IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, December 2025
Impact Factor: 7.67

Comparative Study of Supervised and Unsupervised Learning

Gaikwad Vaishnavi Tribhuvan[1] and Rohini S. Kapse[2]
M.Sc. Computer Science, K.T.H.M College, Nashik[1]
Professor, Department Computer Science and Application, K.T.H.M College Nashik[2]

Abstract: *The field of Artificial Intelligence and Machine Learning is growing very fast and touching almost every industry today. This project work compares the two main approaches of machine learning – Supervised Learning and Unsupervised Learning. Supervised learning needs labeled data and gives very accurate predictions, so it is mostly used for classification and regression tasks. On the other side, unsupervised learning works on data without any labels and tries to find hidden patterns, natural groups or reduce dimensions using clustering and other methods.*

In this study, popular algorithms like Decision Tree, Support Vector Machine, Logistic Regression for supervised part and K-Means, Hierarchical Clustering and PCA for unsupervised part were implemented on standard benchmark datasets (Iris, Breast Cancer Wisconsin and Mall Customers dataset). The experiments were carried out in Python using scikit-learn library

Results clearly show that supervised models achieved high accuracy between 95–97%, whereas unsupervised models gave Silhouette Score in the range of 0.6–0.7 which is quite decent for pattern discovery. When labeled data is available, supervised learning is definitely better, but when labeling is costly or not possible, unsupervised methods become very useful. The study also suggests that future systems can get even better results by combining both approaches in hybrid or semi- supervised models.

Keywords*:* Supervised Learning, Unsupervised Learning, Classification, Clustering, K-Means, SVM, Decision Tree, PCA, Machine Learning Comparison

I. INTRODUCTION

The field of Artificial Intelligence and Machine Learning has grown very fast in the last few years. Today, almost every industry, whether it is medical diagnosis, online shopping, banking fraud detection or self-driving cars, is using some form of machine learning to solve problems and take better decisions. Among all the techniques available, two basic categories are used most widely: Supervised Learning and Unsupervised Learning.

In supervised learning, we train the model using data that already has correct answers (labels) attached to each record. The model learns the relationship between the input features and the given output. Once trained, it can predict the output for completely new data. Common examples are classifying an email as spam or not spam, predicting whether a patient has a disease or not, or estimating the price of a house.

In unsupervised learning, there is no correct answer provided. The algorithm looks at the raw data and tries to find natural groups, patterns or hidden structures on its own. It is very useful when we have a large amount of data but labeling each record manually is costly or impossible. Typical uses are customer segmentation for marketing, grouping similar news articles, or finding unusual transactions in credit card data.

Both approaches are important, but they solve different kinds of problems. Many students and even professionals get confused about when to use which method, how much performance they can expect, and how to explain the results to others. Very few academic projects actually implement both types of algorithms on the same system and compare them side by side with proper charts and numbers.

This project is an attempt to fill that gap. We have implemented popular supervised algorithms (Decision Tree, Support Vector Machine, Logistic Regression) and popular unsupervised algorithms (K-Means, Hierarchical Clustering, PCA)

Copyright to IJARSCT
www.ijarsct.co.in

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 1, December 2025 Impact Factor: 7.67

on standard benchmark datasets. We then measured their performance using suitable metrics and presented the results through tables and graphs so that the differences become clearly visible.

Objectives:

I. To study and implement three supervised learning algorithms on labeled datasets (Iris and Breast Cancer Wisconsin).

II. To study and implement three unsupervised learning algorithms on an unlabeled real-world dataset (Mall Customer Segmentation).

III. To evaluate supervised models using accuracy, precision, recall and F1-score, and unsupervised models using Silhouette Score, Davies-Bouldin Index and variance explained.

IV. To compare both approaches in terms of accuracy, speed, interpretability and practical use-cases.

V. To draw clear conclusions and suggest when to choose which method in real-life projects.

II. LITERATURE SURVEY

A lot of research has already been done on both supervised and unsupervised learning separately, but very few papers directly compare them on the same system with actual code and results.

Several authors have studied supervised learning techniques in detail. Zhang et al. (2018) worked on Decision Trees, SVM and k-NN and showed that when good-quality labeled data is available, these algorithms give very high accuracy in medical diagnosis and image classification tasks. Similarly, Patel and Thakur (2020) tested Logistic Regression and SVM on different datasets and reported accuracy above 90% in most cases. They pointed out that SVM works especially well when classes are clearly separable.

On the unsupervised side, Khan et al. (2019) compared K- Means and Hierarchical Clustering on large datasets and found that K-Means is faster while Hierarchical gives better visual understanding through dendrograms.

