating additional features such as genetic data, patient lifestyle, and treatment adherence could further improve prediction accuracy.
- Model Enhancement: Exploring deep learning models and neural networks could potentially enhance predictive performance and uncover more complex patterns in the data.

References:

Zuluaga, Juan Pablo. (2019). Breast Cancer Diagnosis Using Machine Learning Techniques. 10.13140/RG.2.2.36122.52168.