Mini Project

- Jared Pek (jarpek-4@student.ltu.se)
- Randall Chiang (ranchi-4@student.ltu.se)
- Joel Tham (joetha-4@student.ltu.se)

Video Presentation

https://drive.google.com/file/d/1K4QQObYQOxo2ydEmPbwJWb3g5aPRATir/view?usp=sharing

Introduction

Our mini project will focus on these UCI machine learning datasets:

1. Abalone
2. Bank Marketing
3. Car
4. Mushrooms
5. Wine Quality

The models we will be using to train on each dataset are the unsupervised K-Means clustering algorithm and the supervised logistic regression classification model.

Results

Car:

The Car Evaluation dataset was analyzed using both unsupervised and supervised learning methods. The unsupervised learning model, K-Means Clustering, produced an Adjusted Rand Index (ARI) of **0.0007**. These results indicate that the clusters formed by K-Means do not align well with the true class labels, and the clustering quality is poor. This is likely due to the categorical nature of the dataset and the limitations of using Euclidean distance in K-Means after one-hot encoding. The dataset does not exhibit natural clusters, making K-Means unsuitable for this task.

In contrast, the supervised learning model, Logistic Regression, achieved an overall accuracy of **92.87%**, demonstrating strong performance. The model effectively classified the majority classes (good and unacc), achieving high precision and recall for these categories. The minority classes (acc and vgood) were more challenging to classify, with lower precision and recall due to their smaller sample sizes and potential feature overlap. For example, the acc class achieved a precision of **65%** and a recall of **68.4%**, reflecting some misclassification issues. Despite this, the model maintained a balanced performance across all classes, as evidenced by a macro F1-score of **84.3%** and a weighted F1-score of **92.9%**.

Overall, Logistic Regression significantly outperformed K-Means Clustering for this dataset, demonstrating that the structured and categorical nature of the data is better suited to supervised learning methods. Addressing class imbalance or exploring clustering algorithms tailored for categorical data, such as K-Prototypes, could further improve performance.

Abalone:

The performance metrics of the logistic regression model and the k-means clustering model demonstrate differences in their ability to classify the data. Logistic regression achieved an overall accuracy of 27.51%, with a weighted average precision, recall, and F1-score of 0.20, 0.27, and 0.22, respectively. The model performed moderately well for Class 7, 8, and 10, achieving F1-scores of 0.45, 0.39, and 0.32, respectively. For Class 7, precision and recall were 0.37 and 0.58, reflecting its relatively stronger performance. However, the model struggled significantly with many classes such as Class 3, 4, 12, 14, and 15, where precision, recall, and F1-scores were all 0.0. This indicates a failure to predict these classes, likely due to class imbalance and limited representation of smaller classes.

In contrast, the k-means clustering model performed slightly worse, with an overall accuracy of 26.79%. While the model demonstrated an F1-score of 0.45 for Class 8 and some recognition for Class 10 (F1-score 0.33), it struggled significantly across most other classes. Precision, recall, and F1-scores for many smaller classes were 0.0, highlighting its inability to identify these categories. The reliance on an unsupervised approach, coupled with the presence of overlapping features, likely led to poor clustering and poor classification performance.

Overall, the logistic regression model performs marginally better than the k-means clustering model, particularly for Class 7, 8, and 10, where it demonstrates higher F1-scores. However, both models exhibit severe shortcomings due to class imbalance, linear separability issues, and overlapping features. Improving the models could involve addressing class imbalance with oversampling or undersampling, engineering better features, and exploring non-linear classifiers such as decision trees or ensemble methods. The current results highlight the limitations of both approaches for this dataset and the need for further optimization.

Wine Quality:

The performance metrics of the logistic regression model and the k-means clustering model indicate distinct outcomes in their ability to classify wine quality. Logistic regression achieved an overall accuracy of 53.62%, with a weighted average precision, recall, and F1-score of 0.50, 0.54, and 0.50, respectively. The model demonstrated moderate performance for Class 5 and 6, which represent the majority of the samples, achieving F1-scores of 0.57 and 0.60, respectively. For Class 5, the precision and recall were 0.54 and 0.61, while for Class 6, precision and recall were 0.54 and 0.68, respectively. However, the model struggled significantly with minority classes such as Class 4, 8, and 9, where precision, recall, and F1-scores were 0.0, indicating its inability to predict these underrepresented classes. For Class 3, precision was perfect at 1.0, but the recall was very low at 0.17, leading to a poor F1-score of 0.29.

In contrast, the k-means clustering model performed worse, with an overall accuracy of 47.46% and a weighted average precision, recall, and F1-score of 0.36, 0.47, and 0.41, respectively. While the model showed some ability to classify Class 5 and 6, achieving F1-scores of 0.44 and 0.59, respectively, it completely failed to predict Class 3, 4, 7, 8, and 9, with precision, recall, and F1-scores of 0.0 across these classes. The macro average F1-score of 0.15 highlights the model's inability to generalize well across all classes, further underscoring its poor performance on the minority labels.

