ML PROJECT

Our project implements linear regression and KNN as regressors on a numerical dataset, and logistic regression and KNN as classifiers on an image dataset.

Project Overview

Our project includes implementations of some of the supervised categories of machine learning. We implement linear regression & KNN as regressors on a numerical dataset, also we implement logistic regression & KNN as classifiers on an image data set.

Datasets:

1. **Numerical dataset : Student lifestyle**

2. **Image dataset : Fashion MINST**

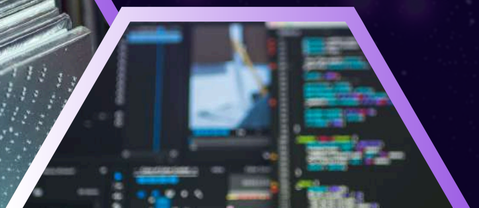
Student lifestyle Dataset

This dataset provides information on the lifestyle of 2,000 students and its correlation with academic performance (GPA). It includes data on study hours, extracurricular activities, sleep, socializing, physical activity, and stress levels, helping to analyze the impact of daily habits on academic performance and student well-being. The data was collected through a Google Form survey from August 2023 to May 2024, primarily reflecting the lifestyles of students in India.

Student lifestyle Dataset

DESCRIPTION

- File Name: Daily_Lifestyle_and_Academic_Performance.csv
- File Format: CSV
- Number of Records: 2000 rows
- Number of Columns: 8 columns
- Column Names: Student ID, Stud- y Hours, Extracurricular Hours, Sleep Hours, Social Hours, Physical Activity Hours, Stress Level, CGPA
- Missing values: there are 50 missing values in the Sleep_Hours_Per_Day column.



Fashion MINST Dataset

Fashion-MNIST is a dataset of 28x28 grayscale images of Zalando's articles, with 60,000 training examples and 10,000 test examples. Each image is associated with a label from 10 classes. It is designed to be a direct replacement for the MNIST dataset, used to benchmark machine learning algorithms. Each image consists of 784 pixels, with pixel values between 0 and 255 indicating lightness or darkness. The dataset includes 785 columns: the first column contains the class labels (representing clothing types), and the remaining columns contain pixel values.

Fashion MINST Dataset

DESCRIPTION

- Each row is a separate image
- Column 1 is the class label.
- Remaining columns are pixel numbers (784 total).
- Each value is the darkness of the pixel (1 to 255)
- Missing : no missing values

Labels: The 5 labels used in our project

Labels	Description
0	T-Shirt/top
1	Trouser
2	Pullover
3	Dress
4	Coat
5	Sandals
6	Shirt
7	Sneaker
8	Bag
9	Ankle boot

Samples :

Numerical validation & testing samples :

600 sample

Numerical training samples:

1400 samples

Image training samples :

60,000 samples

Image validation & Testing sample :

10,000 samples

General comparison between logistic regression & KNN :

- Algorithm
- Working
- Nature
- Training
- Assumptions
- Handling Outliers
- Training Time
- Scalability

Let's get to know it !

Nature

logistic	vs	KNN
Suitable for binary and multiclass classification		Used for both classification and regression.

Training

logistic	vs	KNN
It involves estimating the parameters by minimizing the logistic loss function using techniques like gradient descent		The training process in KNN is minimal, as the algorithm essentially memorizes the training data.

Algorithm:

logistic

vs

KNN

Logistic function is (sigmoid)used

Nearest neighbors are computed using distance metrics

Working

logistic

vs

KNN

Predicts th likelihood of an instance belongs to particular category

Classifies datapoints based on the class of the k-nearest neighbors.

Assumptions

logistic **vs** **KNN**

Assumes a linear relationship between features and the log of odds of variables.

KNN makes no assumptions about the underlying data distribution and is non-linear.

Handling Outliers

logistic **vs** **KNN**

It can be sensitive to outliers which may affect the estimated coefficients.

It is not affected by outliers as the prediction is based on neighbors.

