Image Classification using Logistic Regression and K-Means on the Fruits-360 Dataset

---

1. Introduction

Image classification is a fundamental task in machine learning that aims to automatically assign a class label to an image based on its visual content. While deep learning approaches such as Convolutional Neural Networks (CNNs) are commonly used for image classification, classical machine learning algorithms can still provide valuable insights when combined with appropriate feature extraction and preprocessing techniques.

The main objective of this project is to explore how **traditional machine learning algorithms** perform on image data after suitable preprocessing. Specifically, the project investigates the use of **Logistic Regression** as a supervised classification method and **K-Means** as an unsupervised clustering method on a subset of the Fruits-360 image dataset.

---

2. Dataset Description

The project uses the **Fruits-360 dataset**, which consists of images of fruits captured under controlled conditions. Each image has:

- A fixed resolution of **100 × 100 pixels**
- A plain background
- Consistent lighting conditions

For this project, **five fruit classes** were selected to reduce complexity and improve class separability:

- Class 0
- Class 1
- Class 2
- Class 3
- Class 4

Limiting the number of classes helps classical models better distinguish between categories and avoids excessive overlap in feature space.

---

3. Problem Statement and Objectives

3.1 Problem Statement

Classical machine learning algorithms are not inherently designed to process raw image data. Therefore, it is important to investigate whether such algorithms can still perform effective image classification when images are transformed into numerical feature vectors.

3.2 Objectives

The objectives of this project are:

1. To preprocess image data and convert it into a suitable numerical representation.
2. To apply **Principal Component Analysis (PCA)** for dimensionality reduction.
3. To train and evaluate a **Logistic Regression** classifier on image features.
4. To apply **K-Means clustering** to explore the intrinsic structure of the image data.
5. To compare supervised and unsupervised learning approaches on the same dataset.

---

4. Data Preprocessing

Before applying any machine learning model, several preprocessing steps were performed:

1. **Image Loading:**
   Images were read using OpenCV while handling unreadable or invalid files.
2. **Grayscale Conversion:**
   Images were converted to grayscale to reduce computational complexity and focus on intensity-based features.
3. **Resizing:**
   All images were resized to a fixed resolution of **100 × 100 pixels** to ensure consistency.
4. **Flattening:**
   Each image was flattened into a one-dimensional vector of **10,000 features**.
5. **Normalization:**
   Pixel values were normalized to the range [0, 1] by dividing by 255.
6. **Train-Test Split:**
   The dataset was split into training (80%) and testing (20%) sets using **stratified sampling** to preserve class balance.

---

5. Dimensionality Reduction using PCA

Image data is inherently high-dimensional, which can negatively impact the performance of classical machine learning models. To address this issue, **Principal Component Analysis (PCA)** was applied.

- PCA was configured to retain **95% of the total variance**.
- Feature dimensionality was reduced from **10,000 to 137 features**.

Why PCA?

- Reduces computational cost
- Removes redundant and noisy features
- Mitigates the curse of dimensionality
- Improves model generalization

---

6. Logistic Regression (Supervised Learning)

6.1 Model Description

Logistic Regression is a linear supervised learning algorithm commonly used for classification tasks. In this project, it was used as a **baseline classifier** to evaluate how well a classical model can separate image classes after PCA.

6.2 Training Procedure

The Logistic Regression model was trained using:

- PCA-transformed features
- StandardScaler applied only on training data
- Maximum iterations set to ensure convergence

6.3 Evaluation Metrics

The model was evaluated using:

- Accuracy
- Precision

- Recall
- F1-score
- Confusion Matrix

6.4 Results

The Logistic Regression model achieved:

- **Accuracy ≈ 99.8%**
- Only one misclassification across all test samples

This high performance is attributed to the controlled nature of the dataset and the strong linear separability of classes after preprocessing.

---

7. K-Means Clustering (Unsupervised Learning)

7.1 Purpose

Unlike Logistic Regression, K-Means does not use class labels during training. It was applied to:

- Explore the intrinsic structure of the dataset
- Analyze whether images naturally form meaningful clusters

7.2 Methodology

- K-Means was applied to the **PCA-reduced feature space**
- The number of clusters was set to 5 to match the number of classes

7.3 Evaluation

Since K-Means is unsupervised, classification accuracy is not an appropriate metric. Instead, clustering quality was evaluated using the **Silhouette Score**.

7.4 Results

- **Silhouette Score ≈ 0.28**

This score indicates moderate cluster separation, which is expected for image data due to visual similarities between different fruit classes.

---

8. Comparison Between Logistic Regression and K-Means

Aspect	Logistic Regression	K-Means
Learning Type	Supervised	Unsupervised
Uses Labels	Yes	No
Task	Classification	Clustering
Performance	Very High	Moderate
Purpose	Accurate prediction	Data exploration

The results demonstrate that supervised learning significantly outperforms unsupervised clustering for image classification tasks.

