

Name	ID	Department	
مونيا ايمن فوزي ذكي	20210975	CS	3rd year
ملك اشرف عبدالحميد	20210955	CS	3rd year
نادين طارق محمد	20210987	AI	3rd year
ميريام سامح فهم	20210983	CS	3rd year
ندى حاتم مهدى عيسى	20210994	CS	3rd year
عبدالله علي محمد مصطفى	٢٠٢١٠٥٥٢	CS	3rd year

DATASET

plant pathology 2020 - fgvc 7

<https://www.kaggle.com/c/plant-pathology-2020-fgvc7>

About Dataset:

The Plant Pathology dataset encompasses a wide array of plant leaf images, showcasing diverse instances of diseases such as rust, scab, and multiple diseases, along with representations of healthy leaves. It includes images exhibiting variations in disease severity, leaf angles, lighting conditions, and background clutter, simulating real-world scenarios encountered in agricultural settings.

This dataset proves instrumental in a multitude of applications:

Disease Identification: Researchers leverage this dataset to train machine learning models for accurate identification and classification of plant diseases from images.

Model Evaluation: It serves as a benchmark for assessing the performance of various computer vision algorithms and techniques in detecting and categorizing plant diseases.

Agricultural Impact: The dataset's utility extends to aiding in the development of AI-powered solutions that empower farmers with early disease detection capabilities, facilitating prompt intervention and optimized crop management practices.

Labels:

Each image in the dataset is associated with specific labels embedded in the file name, structured as follows:

Disease annotations are incorporated, indicating the type of disease or the healthy state of the plant leaves.

The labels specify the particular disease present, including 'healthy', 'rust', 'scab', 'multiple_diseases'.

Number of Classes: There are four classes in this dataset, represented by 'healthy', 'multiple_diseases', 'rust', and 'scab'.

Total Number of Samples in the Dataset: The dataset comprises a total of 3642 samples.

Size of Each Image: The images are resized to a final size of (64, 128) pixels.

Number of Samples Used in Training, Validation, and Testing:

- **Training set:** 700 samples
- **Test set:** 300 samples

Implementation details :

We used two algorithms K-Means and Logistic Regression

Feature Extraction Phase:

Histogram of Oriented Gradients (HOG)

The feature extraction phase in this project employs the Histogram of Oriented Gradients (HOG) technique to extract discriminative features from images.

Number of Features Extracted:

Upon applying the HOG method to each image in the dataset, a total of 2916 features were extracted from each image.

Feature Names and Dimension:

Feature Names: The extracted features are denoted as 'HOG_feature_i', where 'i' represents the index of the feature.

Dimension of Resulted Features: Each image yields a feature vector of size (2916,). This means that for every image processed, a feature vector with 2916 elements is generated, representing the HOG descriptors.

Class names are encoded into numeric labels using **label_encoder**.

Data is divided into training (**X_train**, **y_train**) and testing (**X_test**, **y_test**) sets.

into 20% test and 80% train by `train_test_split` algorithm.

train_data stores training features and labels, while **test_data** holds testing features and labels.

- **Training Data:**
 - **train_data** combines features (**X_train**) and their respective labels (**y_train**).
 - **X_train** holds features used for model training.
 - **y_train** contains corresponding labels for the features in **X_train**.
- **Testing Data:**
 - **test_data** merges features (**X_test**) and their associated labels (**y_test**).
 - **X_test** includes features for testing the trained model.
 - **y_test** stores labels linked to the features in **X_test**

KMEANS

Goal:

The goal of this code is to perform clustering analysis on extracted HOG (Histogram of Oriented Gradients) features using the K-means clustering algorithm. It aims to find an optimal number of clusters and visualize the clustering results.

Code Explanation:

Dimensionality Reduction and Feature Scaling:

Uses Principal Component Analysis (PCA) to reduce the dimensionality of the HOG features to 2D.

Scales the reduced features using StandardScaler for uniform scaling.

Finding Optimal Number of Clusters (K):

Iterates through different values of k (number of clusters) and computes the silhouette score for each k-value.

Prints the silhouette scores for different k-values to identify the optimal number of clusters.

Performing K-means Clustering:

Utilizes KMeans algorithm to perform clustering with the identified optimal k.

Maps cluster labels to specific colors for visualization.