Liu and Wang (2021) used PCA for dimensionality reduction and showed that it can keep more than 90% of the information while removing noise – very useful before applying any clustering algorithm.

Some researchers have tried direct comparison. Singh et al. (2022) took classic datasets like Iris and Wine and applied both types of algorithms. They concluded that supervised models always beat unsupervised ones in prediction accuracy, but unsupervised methods are better when we just want to explore the data without any labels. Chawla (2020) talked about semi-supervised learning – using a small labeled portion and a large unlabeled portion together – and showed that performance improves a lot when labeling the entire data is costly.

After reading these papers, we noticed the following gaps:

- Most studies use different machines, different versions of libraries, or different datasets, so fair comparison is difficult.
- Many papers only show theoretical advantages/disadvantages instead of actual running time and scores.
- Very few student-level projects show complete code, graphs and tables side by side.

Because of these reasons, we decided to implement everything on the same laptop using scikit-learn, run all algorithms on standard benchmark datasets, record exact accuracy and clustering scores, and present everything through clear tables and charts. This way the differences become visible to anyone – even to those who are new to machine learning.

III. DATASETS AND PRE-PROCESSING

Datasets Used

We selected three standard, publicly available datasets so that anyone can download and repeat our experiments.

- iris dataset has three flower species – very clean and small, perfect for beginners.
- Breast Cancer dataset tells whether a tumor is malignant or benign – real medical data with 30 measurements.
- Mall Customers dataset contains Age, Annual Income, Spending Score, Gender – no labels given, ideal for customer segmentation.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 1, December 2025 Impact Factor: 7.67

Data Cleaning and Pre-processing Step

We followed the same steps for all datasets so the comparison remains fair.

1. Loaded the data using pandas
2. Checked for missing values – luckily none of the chosen datasets had missing entries.
3. For Mall Customers dataset, converted Gender (Male/Female) to numbers (0/1) using LabelEncoder.
4. Scaled all numerical features using StandardScaler() from scikit-learn because algorithms like SVM and K- Means are sensitive to different scales (e.g., income is in thousands, age is 18–70).
5. For supervised datasets, separated features (X) and target label ((y).
6. Split the labeled datasets into 70% training and 30% testing using train_test_split (random_state=42 for reproducibility).

IV. METHODOLOGY AND EXPERIMENTAL SETUP

Supervised Learning Algorithms Implemented

We used three popular and easy-to-understand algorithms from scikit-learn:

I. Decision Tree Classifier

• Works by repeatedly splitting the data based on feature values
• Very easy to visualize and explain
• Code used:

```
from sklearn.tree import DecisionTreeClassifier
dt = DecisionTreeClassifier(max_depth=5, random_state=42) dt.fit(X_train, y_train)
```

II. Support Vector Machine (SVM)

• Finds the best possible line/plane that separates classes with maximum margin
• Works really well on small-to-medium datasets
• Code used:

```
from sklearn.svm import SVC
svm=SVC(kernel='rbf', C=1.0, random_state=42) svm.fit(X_train, y_train)
```

III. Logistic Regression

• Though the name says regression, it is used for binary/multiclass classification
• Gives probability scores along with prediction
• Code used:

```
from sklearn.linear_model import LogisticRegression
lr = LogisticRegression(max_iter=200,random_state=42) lr.fit(X_train, y_train)
```

All models were trained using 10-fold cross-validation and final accuracy was checked on the 30% test set

Unsupervised Learning Algorithms Implemented

Unsupervised Learning Algorithms Implemented

I. K-Means Clustering

• Partitions data into k clusters by minimizing distance to cluster centers
• We used Elbow method and Silhouette score to choose best k (k=5 for Mall dataset)
• Code:

```
from sklearn.cluster import KMeans kmeans = KMeans(n_clusters=5, random_state=42)
clusters = kmeans.fit_predict(X_scaled)
```

II. Hierarchical Clustering

• Builds a tree of clusters (dendrogram)

Copyright to IJARSCT
www.ijarsct.co.in
DOI: 10.48175/568

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 1, December 2025 Impact Factor: 7.67

• No need to specify number of clusters beforehand
• Code:

```
from sklearn.cluster import AgglomerativeClustering
hc = AgglomerativeClustering(n_clusters=5, linkage='ward') hc_labels = hc.fit_predict(X_scaled)
```

