Internship Project Report

Mahindra University

INTERNSHIP REPORT

on

"Machine Learning"

(Machine Data Clustering Using K-Means and DBSCAN)

Submitted By:

M. Mani Krishna

Sofiya Sultana

Submitted to:

Mahindra University

Under the Guidance of

Prof. Dr. Arun k Pujari

HOD of AI & CSE, Adviser & Professor Emeritus

Assistant Professor

Tauheed Ahmed

Shabnam Samima

Duration:19-05-2025 To 02-07-2025

2nd July 2025

DECLARATION

I hereby declare that the presented report entitled
"Clustering and Classification Techniques on Labelled Data Sets"
carried out at **Mahindra University**, has been uniquely prepared by me and my team mate.

This report is submitted in partial fulfilment of the requirements for the award of the **M.Sc. in Data Science** as part of the **Internship** at **Mahindra university**

I also declare that this report has not previously formed the basis for the award of any degree, diploma, associateship, fellowship, or any other similar title of any university or institute.

(Signature)

Mani krishna

Sofiya Sultana

MSc. Data science

Place: Hyderabad

Certificate of Approval

(From Faculty Mentor)

This is to certify that **Mr. M. Mani Krishna**, Roll Number **2024067227 and Sofiya Sultana ,**Roll Number **2024100116**, are students of **Gitam University**, has successfully carried out the dissertation work presented in this report titled **“Clustering and Classification Techniques on Labelled Datasets”** for the award of **M.Sc. in Data Science** for the academic batch **2023–25**, under my supervision and guidance.

Prof. Dr. Arun k Pujari

Date: July 2nd 2025

Acknowledgment

Summer internship training is a crucial component of the **M.Sc. in Data Science** program, and I am grateful to have had the opportunity to undertake my project at **Mahindra University**.

It gives me immense pleasure to present this project report on
“Clustering and Classification Techniques on Labelled Datasets.”
This project would not have been possible without the valuable guidance, support, and encouragement of several individuals, to whom I express my heartfelt gratitude.

I am deeply thankful to my industry mentor at **Tauheed Ahmed** and **Shabnam Samima** for providing guidance, insights, and constant support throughout the internship period. Their advice and encouragement played a significant role in shaping this project.

I would also like to extend my sincere thanks to my faculty mentor **Motahar Reza** at **Gitam University** for their regular supervision and valuable feedback during the course of the project. Their expertise and constructive suggestions have greatly contributed to the successful completion of this work.

I am grateful to all the staff and members of **Mahindra University** who provided their time, cooperation, and support during my internship.

Finally, I express my deepest gratitude to my family and friends for their encouragement and moral support throughout this journey.

Thanking you,

M. Mani Krishna
Roll No.: 2024067227,

Sofiya Sultana

Roll No.: 2024100116

Abstract

This project focuses on the application of clustering and classification techniques on labelled datasets with the aim of uncovering hidden patterns, improving data-driven decision-making, and providing actionable insights. Clustering was performed using **K-Means**, **Co-occurrence Matrix**, and **Consensus Clustering** to identify natural groupings and relationships within the data. These methods enabled the detection of underlying structures, similarities, and anomalies that are not immediately visible through conventional analysis.

For classification, the **Decision Tree** algorithm was utilized to predict class labels, identify key decision rules, and interpret the contribution of different features in determining outcomes. This approach provided transparency and interpretability, making it easier to understand the decision boundaries and classification logic applied to the dataset.

The project involved extensive data pre-processing, including handling of missing values, normalization, and feature selection to enhance model performance. The implementation was carried out using Python-based frameworks, including **Pandas**, **Scikit-learn**, and **Matplotlib**, which facilitated efficient data manipulation, model development, and result visualization.

The performance of the models was evaluated using standard metrics such as **accuracy**, **precision**, **recall**, and **F1-score** to assess their effectiveness and robustness. The integration of clustering and classification techniques proved valuable in applications such as fault detection, customer segmentation, and predictive maintenance, demonstrating the versatility and power of machine learning approaches in addressing real-world problems.

This work was conducted as part of the **M.Sc. Data Science internship at Mahindra University**, contributing to the practical understanding and application of advanced data science methodologies.

TABLE OF CONTENT

Project Report: Clustering and Classification Techniques on Labelled Datasets

1. Introduction

This project explores various machine learning techniques, focusing primarily on clustering methods and decision tree classification across several labeled datasets. The study begins with fundamental preprocessing and data cleaning, followed by clustering using K-Means, analysis through co-occurrence and co-consensus matrices, and finally concludes with decision tree modeling. The datasets used include:

- Iris Dataset
- Glass Classification Dataset
- Wine Quality Dataset
- Cancer Dataset
- Car Evaluation Dataset
- Yeast Dataset
- Video Game Dataset
- Plant Communication Dataset

The objective is to understand how clustering can be applied to labeled data and to compare it with supervised learning techniques like decision trees.

2.Vision

Our vision is to be a leading contributor in the field of Data Science and Artificial Intelligence, driving innovation in machine learning solutions that empower organizations to make data-driven decisions. We strive to deliver impactful insights through advanced clustering and classification techniques, fostering a future where intelligent data analysis shapes industries and enhances lives globally.

We aim to be recognized for providing high-quality, interpretable, and reliable machine learning solutions, while nurturing talent and promoting ethical, responsible AI practices.

3.Mission

Our vision is to become a trusted enabler of data-driven innovation, equipping learners and professionals with cutting-edge skills in Data Science, Machine Learning, and Artificial Intelligence, and contributing to industries through impactful, ethical, and interpretable solutions that advance society.