Training Time:

logistic **vs** **KNN**

Faster training times, especially for large datasets.

Training time is negligible as it follows lazy learning approach.

Scalability:

logistic **vs** **KNN**

Performs well with large dataset.

Computationally expensive as the size of the dataset increases.

General comparison between linear regression & KNN (as pros. & cons.) :

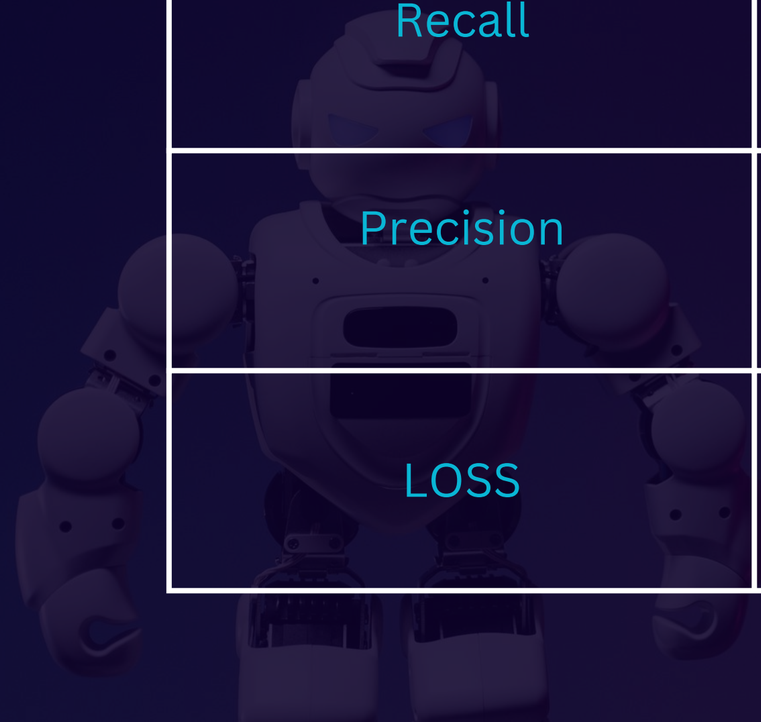
	Pros	Cons
KNN Regression	- Simple and easy - Sensitive to outliers and complex features - Versatile (Regression and classification)	- Not usable for big datasets - Missing values big problem - Equal importance to all features
Linear Regression	- Fast and efficent - Rich statistical insights - Generalist algorithm	- Not suitable for outliers - Only for linear problems - Prone to overfitting

Linear regression & KNN results comparison (applied on the numerical dataset) :

	Linear regression	KNN
Mean absolute error	0.163885782654682 62	0.188765
Mean squared error	0.042725375452414 33	0.055085709999999 96
R^2 score	0.5285578929722119	0.4099849863947047

Logistic regression VS KNN results (applied on the image dataset) :