---

9. Limitations

Despite the strong results, this project has several limitations:

- Flattened pixel features do not capture spatial relationships.
- The dataset is highly controlled and does not reflect real-world image conditions.
- Grayscale conversion removes color information that could improve performance.

---

10. Future Work

Potential improvements include:

- Using color-based features or handcrafted descriptors
- Applying convolutional neural networks (CNNs)
- Testing the model on real-world images with complex backgrounds
- Exploring other dimensionality reduction techniques

---

11. Conclusion

This project demonstrated that classical machine learning algorithms can perform effective image classification when combined with appropriate preprocessing and dimensionality reduction techniques. Logistic Regression achieved excellent performance after PCA, while K-

Means provided valuable insights into the data's structure. The comparison highlights the importance of supervision in achieving high classification accuracy and provides a strong foundation for transitioning to more advanced deep learning approaches.

Fruits-360 على مجموعة بيانات K-Means تصنيف الصور باستخدام الانحدار اللوجستي وخوارزمية

---

1. المقدمة

يُعد تصنيف الصور من المهام الأساسية في مجال تعلم الآلة، حيث يهدف إلى إسناد فئة (تصنيف) لكل صورة اعتمادًا على محتواها البصري. وعلى الرغم من أن تقنيات التعلم العميق مثل الشبكات العصبية الالتفافية (CNNs) تُستخدم على نطاق واسع في تصنيف الصور، إلا أن خوارزميات تعلم الآلة التقليدية ما زالت قادرة على تقديم نتائج مفيدة عند دمجها مع تقنيات مناسبة لاستخلاص السمات والمعالجة المسبقة.

يهدف هذا المشروع إلى دراسة أداء **خوارزميات تعلم الآلة التقليدية** عند تطبيقها على بيانات الصور بعد إجراء المعالجة المسبقة المناسبة. وبشكل خاص، يستكشف المشروع استخدام **الانحدار اللوجستي (Logistic Regression)** كخوارزمية تعلم مُراقَب، واستخدام **K-Means** كخوارزمية تعلم غير مُراقَب، وذلك باستخدام جزء من مجموعة بيانات Fruits-360.

---

2. وصف مجموعة البيانات

يعتمد المشروع على **مجموعة بيانات Fruits-360**، والتي تحتوي على صور لفواكه تم التقاطها في ظروف مُتحكم فيها، حيث تتميز الصور بما يلي:

- دقة ثابتة **100 × 100 بكسل**
- خلفية بسيطة وموحدة
- إضاءة ثابتة ومتجانسة

تم اختيار **خمس فئات فقط من الفواكه** في هذا المشروع بهدف تقليل التعقيد وتحسين القدرة على الفصل بين الفئات. ويساعد تقليل عدد الفئات خوارزميات تعلم الآلة التقليدية على التمييز بشكل أفضل بين الأصناف المختلفة وتقليل التداخل في فضاء السمات.

---

3. مشكلة البحث وأهداف المشروع

3.1 مشكلة البحث

خوارزميات تعلم الآلة التقليدية ليست مصممة بطبيعتها للتعامل مع الصور الخام. لذلك، من المهم دراسة مدى قدرتها على تصنيف الصور بدقة بعد تحويلها إلى تمثيل عددي مناسب.

3.2 أهداف المشروع

يهدف هذا المشروع إلى:

1. معالجة بيانات الصور وتحويلها إلى متجهات رقمية مناسبة.
2. تطبيق **تحليل المكونات الرئيسية (PCA)** لتقليل الأبعاد.
3. تدريب وتقييم نموذج **الانحدار اللوجستي** لتصنيف الصور.
4. تطبيق **K-Means** لاكتشاف البنية الداخلية للبيانات بدون استخدام التصنيفات.
5. مقارنة أساليب التعلم المُراقَب وغير المُراقَب على نفس مجموعة البيانات.

---

4. المعالجة المسبقة للبيانات

قبل تطبيق أي نموذج لتعلم الآلة، تم تنفيذ الخطوات التالية:

1. **قراءة الصور:**
   تم استخدام مكتبة OpenCV لقراءة الصور مع التعامل مع الملفات التالفة أو غير القابلة للقراءة.
2. **التحويل إلى التدرج الرمادي:**
   تم تحويل الصور إلى Grayscale لتقليل التعقيد الحسابي والتركيز على شدة الإضاءة بدلًا من الألوان.
3. **تغيير الحجم:**
   تم توحيد أبعاد جميع الصور إلى **100 × 100 بكسل**.
4. **تسطيح الصور (Flattening):**
   تم تحويل كل صورة إلى متجه أحادي البعد يحتوي على **10,000 سمة**.
5. **التطبيع (Normalization):**
   تم تطبيع قيم البكسل إلى المدى [0,1].

