**Applied Machine Learning**

**Assignment 2**

Submitted By

Mallidi Akhil Reddy

2024TM93056

# Problem Definition:

The goal of this assignment is to perform a classification analysis on two real-world datasets obtained from the UCI Machine Learning Repository:

**Banknote Authentication Dataset:** The goal here is to classify whether a banknote is authentic or not.

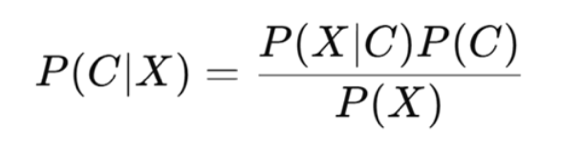
**Haberman's Survival Dataset:** The goal here is to predict patient survival status based on surgery data.

The following machine learning algorithms are employed for both tasks:

1. Naive Bayes
2. Logistic Regression
3. Support Vector Machine (SVM)
4. Random Forest

# Mathematical Details:

**Naive Bayes:** Naive Bayes is probabilistic classifier algorithm based on Bayes’ Theorem. It assumes independence among features.

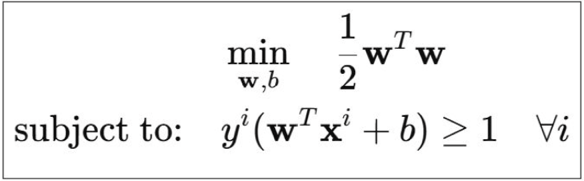


**Logistic Regression:** It’s a linear model algorithm for binary classification. In this Model parameters are estimated using maximum likelihood estimation.

A mathematical equation with numbers and symbols

AI-generated content may be incorrect.

**Support Vector Machine (SVM):** This algorithm Maximizes the margin between classes.

****

# Dataset Description:

**Banknote Authentication Dataset:**

* Number of features: 4 (variance, skewness, curtosis, entropy)
* Number of instances: 1372
* Number of classes: 2 (authentic, counterfeit)
* Preprocessing steps:

1. Scaling the features using StandardScaler to have zero mean and unit variance. This can help improve the performance of algorithms like Logistic Regression and SVM.
2. No missing values were found in the dataset.

**Haberman's Survival:**

* Number of features: 3 (age, year of operation, number of positive axillary nodes)
* Number of instances: 306
* Number of classes: 2 (survived 5 years or more, died within 5 years)
* Preprocessing steps:

1. Scaling the features using StandardScaler, as it is important because the features have different ranges, and scaling can prevent features with larger ranges.
2. No missing values were found.

# Experiments:

**Experimental Setup and Train-Test Split:**

For each dataset, I used an 80/20 train-test split. This means that 80% of the data was used for training the models, and 20% was held out for evaluating their performance.

**Hyperparameter Tuning:**

**Naive Bayes:** Naive Bayes generally doesn't have hyperparameters that require extensive tuning.

**Logistic Regression:**

* Hyperparameter: Regularization strength (C)
* Tuning method: Grid search
* Parameter range: C = [0.001, 0.01, 0.1, 1, 10, 100]
* Regularization type: l2

**SVM:**

* Hyperparameters: Kernel type, C (regularization), gamma (for RBF kernel)
* Tuning method: Grid search
* Parameter ranges:
  + Kernel: ['linear', 'rbf']
  + C: [0.1, 1, 10, 100]
  + gamma (for RBF): [0.001, 0.01, 0.1, 1]

# Results on Test Data:

*Metrics are rounded to 4 decimals*

**Banknote Authentication Dataset:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Macro F1-Score** | **Macro Precision** | **Macro Recall** |
| Naive Bayes | 0.8582 | 0.856 | 0.8571 | 0.8551 |
| Logistic Regression | 0.9855 | 0.9853 | 0.9841 | 0.9869 |
| Support Vector Machine (SVM) | 1.0 | 1.0 | 1.0 | 1.0 |

**Haberman Survival Dataset:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Macro F1-Score** | **Macro Precision** | **Macro Recall** |
| Naive Bayes | 0.7581 | 0.57 | 0.686 | 0.572 |
| Logistic Regression | 0.7581 | 0.5338 | 0.7147 | 0.5516 |
| Support Vector Machine (SVM) | 0.7258 | 0.5127 | 0.5772 | 0.5299 |

# Conclusion:

The classification of the Banknote Authentication dataset yielded remarkably high performance across all evaluated models. Support Vector Machine (SVM) achieved a perfect score with 100% accuracy, F1-Score, precision, and recall on the test data. Logistic Regression also demonstrated strong performance with an accuracy of 0.9855, a macro F1-Score of 0.9853, a macro precision of 0.9841, and a macro recall of 0.9869. Naive Bayes, while slightly lower, still provided a respectable accuracy of 0.8582, a macro F1-Score of 0.856, a macro precision of 0.8571, and a macro recall of 0.8551.

The classification of the Haberman's Survival dataset proved to be a more challenging task. All three models exhibited lower performance compared to the banknote dataset. Logistic Regression and Naive Bayes achieved the same accuracy of 0.7581. However, Logistic Regression showed a slightly better macro F1-Score (0.5338) compared to Naive Bayes (0.57), although Naive Bayes had a higher macro precision (0.686 vs 0.7147 for Logistic Regression) but a lower macro recall (0.572 vs 0.5516 for Logistic Regression). SVM performed slightly worse on this dataset with an accuracy of 0.7258 and a macro F1-Score of 0.5127.

# Future Work:

Incorporate feature selection or transformation like PCA to explore latent patterns.

Employing stratified k-fold cross-validation to ensure stable performance estimates across imbalanced data splits.