DEPARTMENT OF DECISION SCIENCE

FACULTY OF BUSINESS

UNIVERSITY OF MORATUWA

SEMESTER 04

DA2111 - Statistical and Machine Learning

Assignment

Wine Quality Classification Report

Table of Contents

1. Data Set Selection and Associated Task

- **Dataset**: Wine Quality (from UCI Machine Learning Repository)
- **Task**: Classification
- **Description**: This project involves predicting wine quality based on various chemical properties.
- **Objective**: Classifying wine quality can help producers maintain consistent quality and support quality control efforts.

2. Problem Definition

- **Target Variable**: quality (predicting high or low quality based on chemical features)
- **Predictive Features**:
    - Chemical characteristics (e.g., fixed acidity, volatile acidity, citric acid, density, pH, sulphates, alcohol).
- **Research Question**: "Can wine quality be accurately predicted using chemical properties, and which machine learning model offers the highest accuracy?"

3. Data Exploration and Preprocessing

- **Dataset Summary**:
  - The red wine dataset has 1,599 entries, and the white wine dataset has 4,898 entries.
  - Both datasets contain 13 columns, including 12 features and the target variable (quality).

- **Handling Missing Values(Handling NULL Values)**:
  - No missing values were detected, so preprocessing could proceed without imputation.

- **Outlier Detection and Removal**:
  - Applied Interquartile Range (IQR) method to remove extreme values in continuous variables. Outliers were identified and removed, improving data quality and model robustness.

- **Encoding Target Variable**:
  - Transformed quality into a binary classification (1 = high quality, 0 = low quality), using a threshold to distinguish between high and low quality wines.

- **Scaling**:
  - Applied StandardScaler to normalize features, enabling better performance in models sensitive to scale differences.

4. Exploratory Data Analysis (EDA)

- **Feature Distribution**:
  - Histograms were used to visualize distributions. Features like residual sugar and citric acid displayed right-skewed distributions, indicating a minority of wines with high values.
- **Quality Distribution**:
  - Observed imbalance between high and low quality, particularly in red wines, where low-quality samples dominate.
- **Correlation Analysis**:
  - Generated a heatmap to assess feature correlations.
    - Notable correlations:
      - Strong positive correlation between alcohol and quality.
      - Moderate negative correlation between volatile acidity and quality, suggesting lower acidity aligns with higher quality.

5. Model Selection

- **Chosen Models and Justification**:
  - **Logistic Regression**: Provides a simple baseline with linear decision boundaries.
  - **Decision Tree**: Capable of handling non-linear relationships and easily interpretable.
  - **Random Forest**: Ensemble model that aggregates multiple decision trees, enhancing accuracy and robustness.
  - **Support Vector Machine (SVM)**: Effective for binary classification, useful for capturing non-linear relationships with kernel functions.

6. Model Training and Evaluation

- **Data Splitting**:
    - Data split into training (80%) and testing (20%) sets for validation.
- **Class Imbalance Handling**:
    - Given the imbalance between high- and low-quality wines, I applied class balancing techniques to improve model performance:
        - **Class Weights**: For Logistic Regression, Decision Tree, and SVM, I used class_weight='balanced' to assign higher weights to the minority class, ensuring the model gives equal attention to both classes during training.
        - **Random Forest**: The class_weight='balanced' parameter was also set, which adjusted weights dynamically based on class frequency, reducing bias toward the majority class.
- **Evaluation Metrics**:
    - For binary classification of quality, used accuracy, precision, recall, F1-score, and confusion matrices.
- **Performance**:
    - **Logistic Regression**: 71% accuracy; served as a strong baseline.
    - **Decision Tree**: Achieved 82% accuracy with good interpretability.
    - **Random Forest**: Reached 88% accuracy, excelling due to ensemble averaging.
    - **SVM**: Performed at 74% accuracy; showed potential with further feature engineering.
- **Overfitting Check**:
    - Compared train and test accuracy; Random Forest showed slight overfitting, whereas other models maintained balanced performance.

7. Hyperparameter Tuning

- **Random Forest Tuning**:
  - Parameters tuned using GridSearchCV:
    - n_estimators: [50, 100, 200]
    - max_depth: [None, 10, 20, 30]
  - Optimal parameters: n_estimators=100, max_depth=30
- **Results After Tuning**:
  - Accuracy of Random Forest improved to 88.36% on the test set, indicating enhanced performance and stability.

8. Comparison and Conclusion

- **Model Comparison**:
  - **Random Forest (tuned)**: Best performance (88.36%) with minimal overfitting.
  - **Decision Tree**: Performed well (82% accuracy) but was less robust than Random Forest.
  - **SVM**: Moderate accuracy; potentially useful with refined feature selection.
  - **Logistic Regression**: Effective baseline, though limited by linear assumptions.
- **Conclusion**:
  - Random Forest proved to be the most effective model for predicting wine quality, highlighting the advantage of ensemble learning for complex data.

9. Challenges and How They Were Addressed

- **Outlier Detection**:
  - Challenge: High variability among chemical properties.
  - Solution: IQR method effectively removed outliers, stabilizing model performance.
- **Class Imbalance**:
  - Challenge: More low-quality samples, especially in red wine.
  - Solution: Evaluated metrics carefully and could consider resampling if further balancing is needed.
- **Hyperparameter Tuning**:
  - Challenge: Time-intensive for Random Forest.
  - Solution: GridSearchCV enabled efficient tuning, balancing accuracy and training time.

10. Final Thoughts

- **Best Model Selection**:
  - The tuned Random Forest model was the optimal choice, balancing accuracy and generalizability for predicting wine quality.
- **Insights Gained**:
  - Comprehensive preprocessing, scaling, and iterative tuning are essential for enhancing classification model performance.
  - Ensemble models like Random Forest significantly improve accuracy in complex datasets.