The UNIVERSITY of OKLAHOMA

Biomanufacturing Data Analytic/ML Research Project

Project Title: Machine Learning Analysis of Dorothea Dataset
Name: PAUL OKAFOR
Sooner ID: 113585670

Professor's Name: Talayeh Razzaghi (Professor of Industrial and Systems Engineering)
Date: December 06, 2023

1. Project Overview and Data Description

This project employs machine learning techniques to classify chemical compounds from the Dorothea dataset as either active or inactive based on structural features. The dataset is part of the NIPS 2003 feature selection challenge and plays a vital role in drug discovery research. The approach involves data preprocessing, model building, and evaluation, including Logistic Regression, Support Vector Machines, and Decision Tree. The project aims to optimize model performance and assess feature significance for prediction.

The Dorothea dataset, a component of drug discovery, involves classifying chemical compounds based on structural features. It contains 1950 instances and 100,000 features, with a mix of real and probe variables. Active compounds are mapped to +1, and inactive compounds to -1. The data are split into training, validation, and test sets. Notably, the dataset includes distractor features with no predictive power. This project utilizes this dataset to develop and evaluate machine-learning models for drug discovery applications.

Table 1.1. Summary of Training and Validation Datasets

Dataset	Total Instances	Active Compounds	Inactive Compounds
Training Set	800	78	722
Validation Set	350	34	316

2. Data Preprocessing

a. Data Cleaning: Check for missing values (Task 1)

The data preprocessing for the Dorothea dataset revealed no missing values, ensuring data completeness.

b. Data Transformation (Standardization) (Task 2)

Regarding data transformation, standardization was deemed unnecessary due to the dataset's binary nature, consisting of 0s and 1s. These values represent the absence or presence of features rather than quantitative measurements, making standardization irrelevant. Furthermore, as the dataset is sparse, with the majority of values being zeros, standardization would disrupt the sparsity, which is a critical characteristic of the data. This approach was taken to preserve the dataset's integrity and unique structure.

c. Data Imbalance

The Dorothea dataset exhibits a significant imbalance between active and inactive compounds. The chart below indicates a greater number of inactive compounds (-1) compared to active compounds (1). Specifically, the training set shows a substantial majority of inactive instances, while the test set has a relatively smaller, yet still significant, imbalance.

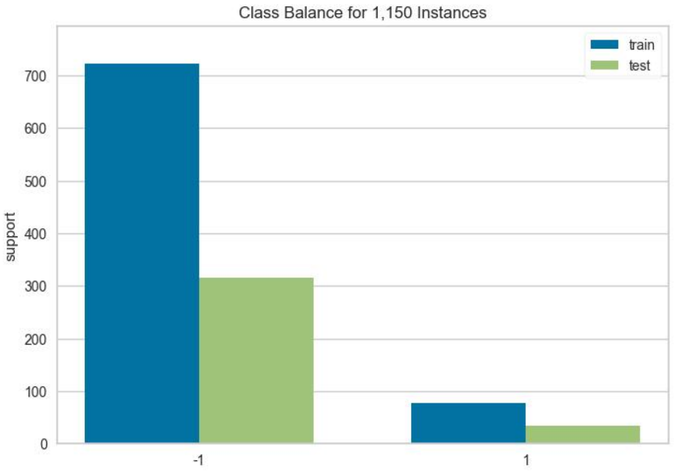

Figure 1.1. Class Imbalance

To tackle the class imbalance evident in the Dorothea dataset, the project adopts Stratified K-fold cross-validation during the model building phase. This technique enhances the model's ability to generalize by ensuring each fold reflects the original distribution of active and inactive compounds, thus preserving the dataset's imbalance ratio while allowing for an equitable learning process.

3. Model Building and Evaluation (Task 3 and 4)

In the model building phase, three distinct strategies were executed to refine the predictive capabilities of our models on the Dorothea dataset. Initially, we employed a straightforward 10-fold cross-validation technique to train three different classification algorithms, focusing on baseline performance without the influence of hyperparameter tuning. Subsequently, we integrated hyperparameter optimization into the 10-fold cross-validation process to enhance the model's predictive accuracy. Post optimization, models were retrained on the full training set.

Lastly, feature selection was applied to isolate the most impactful predictors, followed by a final retraining of the models, enabling us to scrutinize the influence of selected features on model performance.