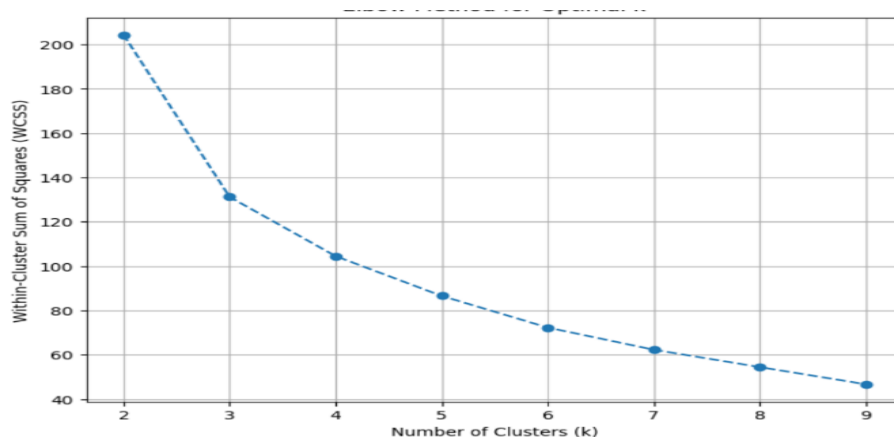
Generates a scatter plot of the PCA-reduced HOG features, marking cluster labels with distinct colors and centroids with yellow markers.

Elbow Method for Optimal K:

Computes the Within-Cluster Sum of Squares (WCSS) for a range of k-values to visually determine the optimal number of clusters using the Elbow Method.

Result:

The average silhouette score for clustering: 0.3394225139687757



Logistic Regression

Goal:

The goal of this code is to perform a classification task using Logistic Regression on Histogram of Oriented Gradients (HOG) features. It aims to evaluate model performance, including accuracy metrics, confusion matrix visualization, and decision boundary plotting.

Code Explanation:

Data Preparation:

Scales the HOG features using StandardScaler and applies PCA for visualization.

Defines class labels and initializes Logistic Regression for a multiclass classification problem.

Model Training and Evaluation:

Fits the Logistic Regression model on the scaled HOG features and evaluates its accuracy on the entire dataset.

Computes classification report metrics (precision, recall, f1-score) and prints the results.

Cross-Validation:

Performs cross-validation using 5 folds on the entire dataset to assess model performance across different subsets.

Prints the cross-validation scores and calculates the mean accuracy.

Confusion Matrix Visualization:

Creates a confusion matrix for the entire dataset and visualizes it using a heatmap.

Decision Boundary Visualization:

Fits the Logistic Regression model on PCA-transformed HOG features and generates decision boundaries.

Plots decision boundaries and the entire dataset to visualize the model's classification areas.

Model Details:

Cross-Validation: The code uses 5-fold cross-validation to evaluate the model's performance.

Hyperparameters:

solver: 'saga'

max_iter: 1000

penalty: 'l2'

Result:

