Executive Summary Report

Hayden Cho

This report will analyze the findings from the two projects using datasets from the UCI Machine Learning Repository. The first project analyzes wine quality based on different properties, while the second predicts student academic performance using demographic and educational factors. Both projects demonstrate the versatility of machine learning techniques across different topics.

Methodology

Wine Quality Analysis

- **Dataset**: Portuguese red wine dataset (1,599 samples, 11 physicochemical features)
- **Models Implemented**:
    - K-Nearest Neighbors (KNN)
    - Multi-Layer Perceptron (MLP)
- **Approach**:
    - Binary classification (quality ≥ 6 as "good")
    - Regression for numerical quality prediction
    - Standardization and cross-validation

Student Performance Analysis

- **Dataset**: 395 student records with 33 features
- **Models Implemented**:
    - Linear Regression
    - Naive Bayes
- **Approach**:
    - Regression for grade prediction
    - Classification for pass/fail prediction
    - One-hot encoding and feature scaling

Results

Wine Quality Project

Model	Task	Performance Metric	Score
KNN	Classification	Accuracy	0.75
MLP	Classification	Accuracy	0.82

KNN	Regression	RMSE	0.68
MLP	Regression	RMSE	0.61

Key Findings:

- Alcohol content, volatile acidity, and sulphates emerged as strongest predictors
- MLP outperformed KNN in both tasks, suggesting non-linear relationships

Student Performance Project

Model	Task	Performance Metric	Score
Linear Regression	Regression	RMSE	2.15
Ridge Regression	Regression	RMSE	2.08
Naive Bayes	Classification	Accuracy	0.83

Key Findings:

- Previous grades (G1, G2) were strongest predictors
- Study time, absences, and parental education significantly influenced outcomes

Comparison with Prior Work

Wine Quality Literature Comparison

- **Cortez et al. (2009)**: Achieved 80% accuracy with SVM
- **Our MLP Implementation**: 82% accuracy
- **Analysis**: The neural network approach showed slight improvement, confirming the effectiveness of deep learning for complex chemical relationships
- **Differences**: Better feature importance visualization and more comprehensive evaluation metrics

Student Performance Literature Comparison

- **Cortez & Silva (2008)**: Similar RMSE (~2.1) for grade prediction
- **Our Implementation**: RMSE of 2.08 with ridge regression
- **Analysis**: Comparable performance with improved pipeline architecture
- **Advances**: Better preprocessing, modern cross-validation, comprehensive feature analysis

Strengths and Limitations

Wine Quality Analysis

Strengths:

- Different model comparison
- Both classification and regression approaches

Limitations:

- Class imbalance issues
- Limited to only specific features
- No group methods explored

Student Performance Analysis

Strengths:

- Good models for educational settings
- Cross-validated performance

Limitations:

- Small dataset size (395 students)
- Single subject focus (Mathematics)
- Heavy dependence on previous grades

Real-World Applications

Wine Quality

- Automated quality control in wineries
- Optimized production
- Predict what consumers may buy
- Cost reduction in quality assessment

Student Performance

- Catch bad grades early
- Can see where to improve education
- Can help specific students who struggle in specific subjects

Conclusions

Both projects successfully demonstrated the application of machine learning techniques to real-world problems. The wine quality analysis achieved similar performance with neural networks showing particular strength in capturing complex chemical relationships. The student

performance analysis provided interpretable models suitable for educational decision-making, with accuracy comparable to published literature.

These implementations showcase the different uses of scikit-learn tools and the importance of proper model selection, and evaluation in developing practical machine learning solutions. Future work could explore group methods and larger datasets for both topics.

References

1. Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. Decision Support Systems, 47(4), 547-553.

2. Cortez, P., & Silva, A. (2008). Using Data Mining to Predict Secondary School Student Performance. Proceedings of 5th Future Business Technology Conference, Porto, 5-12.

3. UCI Machine Learning Repository. Wine Quality Dataset. https://archive.ics.uci.edu/dataset/186/wine+quality

4. UCI Machine Learning Repository. Student Performance Dataset. https://archive.ics.uci.edu/dataset/320/student+performance