III. Principal Component Analysis

• Reduces dimensions while keeping maximum variance
• We reduced Mall dataset from 5 features to 2 for easy plotting
• Code:

```
from sklearn.decomposition import PCA pca = PCA(n_components=2)
X_pca = pca.fit_transform(X_scaled)
print("Variance  explained:",pca.explained_variance_ratio_.sum())
```

Hardware and Software Environment

Laptop: Intel Core i5 8th Gen, 8 GB RAM, Windows 10
Python 3.9, Jupyter Notebook
Libraries: scikit-learn 1.3, pandas, numpy, matplotlib, seaborn

V. RESULTS AND DISCUSSION

After running all the algorithms, we got the following results.
Supervised Learning Results

Algorithm	Dataset Used	Performance / Observation
SVM	Iris, Breast Cancer	Highest accuracy (95–97%); handles
Decision Tree	Iris, Breast Cancer	Good accuracy (93–97%); tends to overfit on small/noisy data
Logistic Regression	Iris, Breast Cancer	Stable results for binary tasks; 95.91–97% accuracy

Unsupervised Learning Results

Algorithm	Dataset Used	Performance / Observation
K-Means Clustering	Customer Segmentation, Mall Data	Silhouette Score 0.6–0.7; well- separated clusters
Hierarchical Clustering	Customer Segmentation, Mall Data	Clear grouping via dendrograms; slower on large datasets
PCA	Customer/Mall datasets	Reduced dimensions; preserved 90–95% variance

Comparative Observation Table

Aspect	Supervised Learning	Unsupervised Learning
Data Requirement	Labeled data	Unlabeled data
Accuracy	High (95–97%)	Moderate (Silhouette Score 0.6–0.7)
Computation	Moderate	Low to Moderate
Interpretability	Easy to interpret	Needs graphs/dendrograms for clarity
Applications	Prediction, classification	Clustering, pattern identification

VI. CONCLUSION

We carried out this project to see the real difference between supervised and unsupervised learning by running actual code on standard datasets. The results are very straightforward – whenever we have proper labels, supervised algorithms like SVM and Decision Tree give excellent accuracy (95–97%) and are perfect for prediction work. But

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 1, December 2025 Impact Factor: 7.67

when labels are missing, K-Means, Hierarchical Clustering and PCA still do a decent job of finding natural groups and reducing dimensions with Silhouette scores around 0.6–0.7 and keeping 90–95% of the information.

So, supervised learning is the clear choice for accurate classification and regression tasks, while unsupervised learning shines in exploration, customer segmentation and anomaly detection. Both approaches are equally important and work even better when combined. This hands-on comparison made everything much clearer than just reading theory, and we feel confident now about choosing the right method for any future project.

VII. FUTURE DIRECTION

Next we want to build semi-supervised models that use only a few labeled records with lots of unlabeled data, try t-SNE/UMAP instead of simple PCA, and move to deep learning (Autoencoders + Neural Networks) on bigger real-world datasets from healthcare and banking. Deploying the final model as a web app would also be a good target.

REFERENCES

[1] Chawla, N. V. (2020). Data mining for imbalanced datasets: An overview. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(2), e1350. https://doi.org/10.1002/widm.1350

[2] Han, J., Kamber, M., & Pei, J. (2022). Data Mining: Concepts and Techniques (4th ed.). Morgan Kaufmann.

[3] Khan, M., Anwar, S., & Naqvi, S. (2019). A comparative analysis of clustering techniques for large datasets. International Journal of Advanced Computer Science and Applications (IJACSA), 10(8), 45–52. https://doi.org/10.14569/IJACSA.2019.010086

[4] Liu, Y., & Wang, H. (2021). Performance evaluation of unsupervised machine learning algorithms for pattern recognition. Journal of Artificial Intelligence and Soft Computing Research, 11(3), 187–198. https://doi.org/10.2478/jaiscr-2021-0012

[5] Patel, D., & Thakur, P. (2020). A study on performance evaluation of supervised machine learning algorithms for classification. International Journal of Computer Applications, 176(35), 12–18. https://doi.org/10.5120/ijca2020920921

[6]Singh, A., Kumar, R., & Gupta, P. (2022). Comparative performance analysis of supervised and unsupervised learning algorithms using benchmark datasets. International Journal of Scientific Research in Computer Science, Engineering and Information Technology,8(2),130–138. https://doi.org/10.32628/CSEIT228222

[7]Zhang, C., Bengio, S., Hardt, M., & Recht, B. (2018). Understanding deep learning requires rethinking generalization. Proceedings of the International Conference on Learning Representations (ICLR). https://doi.org/10.48550/arXiv.1611.03530