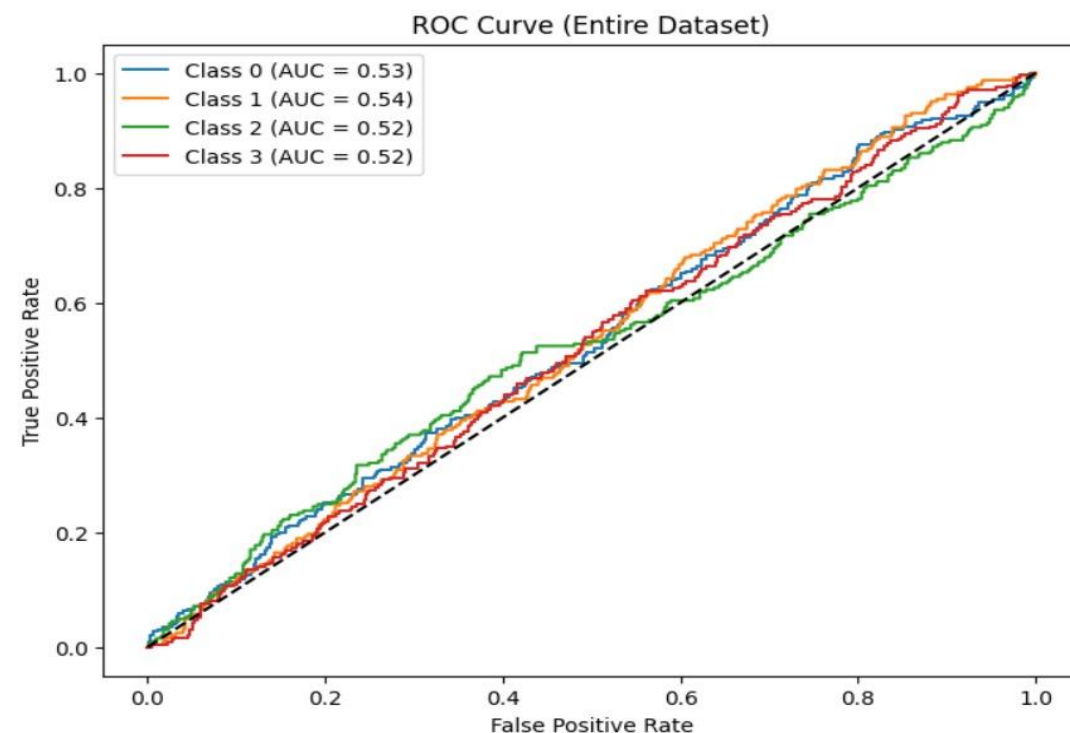
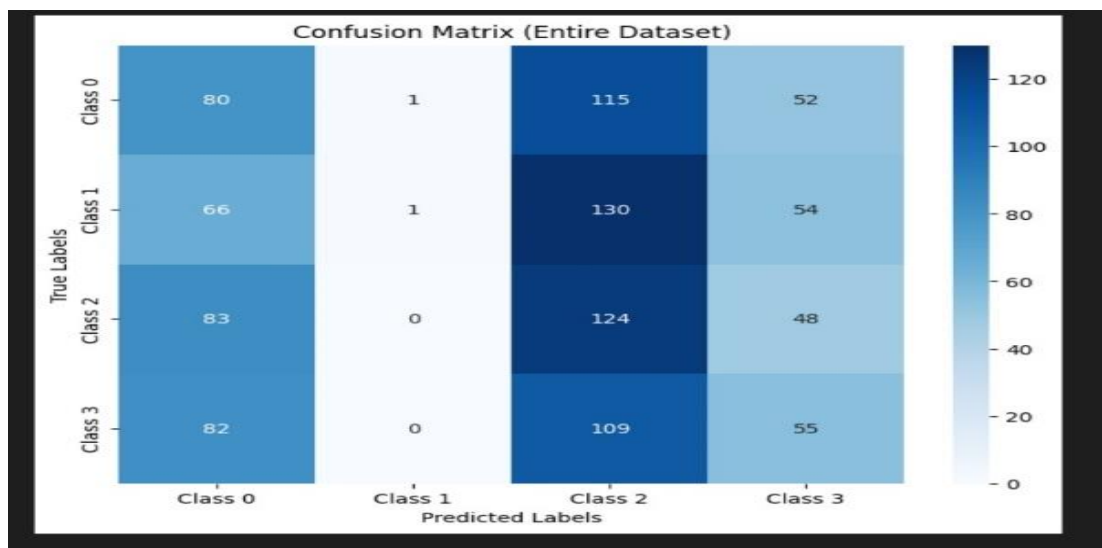
Accuracy: The model achieves an accuracy of approximately 33.77% on the entire dataset.

Classification Report: Provides precision, recall, and f1-score for each class along with macro and weighted averages.

Cross-Validation Scores: Shows the model's performance across different subsets with a mean accuracy of approximately 26.69%.

Confusion Matrix: Visualizes the true and predicted labels' distribution.

Decision Boundaries: Illustrates the model's classification areas in a 2D space.