Overall, the logistic regression model outperformed the k-means clustering model, particularly for the majority classes (Class 5 and 6), where it demonstrated better precision, recall, and F1-scores. However, both models struggled significantly with the minority classes, indicating the need for improvements in handling class imbalance. Addressing this issue through techniques like resampling, cost-sensitive learning, or employing ensemble methods could improve model performance. The results suggest that while logistic regression offers a slight advantage, both models remain limited in their ability to accurately classify all wine quality levels

Bank Marketing:

The performance metrics of the logistic regression model and the k-means clustering model reveal significant differences in their ability to classify the data. Logistic regression achieved a high overall accuracy of 93%, with a strong weighted average precision, recall, and F1-score of 0.94, 0.93, and 0.93, respectively. For class "0," the model demonstrated excellent performance with precision, recall, and F1-score around 0.96-0.97, indicating a robust ability to predict this majority class. However, the model struggled with class "1," achieving only 0.41 precision, 0.50 recall, and an F1-score of 0.45, reflecting challenges in handling the minority class due to potential class imbalance.

In contrast, the k-means clustering model performed poorly, with an overall accuracy of 65% and a substantially lower weighted average F1-score of 0.76. The macro average F1-score of 0.10 highlights the model's inability to effectively classify the data across all classes. While precision and recall for class "0" were moderate at 0.95 and 0.69, respectively, the performance for class "1" was abysmal, with a precision of 0.07, recall of 0.02, and F1-score of 0.03. Additionally, the presence of multiple unused clusters (classes "2" through "7") in the k-means results suggests poor clustering and a failure to capture meaningful groupings in the data.

Overall, the logistic regression model is superior for this dataset, especially when accuracy and handling of the dominant class are critical. However, it still requires improvement in predicting the minority class. The k-means clustering model is not a suitable alternative in this context due to its inability to effectively separate the classes and its reliance on an unsupervised approach that does not leverage label information. Addressing class imbalance and optimizing the logistic regression model might further enhance its performance.

Mushrooms:

The classification results of the Logistic Regression model and the K-Means clustering model present distinct performance patterns, revealing their relative strengths and weaknesses in this dataset. For Logistic Regression, the overall accuracy is 49%, with significant imbalances between precision, recall, and F1-score for the two classes. Class 0 exhibits high precision (0.97) but poor recall (0.19), suggesting that while the model is confident in its predictions for Class 0, it fails to identify most actual instances of this class. Conversely, Class 1 shows high recall (0.99) but low precision (0.42), indicating that most actual instances of Class 1 are captured, but many false positives are included. This imbalance is reflected in the weighted F1-score of 0.42 and suggests the model struggles with proper discrimination in a likely imbalanced dataset.

For the K-Means clustering model, the performance metrics are even less consistent. The overall accuracy is slightly lower at 45%, and the metrics for individual clusters show a stark disparity. Cluster 0 achieves moderate performance with an F1-score of 0.61, supported by reasonable precision (0.56) and recall (0.68). However, Cluster 1 has a high precision of 1.00 but extremely low recall (0.04), indicating that the cluster is highly specific but captures almost none of the relevant instances. Clusters 2, 3, 5, and 7 are entirely ineffective, with zero instances classified, resulting in F1-scores of 0. The macro-average metrics (precision, recall, and F1-score) are all very low, further highlighting the uneven performance across clusters.

In comparison, Logistic Regression demonstrates better capability in identifying patterns relevant to both classes, albeit with substantial room for improvement in balancing precision and recall. K-Means clustering, on the other hand, struggles to form meaningful clusters that align well with the dataset's underlying structure, as evidenced by its poor recall and F1-scores across most clusters. This suggests that the data may not exhibit the clear separability required for K-Means to perform effectively. Additionally, the presence of empty clusters in K-Means indicates potential challenges with the choice of the number of clusters or the initialization method.

Overall, the Logistic Regression model is preferable in this case due to its comparatively better performance, albeit limited by class imbalance and misclassification issues. The clustering model might require further optimization or a different clustering technique to achieve meaningful results.

Conclusion

Based on the models created from the 5 datasets, it can be determined that unsupervised models performed much poorer than the supervised ones. This is because k-means clustering is not designed for supervised learning. Logistic regression leverages labeled data to learn the relationship between input features and the binary target variable, optimizing a decision boundary that separates the two classes based on probabilistic modeling. In contrast, k-means is an unsupervised learning algorithm that groups data points into clusters based solely on feature similarity, without considering class labels.

As a result, k-means can misidentify clusters that do not align with the true classes, especially when the data distribution is complex or the clusters are not well-separated. Additionally, k-means assumes clusters are spherical and equally sized, which may not hold in real-world classification tasks. Logistic regression, by incorporating class labels and optimizing for classification accuracy, is typically more effective at binary classification tasks where labeled training data is available.

These are evident through the results of all 5 datasets that we have done in this project.