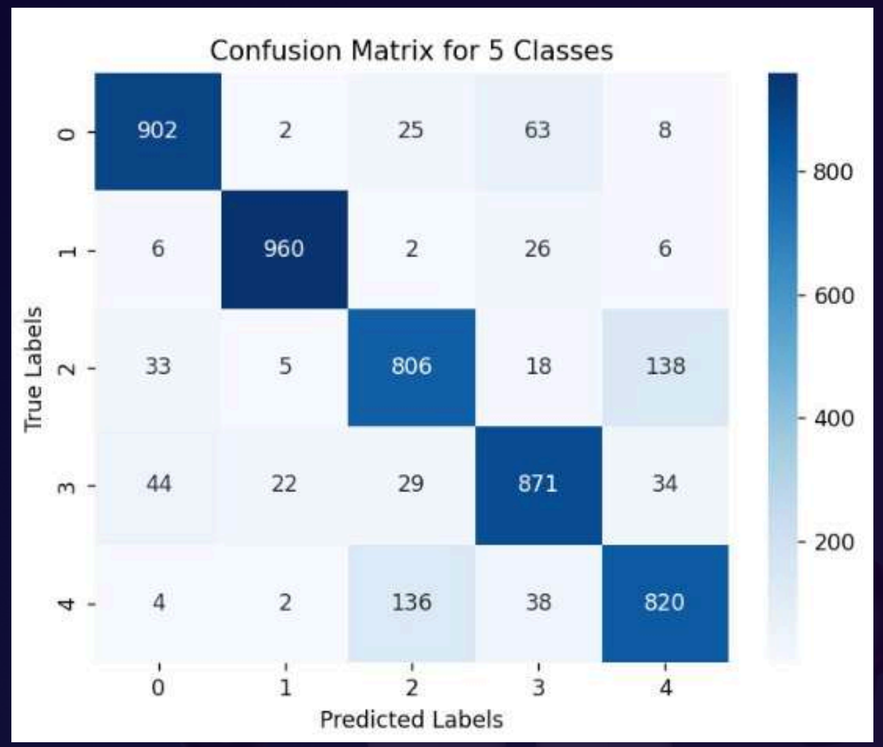
Parameters	Logistic regression	KNN
Accuracy	87.00%	88.76%
Recall	0.87	0.89
Precision	0.87	0.89
LOSS	0.3547	0.7412

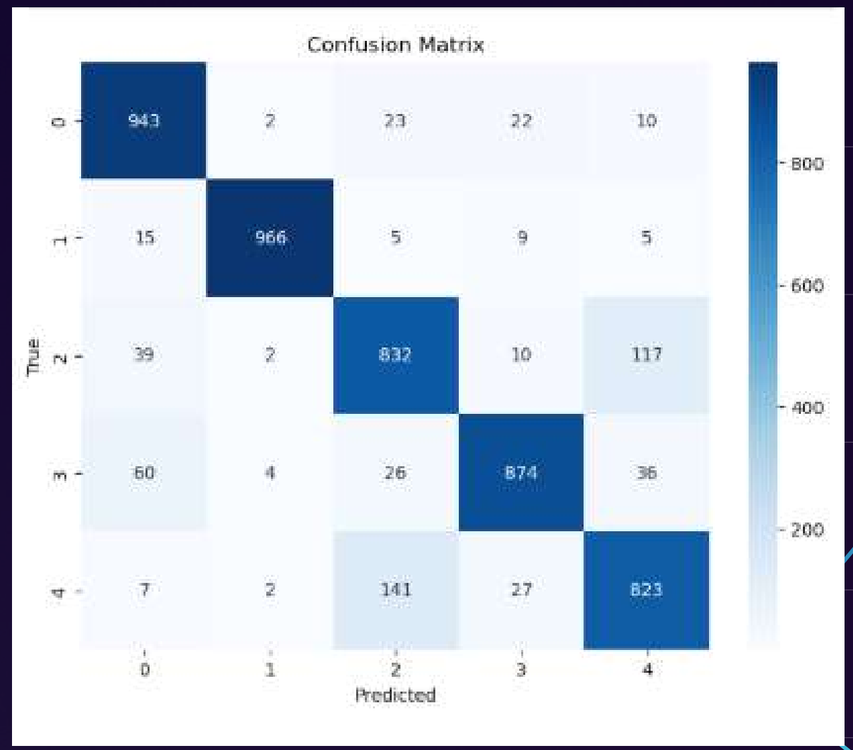


Logistic regression VS KNN results (applied on the image dataset) :

Confusion matrix:

Logistic VS KNN

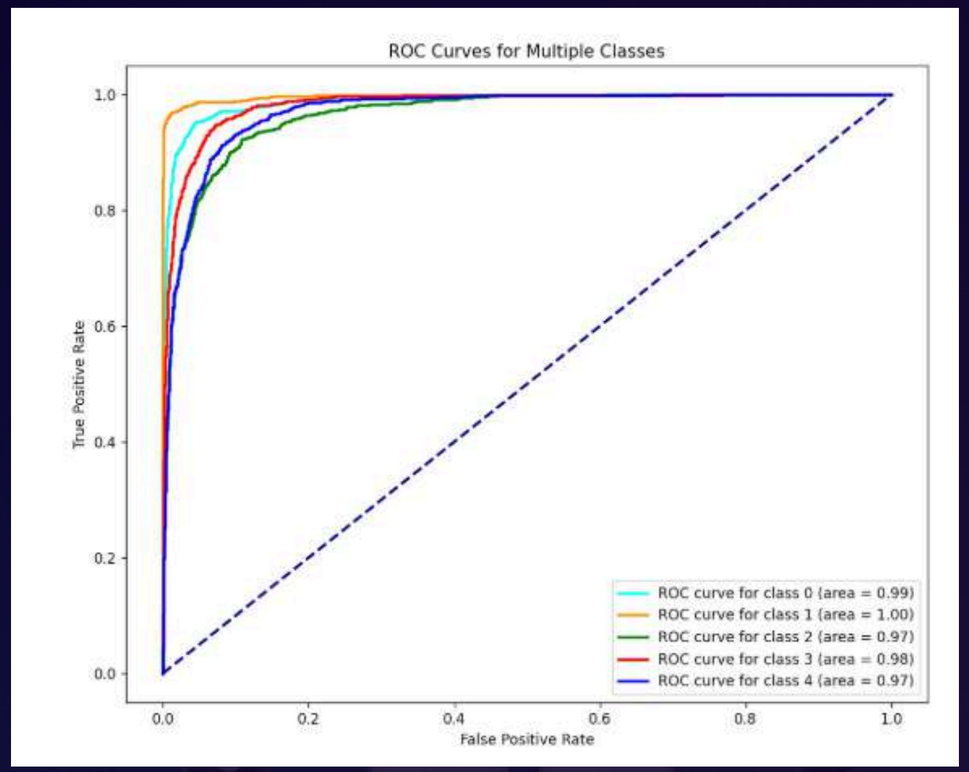


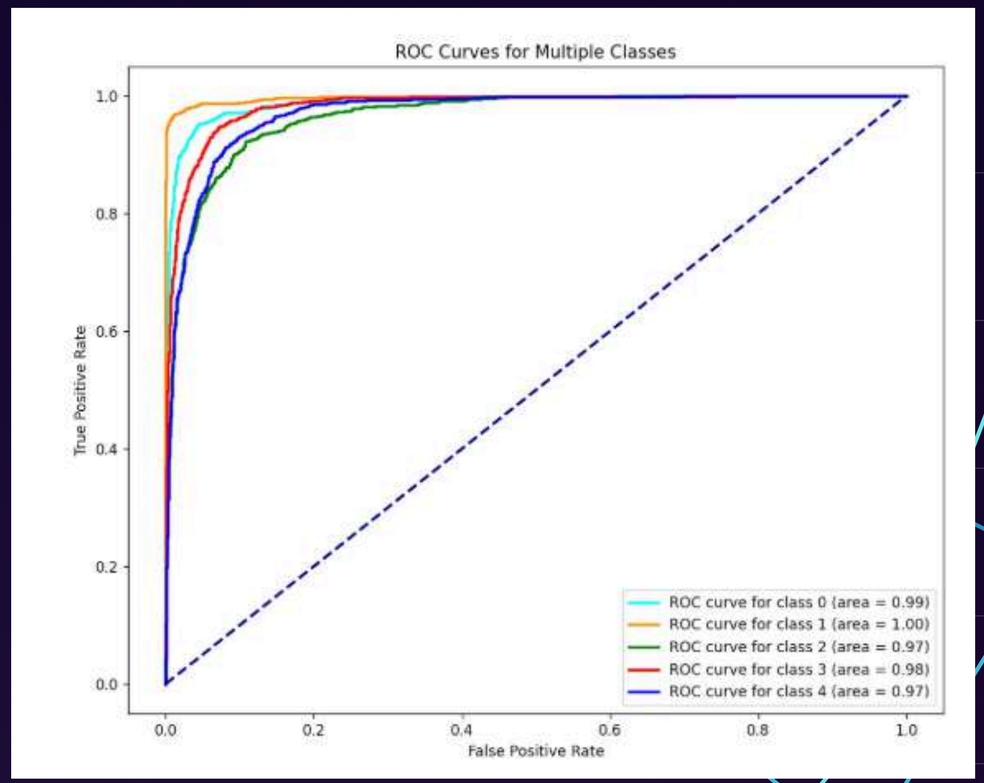


Logistic regression VS KNN results (applied on the image dataset) :

ROC/AUC graph:

