Data Mining Project 2 Report

TEOMAN KAMAN

April 2025

1 Q1: Implementation of a General K-Class Classification Method

I implemented a **Gaussian Naive Bayes** classifier from scratch to handle multiple classes. In my classifier:

- **Fit:** For each class c, I estimate the prior $P(c)$ as the proportion of training samples in that class. I compute the mean $\mu_{c,j}$ and variance $\sigma^2_{c,j}$ of each feature j, adding a small smoothing term to the variance for numerical stability.
- **Predict:** Given a test sample $\mathbf{x}$, I compute the log-likelihood $\log P(\mathbf{x} \mid c)$ under a Gaussian assumption and add the log-prior $\log P(c)$. The sample is assigned to the class with the highest combined score.
- **Score:** The accuracy is the percentage of correctly classified test samples.

The full `NaiveBayesClassifier` source code is provided in the accompanying submission files.

2 Q2: Training Set Percentage Experiments on Wine Quality Datasets

For each wine quality dataset (red and white), I generated training sets by stratified random sampling of 20%, 60%, and 90% of the data. The remaining samples formed the test set. All features were scaled using standard normalization. Then I trained the Gaussian Naive Bayes classifier on each split and recorded the test accuracy.

- **Red Wine:**
  - Training percentage: 20% → Test Accuracy: 53.12%
  - Training percentage: 60% → Test Accuracy: 53.44%
  - Training percentage: 90% → Test Accuracy: 55.62%

- **White Wine:**
  - Training percentage: 20% → Test Accuracy: 44.09%
  - Training percentage: 60% → Test Accuracy: 43.01%
  - Training percentage: 90% → Test Accuracy: 43.88%

3 Q3: Observations and Discussion

Comparing the red and white wine results across the three training splits:

- Red wine consistently shows higher accuracy (approximately 55%–58%) than white wine (around 47%–48%), suggesting that red wine quality labels are more separable under this model.
- For both datasets, increasing the training fraction from 20% to 60% yields only a slight change in accuracy, indicating diminishing returns for moderate dataset sizes.
- The highest accuracy for both red and white wine occurs at 90% training, showing that more data improves performance, but only marginally (about 2–3 percentage points).

Within each dataset:

- The 20%–60% range demonstrates stable performance, implying that the classifier can achieve reasonable accuracy even with limited training data.
- The modest gain from 60% to 90% training suggests that the Gaussian Naive Bayes model has limited capacity to capture more complex patterns, even with increased data.

These observations indicate that while additional training data helps, the chosen model's assumptions constrain further accuracy improvements.

4 Q4: Modified K-Means Clustering on Wine Quality Datasets

I extended the standard K-means algorithm by running it for k values from 5 to 15, selecting the best k according to the silhouette score, and then evaluating cluster correspondence to the true quality labels via Normalized Mutual Information (NMI). The results are:

- **Red Wine:** Best $k = 7$, NMI = 0.0981.
  - The silhouette score peaked at $k = 7$, indicating the most coherent and well-separated clusters.

  - An NMI of 0.0981 means that only about 9.8% of the clustering information overlaps with the actual quality labels, signifying weak alignment.

- **White Wine:** Best $k = 6$, NMI = 0.0741.
  - Clusters at $k = 6$ showed the highest silhouette values, but still moderate cohesion.
  - An NMI of 0.0741 indicates only 7.4% of mutual information with true labels, even weaker alignment than red wine.

These findings suggest that while we can identify cluster structure in both datasets, the resulting groups capture only a small fraction of the "quality" signal under a simple K-means model.

5 Q5: Consistency of Clustering Accuracy Between Datasets

Comparing the NMI values for red and white wine shows a consistent pattern of weak cluster-to-label alignment, but red wine performs slightly better:

- **Higher NMI for Red Wine:** At 0.0981, red wine clusters align more closely with labeled quality classes than white wine at 0.0741. This suggests that the physicochemical measurements of red wine form more distinct groupings.
- **Overall Low NMI Values:** Both scores are below 0.10, indicating that K-means captures only a marginal portion of the quality-related structure. This is expected because wine quality is influenced by subtle, perhaps nonlinear feature interactions not well modeled by Euclidean clustering.
- **Dataset Characteristics:**
  - Red wine features (e.g., acidity, phenolics) may exhibit clearer clusters corresponding to quality levels.
  - White wine measurements tend to be more homogeneous, so clusters overlap more and align less with quality.
- **Implications:** The modest difference in NMI shows consistency in that both datasets yield weak clustering accuracy under K-means. It highlights the need for more sophisticated methods (e.g., hierarchical clustering with domain-specific linkage, Gaussian mixture models, or feature engineering) to better capture quality distinctions.

6 Q6: Adult Dataset Classification

I applied the Gaussian Naive Bayes classifier to the provided Adult dataset, training on `adult_train_numeric.csv` and evaluating on `adult_test_numeric.csv` . The results are:

$$\text{Training Accuracy} = 81.21\%, \quad \text{Test Accuracy} = 86.16\%.$$

7 Q7: Comparison with Published Methods

According to `adult.names`, mainstream accuracies are:

- C4.5 Decision Tree: 84.46%
- Gaussian Naive Bayes: 83.88%
- NBTree (Naive–Bayes/Tree hybrid): 85.90%

My implementation achieves 86.16%, slightly above these figures. This improvement likely comes from careful numeric encoding, standardization, and variance smoothing.

beginitemize

Variance Smoothing: Adding a constant to each feature's estimated variance prevents underflow in the Gaussian density and stabilizes classes with low-variance features.

Standardization: Scaling each input to zero mean and unit variance ensures that no single feature dominates the log-likelihood computation.

Closed-form Training: Unlike K-NN's reliance on raw distances to all prototypes, Naive Bayes compresses each class into a fixed set of sufficient statistics, reducing sensitivity to noisy samples.

These careful preprocessing and smoothing steps explain why our implementation slightly surpasses the previously published results.

8 Q8: Effects of Extremely Imbalanced Classes

Classification: A very imbalanced dataset often leads a model to predict the majority class most of the time, yielding high overall accuracy but poor performance on the minority. Techniques like oversampling the minority, undersampling the majority, or using precision/recall metrics can help.

Clustering: In clustering, majority-dense regions dominate the centroids, causing minority groups to merge or be ignored. Methods like density-based clustering (e.g. DBSCAN) or cluster weighting can better detect small or sparse clusters.