Table 1.2. Characteristics of the selected Classification Algorithms

Classification Algorithms	**Characteristics**
Logistic Regression	- Simple to understand and implement for binary data - Useful to find relationships between features
Support Vector Machine	- Effective with high dimensional data - Relatively computationally cheap
Decision Tree	- Breakdown of complex data into smaller sections - Simple interpretability

3.1. Straightforward 10-fold Cross-validation Technique (Task 3)

The initial results from the straightforward 10-fold cross-validation model building offer valuable insights into the performance of different algorithms on the Dorothea dataset. Logistic Regression, with the highest accuracy, appears to manage the balance between recall and specificity effectively, indicating its potential reliability for the dataset. However, its precision and recall for positive cases (active compounds) are modest. Both Support Vector Machine models with RBF and polynomial kernels show perfect specificity but fail to identify any active compounds, as reflected by zero precision and recall. This could indicate an overfitting to the majority class. In contrast, the Decision Tree provides a more balanced performance with considerable precision and recall, suggesting a better model fit than the SVMs. The G-Mean and ROC AUC values reinforce these interpretations, with Logistic Regression and Decision Tree models showing a better balance between sensitivity and specificity compared to SVM models.

a. Logistics Regression Model

The Logistic Regression model exhibits decent performance with an ROC AUC of 80.09%, indicating good discriminative ability (figure 1.2). However, the Precision-Recall curve suggests room for improvement in precision and recall, especially given the low Average Precision of 38.52%. The confusion matrix shows a higher tendency for true negatives, with moderate false negatives, suggesting some challenges in accurately identifying true positives.

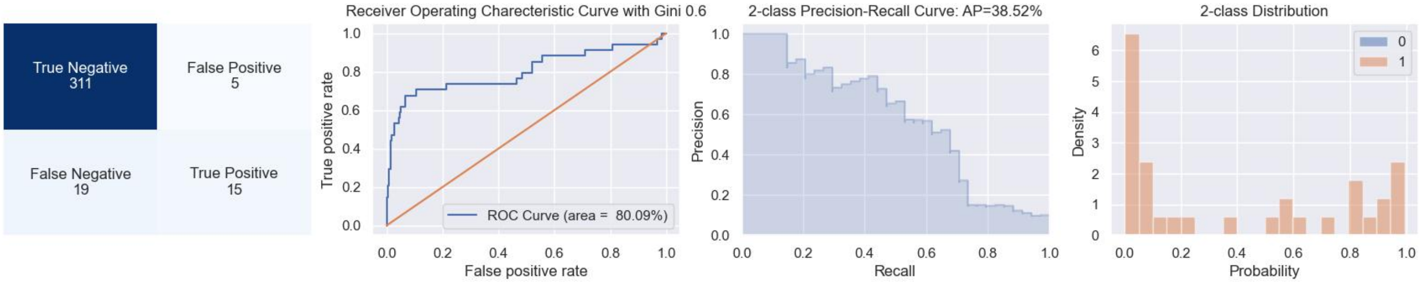

Figure 1.2. Logistic Regression Model

b. Support Vector Machines (rbf and poly kernels)

The SVM model with RBF kernel perfectly identifies all negative cases (true negatives) but fails to recognize any positive cases (true positives), as indicated by the confusion matrix (figure 1.3). Despite a high ROC AUC score of 91.62%, the model's inability to detect any true positives points to a significant issue with recall, necessitating a re-evaluation of model parameters or the approach to handling imbalanced data (figure 1.3).