Logistic **VS** **KNN**

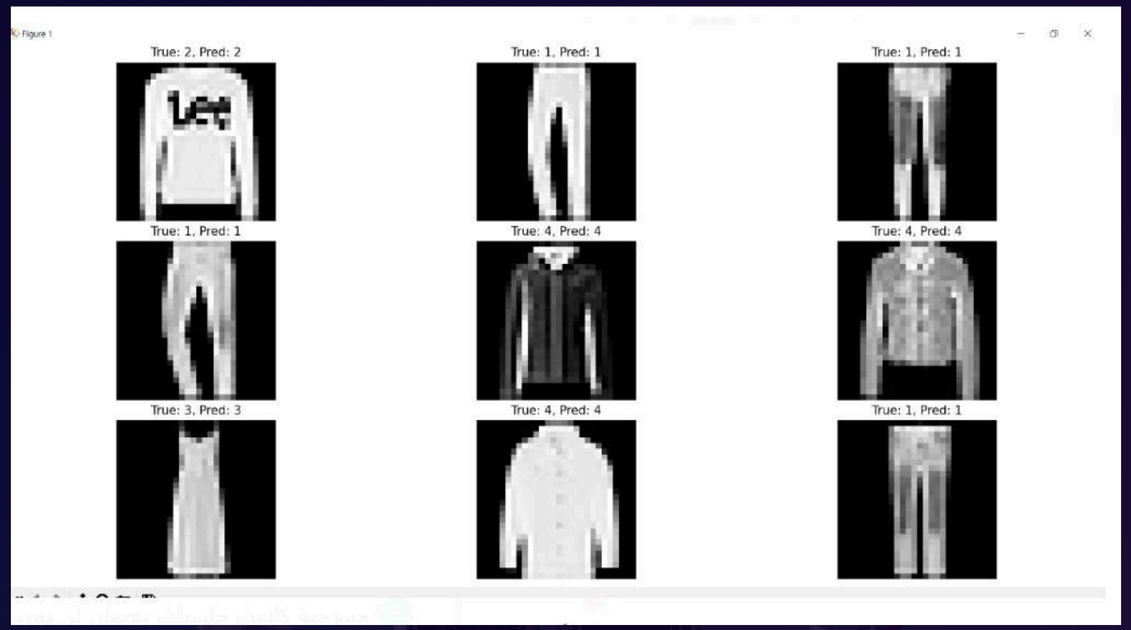


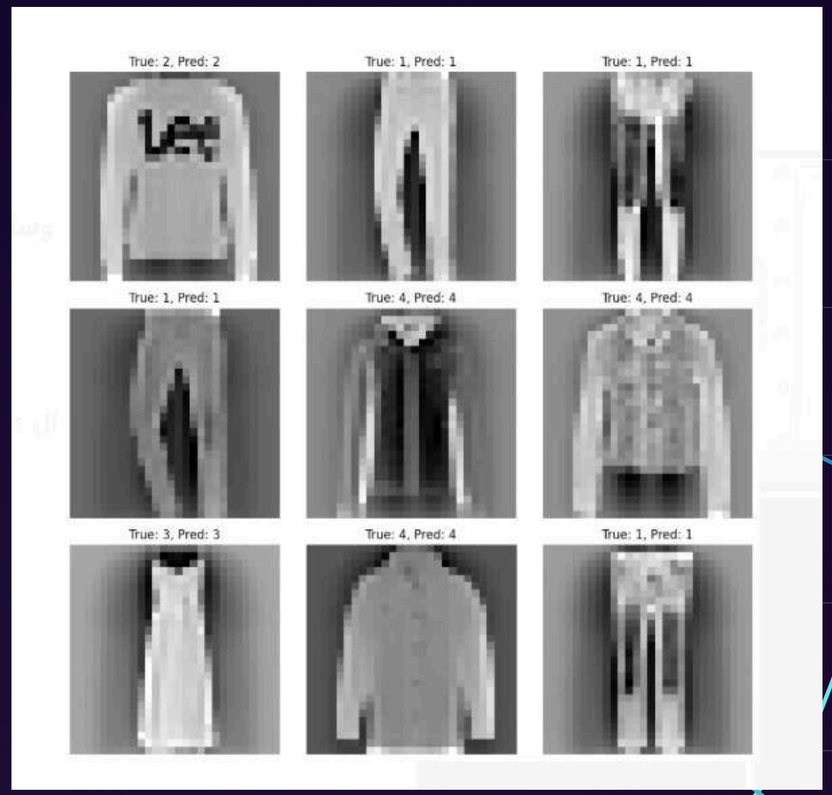


Logistic regression VS KNN results (applied on the image dataset) :

Some of the prediction results:

Logistic **VS** **KNN**





Algorithms :

- Linear regression
- KNN
- Logistic regression

Linear regression

Linear regression is an algorithm that provides a linear relationship between an independent variable and a dependent variable to predict the outcome of future events. It is a statistical method used in data science and machine learning for predictive analysis.

Thus, linear regression is a supervised learning algorithm that simulates a mathematical relationship between variables and makes predictions for continuous or numeric variables such as sales, salary, age, product price, etc.

- Train Time Complexity=O(n*m^2 + m^3)
- Test Time Complexity=O(m)
- Space Complexity = O(m)

KNN

The k-nearest neighbors (KNN) algorithm is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. It is one of the popular and simplest classification and regression classifiers used in machine learning today.

While the KNN algorithm can be used for either regression or classification problems, it is typically used as a classification algorithm, working off the assumption that similar points can be found near one another.

- Train Time Complexity=O(k*n*m)
- Test Time Complexity=O(n*m)
- Space Complexity = O(n*m)

Logistic regression

