CPSC 393: Assignment 1 Report
Jaime Song

Analysis

The dataset consisted of multiple predictor variables and one binary target variable. The predictors were mostly numerical so scaling was necessary to ensure they were on the same scale for modeling. As part of preprocessing, I used standard scaling, which resales each feature to have mean 0 and standard deviation 1. This helps algorithms like SVM and logistic regression converge more effectively and ensures distance based models like KNN are not biased towards larger valued features.

I examined summary statistics and explored relationships among features. Correlation analysis showed that some predictors were moderately correlated, but none were so highly correlated that they needed to be removed. The class distribution in the target variable was fairly balanced, which meant accuracy would be a reasonable performance metric. No major data cleaning or joining was necessary besides handling scaling and splitting into training and testing sets.

Methodology

Built three models using sklearn Pipelines

Support Vector Machine (SVM)

- Tuned regularization parameter C
- Tuned kernel (linear vs RBF)
- Best parameters: C = 25, kernel = linear
- A higher C gave the model for flexibility to classify difficult points correctly, while the linear kernel worked best given the feature relationships

Logistic Regression

- Tuned regularization parameter C
- Best parameter: C = 5
- A moderate C provided a good balance between underfitting (too much regularization) and overfitting (too little regularization)
    - Logistic regression is interpretable, with coefficients representing feature contributions

k-Nearest Neighbors

- Tuned number of neighbors, which controls how many neighbors vote on the class of a new data point
- Best parameter: n_neighbors = 7
- Too few neighbors can make the model too sensitive (high variance), while too many can make it too smooth (high bias)

Results

The models were evaluated on training accuracy, testing accuracy, ROC AUC, and confusion matrices

- Logistic regression achieved the highest test accuracy (77%) and the best ROC AUC (0.832), suggesting it generalizes best
- SVM performed with slightly lower accuracy (74.5%) but still strong ROC AUC (0.829)
- KNN had the highest training accuracy (80%) but the lowest test accuracy (72.5%) which suggests overfitting
- From confusion matrices, logistic regression reduced false negatives compared to the other models, which may be important if missing positive cases has higher cost

Best model: Logistic regression, as it balances accuracy, ROC AUC, and generalization performance

Reflection

Through this project, I learned how to structure machine learning workflows using Pipelines and GridSearchCV in sklearn. I gained hands-on experience with three different models (SVM, Logistic Regression, KNN) and how hyperparameters directly impact their performance. One challenge for me was understanding how hyperparameters actually control the model's behavior. At first, the terminology was confusing, but running experiments and comparing outputs helped solidify my understanding. If I did this assignment again I would spend more time on exploratory data analysis. For example, visualizing feature distributions and checking for nonlinear patterns, because this could guide model selection more effectively. In the future, I would also try more complex models and compare them with simpler models to see if they improve performance further. Overall, this project taught me both the practical coding skills to implement models and the critical thinking needed to evaluate them fairly.