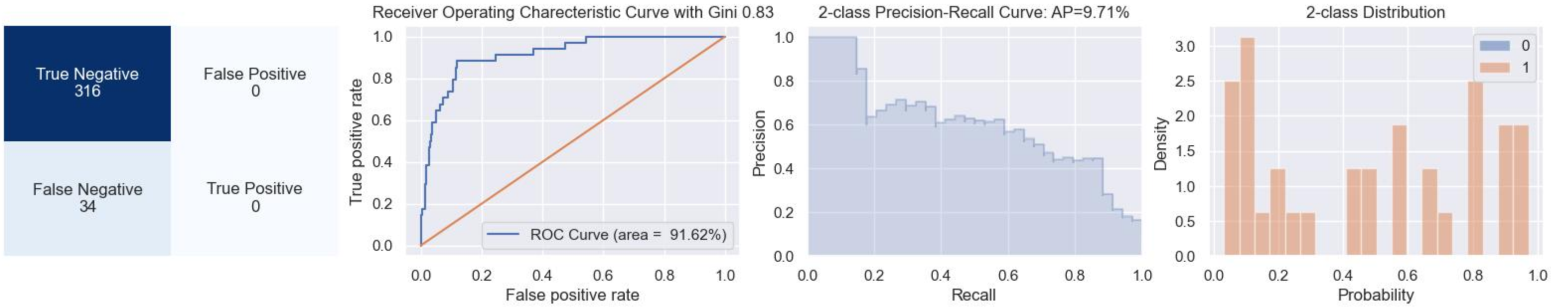

Figure 1.3. Support vector classifier (with rbf kernel)

The SVM with a polynomial kernel displays a high ROC AUC of 91.70%, indicating potential for distinguishing between classes (figure 1.4), yet the confusion matrix reveals it did not identify any positive cases. The Precision-Recall curve, with an Average Precision of 9.71%, confirms the model's difficulty in classifying positive instances, suggesting a need to adjust the kernel or address the data imbalance more effectively (figure 1.4).

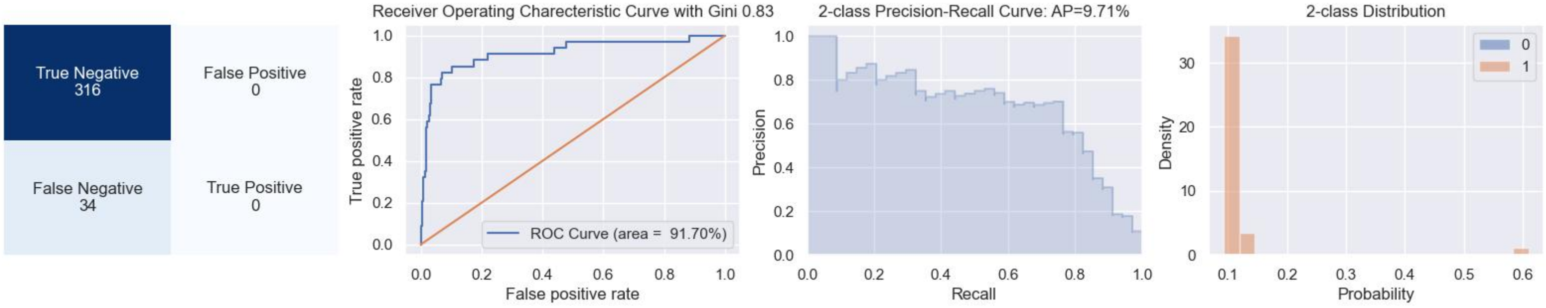

Figure 1.4. Support vector classifier (with poly kernel)

c. Decision Tree Model

The Decision Tree model shows a balanced trade-off between sensitivity and specificity, with an ROC AUC of 77.35% (figure 1.5). It successfully identifies some true positives while maintaining a good number of true negatives but with a modest false positive rate. The Precision-Recall value indicates that precision is relatively balanced with recall, evidenced by an Average

Precision of 39.65%. This model seems to be more effective in recognizing positive cases compared to the previous SVM models, suggesting better handling of the imbalanced dataset (figure 1.5).