Logistic regression is defined as a supervised machine learning algorithm that accomplishes binary classification tasks by predicting the probability of an outcome, event, or observation. This article explains the fundamentals of logistic regression, its mathematical equation and assumptions, types, and best practices.

Logical regression analyzes the relationship between one or more independent variables and classifies data into discrete classes. It is extensively used in predictive modeling, where the model estimates the mathematical probability of whether an instance belongs to a specific category or not.

- Train Time Complexity=O(n*m)
- Test Time Complexity=O(m)
- Space Complexity = O(m)

Used resources :

- https://www.kaggle.com/datasets/steve1215rogg/student-lifestyle-dataset
- https://www.kaggle.com/datasets/zalando-research/fashionmnist
- https://www.geeksforgeeks.org/logistic-regression-vs-k-nearest-neighbors-in-machine-learning/
- https://medium.com/machine-learning-t%C3%BCrkiye/localized-regression-knn-with-local-regression-7b4d302adb85
- https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-linear-regression/#:~:text=Linear%20regression%20is%20an%20algorithm,machine%20learning%20for%20predictive%20anal
- https://www.ibm.com/topics/knn#:~:text=the%20next%20step-,What%20is%20the%20KNN%20algorithm%3F,of%20an%20individual%20data%20point.
- https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-logistic-regression/#:~:text=Logistic%20regression%20is%20defined%20as,outcome%2C%20event%2C%20or%20observation.


Thank You!