Introduction to Machine Learning – Milestone 1 Report
Group: 392893_362668_364129

1. Introduction

The objective of this project was to implement and compare several classical machine learning methods for both classification and regression tasks. We studied three main models: Logistic Regression, Linear Regression, and K-Nearest Neighbors (KNN).

In addition to the base implementation, we improved the provided code skeleton by adding a more complete experimental pipeline including data preprocessing, train/validation/test separation, hyperparameter tuning, L2 regularization, early stopping, confusion matrices, runtime measurements, and optional visualizations.

Our objective was not only to obtain good predictive performance, but also to understand the strengths, weaknesses, and computational behavior of each method.

2. Methodology

Data preprocessing

All input features were normalized using the mean and standard deviation computed only on the training set, in order to avoid information leakage. A bias term was then added for linear models.

Validation strategy

When the `--test` flag was not used, the original training set was divided into:

- 80% training data
- 20% validation data

The validation set was used for model selection and hyperparameter tuning. When `--test` was used, the full training set was used for training and the official test set was used only for final evaluation.

Hyperparameter selection

Hyperparameters were selected through grid search using the validation set. For classification tasks, the best configuration was the one achieving the highest validation accuracy, while for regression tasks, the best configuration minimized the validation mean squared error.

Models

Logistic Regression:
We implemented multi-class logistic regression with the softmax function and optimized it using gradient descent. We performed a grid search over the following hyperparameters:

- learning rate $\in \{0.0001, 0.001, 0.01\}$
- number of iterations $\in \{100, 500, 1000\}$
- L2 regularization strength $\lambda \in \{0.0, 0.01, 0.1\}$

This corresponds to 27 tested combinations. We also added early stopping to stop training when the loss improvement became negligible, reducing unnecessary iterations and training time.

Linear Regression:
We implemented the closed-form solution and added optional L2 regularization (ridge regression). We performed a grid search over:

- L2 regularization strength $\lambda \in \{0.0, 0.01, 0.1, 1.0\}$

The best validation performance was obtained with $\lambda = 0.0$.

KNN:
We implemented KNN for both classification and regression. For each query sample, the K nearest training samples were selected using Euclidean distance. In the classification task, the predicted label was obtained by majority vote among the neighbors. In the regression task, the prediction was the average target value of the neighbors. Since KNN does not learn explicit model parameters, the main hyperparameter is the number of neighbors K. We selected it using validation performance with the following grid:

- $K \in \{1, 3, 5, 7, 9, 11, 15\}$

The same tuning strategy was applied for both classification and regression tasks.

For all models, we additionally measured training and prediction times in order to compare not only predictive performance, but also computational efficiency.

3. Experiments and Results

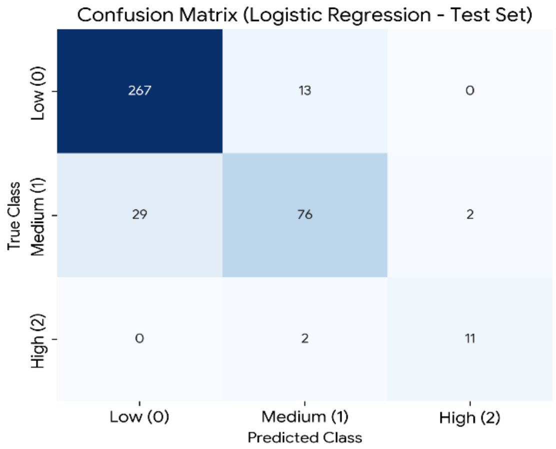
Classification results

Model	Accuracy (Val)	Accuracy (Test)	Macro F1-Score (Test)
Logistic Regression	86.56%	88.50%	0.847
KNN (best K=9)	80.00%	79.75%	0.590

Best Logistic Regression hyperparameters:

- learning rate = 0.001
- max iterations = 1000
- lambda regularization = 0.0

Logistic regression clearly achieved the best classification performance. The confusion matrix showed that most samples were correctly classified, with errors mainly between neighboring classes (Low vs Medium or Medium vs High), which is expected for ordinal addiction levels.



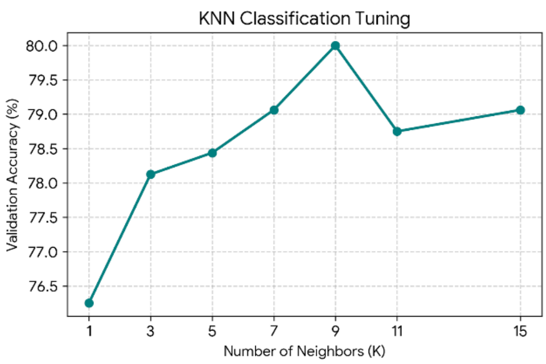
"Figure 1: Confusion matrix for the Logistic Regression model."

These results indicate that gaming behavior and mental health related features contain strong predictive information, and that the classes are relatively well separated.

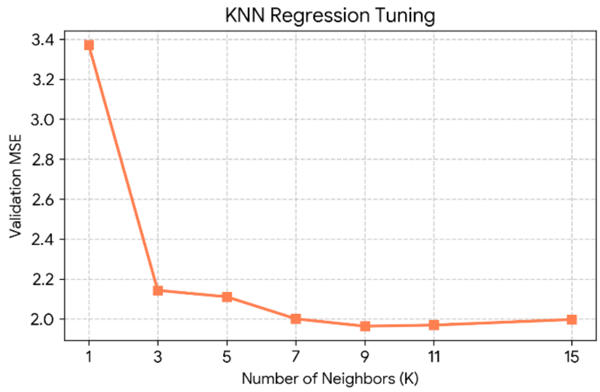
Regression results

Model	Best Hyperparameter	MSE (Val)	MSE (Test)
Linear Regression	λ = 0.0	0.941	0.994
KNN Regression	K = 9	1.964	1.865

Linear regression clearly outperformed KNN for the regression task. This suggests that the relationship between the features and the regression target is mostly linear.



"Figure 2: KNN validation performance across different values of K for classification."



"Figure 3: KNN validation performance across different values of K for regression."

Runtime comparison

Training and prediction times showed important differences:

- **Linear Regression** was the fastest model overall because it uses a direct closed-form solution.
- **Logistic Regression** required moderate training time but had extremely fast prediction once trained.
- **KNN** had almost no training cost, but slower prediction because it computes distances to all stored training samples.

This confirms the expected trade-off between instance-based methods, which shift computation to prediction time, and parametric models, which require more training but allow faster inference.

4. Discussion and Conclusion

Among the classification models tested, Logistic Regression achieved the best performance, reaching 88.5% test accuracy. Since logistic regression is a linear classifier, this suggests that the dataset is approximately linearly separable and that simple linear relationships explain a large part of the target classes.

L2 regularization did not improve performance for logistic or linear regression. This indicates limited overfitting and a relatively clean dataset with informative features.

Linear Regression achieved the best regression results while being extremely efficient computationally. KNN was less competitive on both tasks: with small K values it overfit the training set, while larger K values improved generalization but remained below the linear models.

Overall, this project shows that relatively simple machine learning methods can already obtain strong performance when the data is informative and well structured. Logistic Regression and Linear Regression offered the best compromise between predictive accuracy, simplicity, interpretability, and computational efficiency.