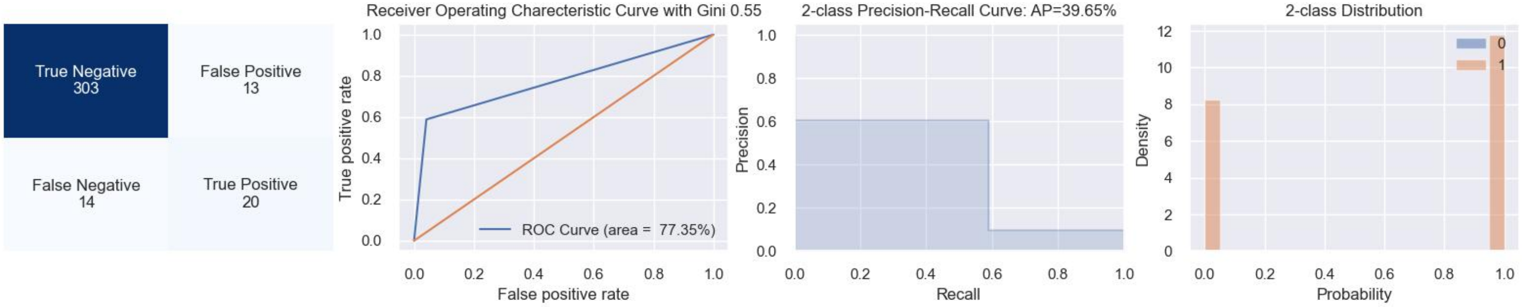

Fig 1.5. Decision Tree Model

Here is the summary of the models.

Table 1.3. Model Summary

Model	Accuracy (%)	Precision	Sensitivity (Recall)	Specificity	G-Mean	ROC AUC
	%	%	%	%	%	%
Logistic Regression	93.14	75.00	44.12	98.42	65.89	80.09
SVM with RBF Kernel	90.29	0.00	0.00	100.00	0.00	91.62
SVM with Poly Kernel	90.29	0.00	0.00	100.00	0.00	91.70
Decision Tree	92.29	60.61	58.82	95.89	75.10	77.35

3.2. Model Building with Hyperparameter Optimization (Task 3)

Hyperparameter optimization via Optuna provided a nuanced view of model performance. For Logistic Regression, a slight decrease in accuracy and precision post-optimization indicates a trade-off for a higher ROC AUC, suggesting improved balance in classification ability. The optimized SVM model with an RBF kernel showed a remarkable precision improvement, signifying a higher trustworthiness in its positive predictions, albeit at the cost of recall (table 1.4). This precision uptick did not translate into an improved F-score, due to the low recall. The Decision Tree's metrics experienced marginal adjustments, with a slight drop in performance across most metrics, indicating the complexity of finding an optimal balance of hyperparameters for this model. Overall, while hyperparameter optimization improved certain aspects of model performance, computational constraints limited the extent of the optimization, which reflects in the modest performance gains.

Here are the optimized results in a table:

Table 1.4. Model Summary after Hyperparameter Optimization

Model	**Accuracy (%)**	**Precision**	**Sensitivity (Recall)**	**Specificity**	**G-Mean**	**ROC AUC**
	%	**%**	**%**	**%**	**%**	**%**
Logistic Regression	92.57	70.00	41.18	98.10	63.56	83.82
SVM (RBF Kernel)	92.29	88.89	23.53	99.68	48.43	92.98
Decision Tree	92.00	65.00	38.24	97.78	61.15	82.28

3.2.1. Model Building after Feature Selection (with optimized Hyperparameters) (Task 4)

Feature selection via the variance threshold method has demonstrated its efficacy in enhancing model performance. By eliminating less informative features, the Logistic Regression model maintained its accuracy while slightly improving precision. The Support Vector Machine with RBF kernel witnessed the most notable enhancement, with significant increases in precision and F-score, reflecting a substantial improvement in the model's ability to make correct predictions on active compounds (table 1.5). The Decision Tree model's performance metrics remained relatively stable, indicating its robustness to feature selection.

The table below summarizes the results after feature selection:

Table 1.5. Model Summary after Feature Selection

Model	**Accuracy (%)**	**Precision**	**Sensitivity (Recall)**	**Specificity**	**G-Mean**	**ROC AUC**
	%	**%**	**%**	**%**	**%**	**%**
Logistic Regression	93.14	77.78	41.18	98.73	63.76	82.30
SVM (RBF Kernel)	94.00	93.33	41.18	99.68	64.07	85.94
Decision Tree	93.14	75.00	44.12	98.42	65.89	80.30

This step of feature selection underscores the importance of having a parsimonious model, which not only simplifies the model but can also lead to better generalization and performance.

4. Conclusion (Task 5)

The exploration of the Dorothea dataset through various machine-learning models has provided a comprehensive understanding of how different algorithms can be tuned and optimized for better predictive performance. In conclusion, the models' performance after feature selection indicates that the Support Vector Machine with RBF kernel stands out due to its high precision and F-score, showing it to be most adept at making accurate predictions when it classifies a compound as active. However, its recall remains modest, indicating room for improvement in identifying all active compounds.