6. **تقسيم البيانات:**
   تم تقسيم البيانات إلى مجموعة تدريب (80%) واختبار (20%) باستخدام التقسيم الطبقي (Stratified Split) للحفاظ على توازن الفئات.

---

5. تقليل الأبعاد باستخدام PCA

تُعد بيانات الصور ذات أبعاد عالية، مما قد يؤثر سلبًا على أداء النماذج التقليدية. لذلك تم استخدام **تحليل المكونات الرئيسية (PCA)**.

- تم الاحتفاظ بـ **95% من التباين الكلي**.
- تم تقليل عدد السمات من **10,000** فقط **إلى 137 سمة**.

لماذا PCA؟

- تقليل التكلفة الحسابية
- إزالة السمات الزائدة والضوضاء
- معالجة مشكلة لعنة الأبعاد
- تحسين القدرة على التعميم

---

6. الانحدار اللوجستي (تعلم مُراقَب)

6.1 وصف النموذج

الانحدار اللوجستي هو خوارزمية تعلم مُراقَب خطية تُستخدم في مهام التصنيف. وقد استُخدم في هذا المشروع كنموذج أساسي (Baseline) لتقييم قدرة النماذج التقليدية على تصنيف الصور بعد تقليل الأبعاد.

6.2 التدريب

تم تدريب النموذج باستخدام:

- السمات الناتجة من PCA
- StandardScaler مطبق فقط على بيانات التدريب
- عدد تكرارات كافٍ لضمان التقارب

6.3 معايير التقييم

- الدقة (Accuracy)
- Precision

- Recall
- F1-score
- مصفوفة الالتباس (Confusion Matrix)

6.4 النتائج

حقق نموذج الانحدار اللوجستي:

- **دقة ≈ 99.8%**
- خطأ تصنيفي واحد فقط في مجموعة الاختبار

ويُعزى هذا الأداء العالي إلى طبيعة البيانات المُتحكم فيها وسهولة الفصل الخطي بين الفئات بعد المعالجة.

---

7. خوارزمية K-Means (تعلم غير مُراقَب)

7.1 الهدف

تم استخدام K-Means لاستكشاف البنية الداخلية للبيانات بدون استخدام التصنيفات الحقيقية.

7.2 المنهجية

- تم تطبيق K-Means على السمات الناتجة من PCA
- عدد العناقيد يساوي عدد الفئات (5)

7.3 التقييم

نظرًا لأن K-Means خوارزمية غير مُراقَبة، لا يُعد استخدام الدقة مناسبًا. لذلك تم استخدام **Silhouette Score**.

7.4 النتائج

- **Silhouette Score ≈ 0.28**

وهي قيمة مقبولة في حالة بيانات الصور نظرًا للتشابه البصري بين الفئات المختلفة.

---

8. المقارنة بين الانحدار اللوجستي و K-Means

المقارنة	الانحدار اللوجستي	K-Means
نوع التعلم	مُراقَب	غير مُراقَب
استخدام التصنيفات	نعم	لا

المقارنة	الانحدار اللوجستي	K-Means
الهدف	تصنيف	تجميع
الأداء	مرتفع جدًا	متوسط
الاستخدام	تنبؤ دقيق	استكشاف البيانات

---

9. القيود

- استخدام البكسلات المسطحة لا يحافظ على العلاقات المكانية داخل الصورة.
- البيانات مُتحكم فيها ولا تمثل صور العالم الحقيقي.
- فقدان معلومات الألوان بسبب التحويل إلى Grayscale.

---

10. الأعمال المستقبلية

- استخدام سمات لونية أو واصفات يدوية
- تطبيق الشبكات العصبية الالتفافية (CNNs)
- اختبار النموذج على صور حقيقية بخلفيات معقدة
- تجربة تقنيات أخرى لتقليل الأبعاد

---

11. الخلاصة

أظهر هذا المشروع أن خوارزميات تعلم الآلة التقليدية يمكن أن تحقق أداءً عاليًا في تصنيف الصور عند دمجها مع المعالجة المسبقة المناسبة وتقنيات تقليل الأبعاد. حقق الانحدار اللوجستي نتائج ممتازة بعد استخدام PCA وفرت K-Means ، بينما فهمًا جيدًا للبنية الداخلية للبيانات. تؤكد النتائج أهمية التعلم المُراقَب في مهام التصنيف وتمثل أساسًا قويًا للانتقال إلى تقنيات التعلم العميق مستقبلًا.