# **Computers and Artificially intelligence**

## ***AI330\_ Machine Learning Projects\_Fall2023***

**Team information:**

|  |  |
| --- | --- |
| **Name** | **ID** |
| 1. **مريم محمود سيد محمود** | **20210926** |
| 1. **مريم ياسر عمر شندي** | **20210931** |
| 1. **مريم عزت حسن همام** | **20210916** |
| 1. **مريم مدحت الليثي** | **20210928** |
| 1. **نورا على عبده محمود** | **20211010** |
| 1. **احمد صابر محمود محمد** | **20210063** |

1. **General Information on datasets:**

|  |  |
| --- | --- |
| **Numerical dataset** | **Image dataset** |
| **Name of dataset:** Flight price prediction  Dataset link**:**  **[https://www.kaggle.com/datasets/shubhambathwal/flight-price-prediction](https://www.kaggle.com/datasets/shubhambathwal/flight-price-prediction%09)** | **Name of image dataset:** Tomato Detection  Dataset link**:**  **<https://www.kaggle.com/datasets/nexuswho/tomato-detect>** |
| **Total number of samples:** 300153 rows  11 columns | **Total number of samples:** 16764 |
| **Linear Regression:**  **Train:**  **x\_train**= (208630, 10) **y\_train**= (208630,)  **Test:**  **x\_test** = (89413, 10) **y\_test=** (89413,)  **KNN Regressor:**  **Train:**  **x\_train**= (208630, 10) **y\_train**= (208630,)  **Test:**  **x\_test** = (89413, 10) **y\_test=** (89413,)  **Total number of samples=** 300154 rows & 12 columns | **K-means:**  **Train:**  **x\_train= 9000**  **Logistic Regression:**  **Train:**  **x\_train= 7000**  **y\_train=7000**  **Test:**  **x\_test = 1405 y\_test= 1405** |
|  | **Labels:**  **Logistic regression:** same number as images.  **KMeans:** no labels are used.  **Number of classes:**  2 (fresh and rotten) |

# **Implementation details:**

* **KMeans**

# **Feature extraction:**

1. Name of feature extraction used: SIFT Feature Extractor
2. Number of extracted features: 9000\*128
3. The dimension of resulted features: (9000, 16384)

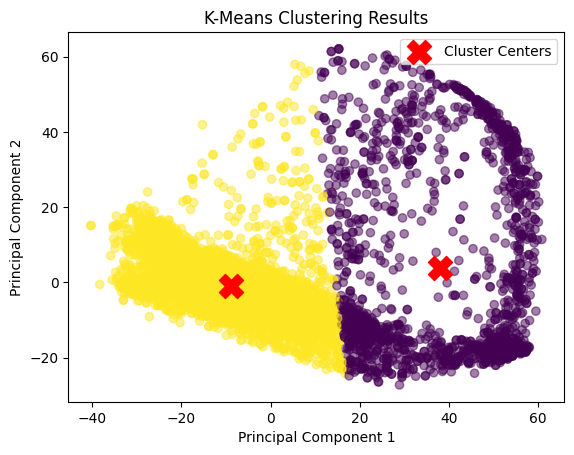
* **Hyperparameters:**

**- kmeans = KMeans (n\_clusters=2 , random\_state= 42)**

1.  n\_clusters: specifies the number of clusters.

2. random\_state: determines the random seed for the initialization of the K-means.

* **Diagram:**



* **Silhouette Score = 0.6198**
* **Logistic Regression**

# **Feature extraction:**

1. Name of feature extraction used: SIFT Feature Extractor
2. Number of extracted features: 5460\*128
3. The dimension of resulted features: images (5460, 16384)

Labels (1104, 16384)

* **Hyperparameters:**

**- model = LogisticRegression (random\_state=42, max\_iter=1000)**

1. max\_iter: It determines the maximum number of iterations for

the solver to converge.

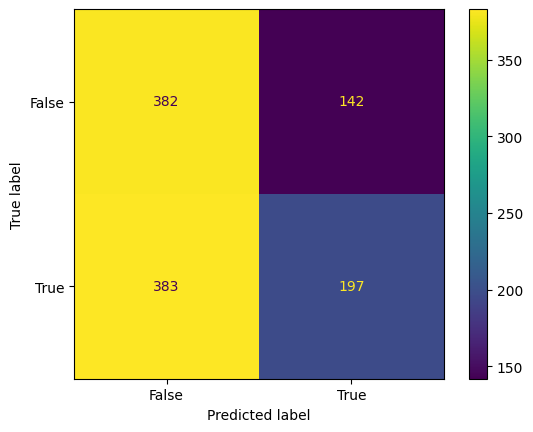
2. random\_state: determines the random seed for the initialization of

the K-means.

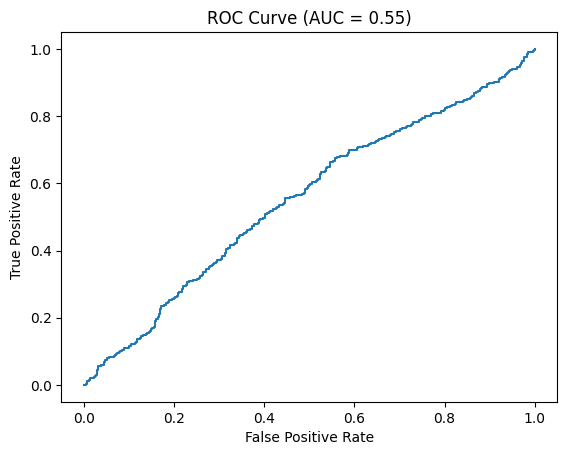
* **descriptor\_size = 128**

1. It specifies the desired size of the SIFT descriptors.

* **Diagrams:**
* **Confusion matrix:**



* **ROC Curve:**



* **Model score: 52.4456%**
* **Linear Regression**
* **Hyperparameters:**

**- No hyperparameters is used**

* **Diagram:**

**A graph showing the difference between a price and a price

Description automatically generated**

* **Optimization:**

**- Data scaling : using Minmax Scaler**

**- Date Preprocessing : encode object data**

* **Model score:**

**- Training set : 0.9047818372733851**

**- Test set : 0.9045280699666446**

* **KNN Regressor**
* **Hyperparameters:**

**-** **KNN = KNeighborsRegressor(n\_neighbors=5)**

1. n\_neighbors: Number of neighbors to consider in the KNN model

* **Diagram:**

**A graph showing a comparison of the price of a product

Description automatically generated with medium confidence**

* **Optimization:**

**- Data scaling: using Minmax Scaler**

**- Date Preprocessing: encode object data**

* **Model score:**

**- Training set: 0.9833754447848447**

**- Test set: 0.9738976560523219**