D7041E - Miniproject

Eriksson, Elliot
lelrek-1@student.ltu.se

Larsson, Sune
sunlar-3@student.ltu.se

January 2025

1 Introduction

Click here to view the project on Github.

2 Addressed grading criteria

The grade criteria we focused on were for grade 3: *Develop 1 unsupervised and 1 supervised classification model for 5 datasets of your choice from 121 UCI datasets. Report accuracy results.*

3 Description of datasets used

The datasets we used are included in table 1. We utilised the Python script provided at the website to import the datasets.

Dataset with link
Iris
Wine
Breast Cancer Wisconsin (Diagnostic)
Optical Recognition of Handwritten Digits
Wine Quality

Table 1: Table over datasets used in the project (links to uci websites)

4 Description of the models used

Unsupervised learning: K-means Supervised learning: K Nearest Neighbours (KNN)

The K-means was implemented by utilising the sklearn library KMeans. The KNN was taken from lab 1 and modified to work with the datasets.

5 Description of the experimental methodology

We split the datasets into 3 groups train, validation, test, where the data was split $train = 60\%, validation = 20\%, test = 20\%$

We used LabelEncoder to label the target data, and StandardScaler to scale the data, both from sklearn.

For cross-validation we utilised the cross-validation we created in lab 1 for KNN and implemented a similar one for k-means but used the silhouette score to find the best k value.
The k values tested was $[2 - 9]$

6 Experimental result

Dataset	**Model**	**Best k**	**Validation Accuracy (%)**	**Test Accuracy (%)**
Iris	k-NN	4	100.00	96.67
	k-Means	2	50.00	70.00
Wine	k-NN	3	94.44	94.44
	k-Means	3	97.22	94.44
Breast Cancer	k-NN	4	94.74	94.74
	k-Means	2	89.47	79.82
Digits	k-NN	6	96.89	98.13
	k-Means	8	29.98	29.80
Heart Failure	k-NN	5	63.33	71.67
	k-Means	2	61.67	71.67
Wine Quality	k-NN	2	58.35	58.38
	k-Means	3	47.50	45.08

Table 2: Performance of k-NN and k-Means models on various datasets.

7 Conclusion

7.1 Conclusion of the Experimental Results

The experiments highlighted clear differences between k-NN and k-Means across various datasets:

k-NN performed exceptionally well on simpler datasets like Iris and Wine, with validation accuracies exceeding 90%. However, its performance dropped for complex datasets like Wine Quality and Heart Failure, where noise and overlapping classes posed challenges. k-Means, being a clustering algorithm, struggled in supervised classification. It performed reasonably on simpler datasets but faltered significantly on complex ones like Digits and Wine Quality.

7.2 Preprocessing and Training:

Data Preparation: Datasets were split into training, validation, and test sets (60/20/20), standardized for distance-based models, and labels were encoded for compatibility. k-NN: Used cross-validation to tune k, with strong results for clean datasets. k-Means: Determined optimal clusters (k) using silhouette scores but showed limited adaptability for supervised tasks.