Numerical dataset :

a) students performance in exams

Link :

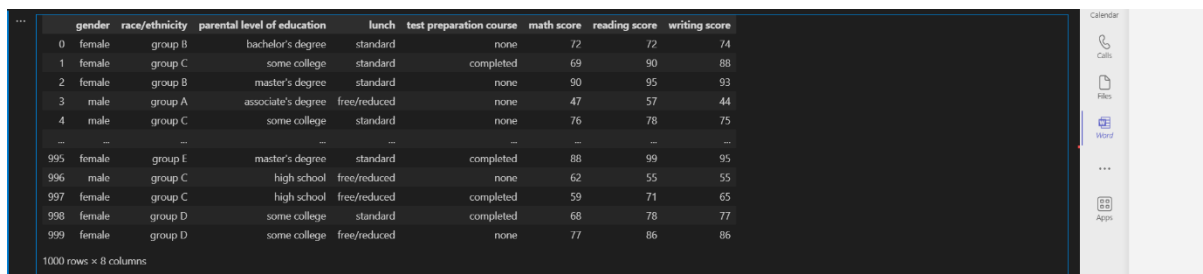
<https://www.kaggle.com/datasets/spscientist/students-performance-in-exams>

Total num of sample : 1001

Testing data: 250

Training data :751

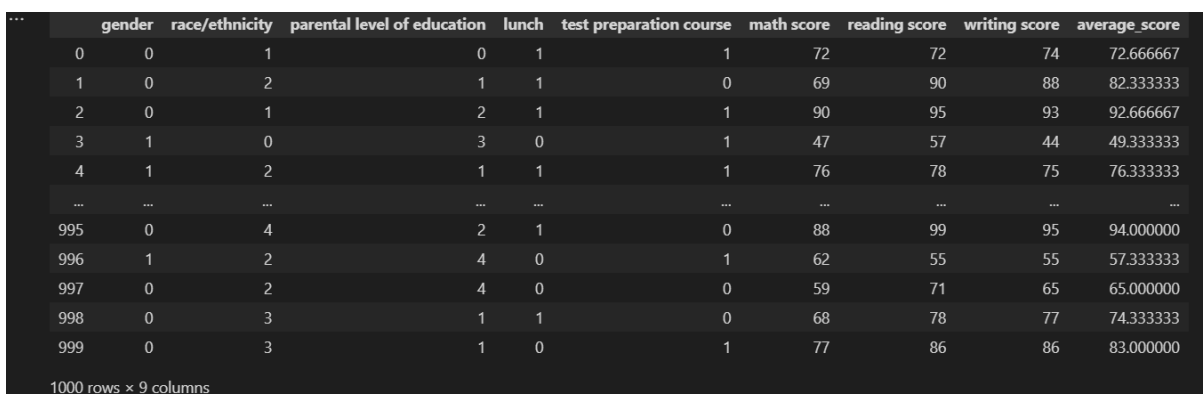
b) data before encoding:



The screenshot shows a preview of a dataset with 1000 rows and 8 columns. The columns are: gender, race/ethnicity, parental level of education, lunch, test preparation course, math score, reading score, and writing score. The data is displayed in a table format with a dark background and white text. The first few rows are visible, showing various student profiles and their scores.

	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
0	female	group B	bachelor's degree	standard	none	72	72	74
1	female	group C	some college	standard	completed	69	90	88
2	female	group B	master's degree	standard	none	90	95	93
3	male	group A	associate's degree	free/reduced	none	47	57	44
4	male	group C	some college	standard	none	76	78	75
...
995	female	group E	master's degree	standard	completed	88	99	95
996	male	group C	high school	free/reduced	none	62	55	55
997	female	group C	high school	free/reduced	completed	59	71	65
998	female	group D	some college	standard	completed	68	78	77
999	female	group D	some college	free/reduced	none	77	86	86

after encoding:



The screenshot shows the same dataset after encoding categorical variables into numerical values. The columns are: gender, race/ethnicity, parental level of education, lunch, test preparation course, math score, reading score, writing score, and average_score. The data is displayed in a table format with a dark background and white text. The first few rows are visible, showing the encoded values for the categorical variables.

	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score	average_score
0	0	1	0	1	1	72	72	74	72.666667
1	0	2	1	1	0	69	90	88	82.333333
2	0	1	2	1	1	90	95	93	92.666667
3	1	0	3	0	1	47	57	44	49.333333
4	1	2	1	1	1	76	78	75	76.333333
...
995	0	4	2	1	0	88	99	95	94.000000
996	1	2	4	0	1	62	55	55	57.333333
997	0	2	4	0	0	59	71	65	65.000000
998	0	3	1	1	0	68	78	77	74.333333
999	0	3	1	0	1	77	86	86	83.000000

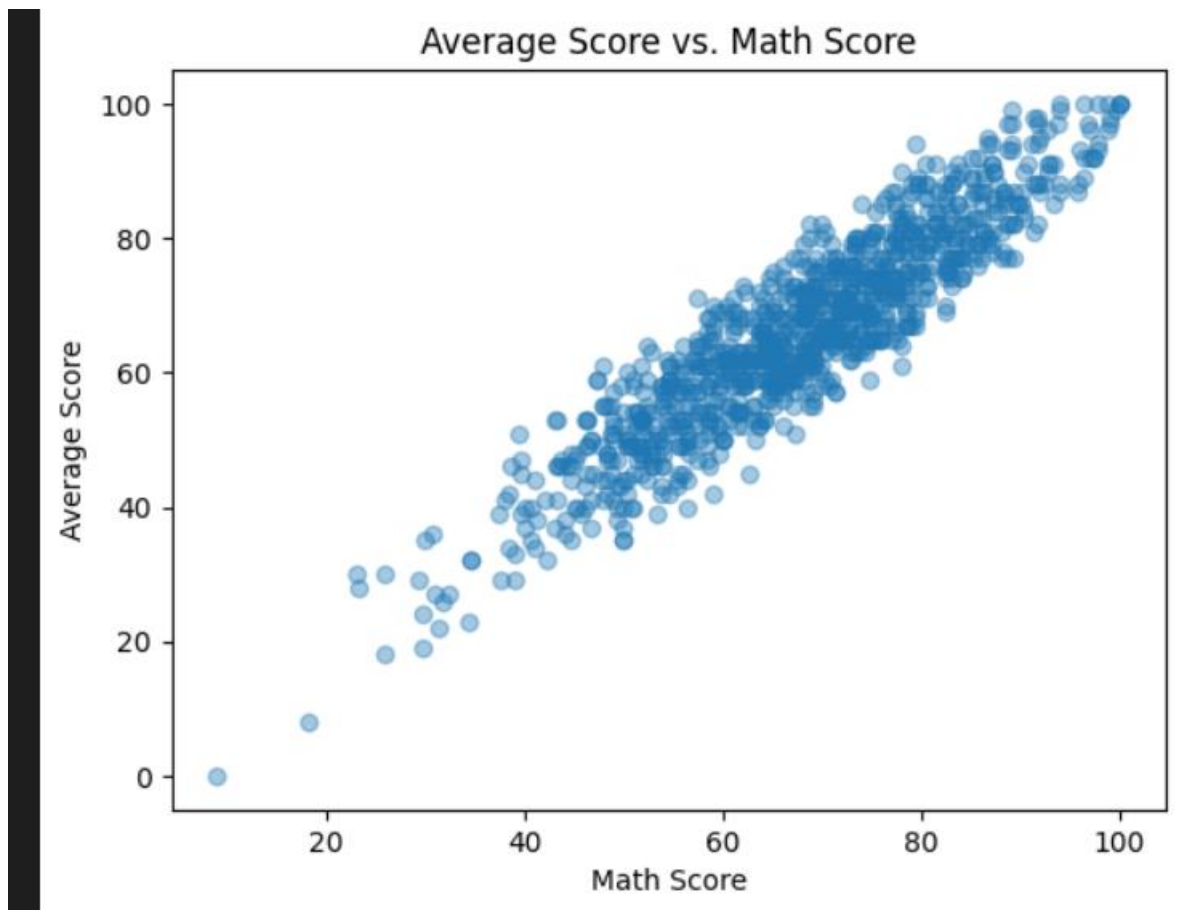
a) Linear Regression

b) Goal : predict students grade

c) Code : used linear regression from sklearn.linear_model

e) Result: R-squared : 100 %





b) K-nearest neighbor model

Goal : predict students grades

Selecting the best K and parameters using gridsearch I get these results

Number of Neighbors in k-Nearest Neighbors (KNN):* The value of k in KNN is a hyperparameter that determines the number of neighbors considered when making predictions

Standardization, also known as z-score normalization or zero-mean normalization, is a preprocessing technique used in statistics and machine learning to transform the features of a dataset to have a mean of 0 and a standard deviation of 1. This process makes it easier to compare and interpret the data, especially when the features originally have different scales or units

c) accuracy of linear regression is 100%

accuracy of knn is 99.15%