4.Internship Objectives

- To apply theoretical knowledge of data science and machine learning to solve real-world problems using clustering and classification techniques.

- To gain hands-on experience in working with labelled datasets and implement models such as K-Means, Co-occurrence Matrix, Consensus Clustering, and Decision Trees.
- To develop skills in data pre-processing, feature selection, model evaluation, and visualization using tools like Python, Scikit-learn, and Pandas.
- To enhance interpersonal, communication, and time management skills through collaboration and reporting during the internship.
- To understand professional ethics, data privacy, and the responsibilities associated with working in the field of data science.
- To contribute meaningfully to the internship organization by delivering insights that support data-driven decision-making.

Technology Stack

Python :

Python is a widely used high-level programming language for general-purpose programming, created by Guido van Rossum and first released in 1991. An interpreted language, Python has a design philosophy which emphasizes code readability (notably using whitespace indentation to delimit code blocks rather than curly braces or keywords), and a syntax which allows programmers to express concepts in fewer lines of code than possible in languages such as C++ or Java. The language provides constructs intended to enable writing clear programs on both a small and large scale. Python features a dynamic type system and automatic memory management and supports multiple programming paradigms, including objectoriented, imperative, functional programming, and procedural styles. It has a large and comprehensive standard library

NumPy

NumPy is a Python library used for working with arrays.It also has functions for working in domain of linear algebra, fourier transform, and matrices.It is an open source project and you can use it freely.NumPy stands for Numerical Python.In Python we have lists that serve the purpose of arrays, but they are slow to process.NumPy aims to provide an array object that is up to 50x faster than traditional Python lists.The array object in NumPy is called ndarray, it provides a lot of supporting functions that make working with ndarray very easy.Arrays are very frequently used in data science, where speed and resources are very important.Data Science is a branch of computer science where we study how to store, use and analyze data for deriving information from it.NumPy arrays are stored at one 12 | Page continuous place in memory unlike lists, so processes can access and manipulate them very efficiently.This behavior is called locality of reference in computer science.This is the main reason why NumPy is faster than lists. Also it is optimized to work with latest CPU architectures.NumPy is a Python library and is written partially in Python, but most of the parts that require fast computation are written in C or C++.

Pandas

Pandas is a Python library used for working with data sets.It has functions for analyzing, cleaning, exploring, and manipulating data. The name "Pandas" has a reference to both "Panel Data", and "Python Data Analysis" and was created by Wes McKinney in 2008.Pandas allows us to analyze big data and make conclusions based on statistical theories. Pandas can clean messy data sets, and make them readable and relevant. Relevant data is very important in data science. Data Science is a branch of computer science where we study how to store, use and analyze data for deriving information from it. Pandas gives you answers about the data. Like - Is there a correlation between two or more columns, average value, Max value, Min value. Pandas are also able to delete rows that are not relevant, or contains wrong values, like empty or NULL values. This is called cleaning the data.

Machine Learning

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves.

Applications of Machine Learning :

➢ Image Processing

▪ Optical Character Recognition (OCR)

▪ Self-driving cars

▪ Image tagging and recognition

➢ Robotics

▪ Industrial robotics

▪ Human simulation

➢ Data Mining

▪ Association rules

▪ Anomaly detection

▪ Grouping and Predictions

➢ Video games

▪ Pokémon

▪ PUBG

➢ Text Analysis

▪ Spam Filtering

▪ Information Extraction

▪ Sentiment Analysis

➢ Healthcare

▪ Emergency Room & Surgery

▪ Research

▪ Medical Imaging & Diagnostics

5. Data Preprocessing and Cleaning

Before applying any machine learning techniques, each dataset underwent systematic preprocessing:

- **Handling Missing Values**: Missing data entries were either filled using imputation strategies (mean/median/mode) or removed, depending on the dataset.
- **Encoding Categorical Features**: Label encoding or one-hot encoding was used as per the nature of the categorical columns.

- **Feature Scaling**: StandardScaler was used to normalize the data, ensuring fair distance-based clustering.
- **Exploratory Data Analysis (EDA)**: Summary statistics and visual plots (histograms, box plots) helped understand the feature distributions.

3. Dataset-wise Detailed Analysis

- Iris Dataset

Preprocessing

- No missing values were present in the dataset.
- StandardScaler was used to normalize the numerical features, ensuring equal weight in distance-based clustering.
- PCA was optionally applied for 2D visualization.

K-Means Clustering

- K-Means was applied with k=3 to reflect the three species.
- The resulting clusters aligned closely with the true labels, especially for the Setosa class which formed a distinct group.
- The cluster-to-class matrix confirmed that:
- Cluster 0 mostly matched Setosa.
- Clusters 1 and 2 split between Versicolor and Virginica due to their feature overlap.

DBSCAN Clustering

- Unsupervised clustering method
- Groups dense points, marks outliers as noise
- No need to predefine cluster count
- eps (radius): 0.5 (chosen via k-distance plot)
- min_samples**:** 5

Visualization (PCA-reduced 2D Plot)

- **Cluster 0:** Setosa (clear separation)

- **Cluster 1:** Mixed Versicolor/Virginica
- **Cluster -1:** Noise (outliers)

Clustering Stability via Co-Occurrence and Co-Consensus

- K-Means was run 10 times with different random seeds.
- A co-occurrence matrix was created to show how often each pair of samples was clustered together.
- This matrix was then treated as a new similarity-based feature set.
- Co-consensus clustering using K-Means on this transformed data improved the consistency, especially in separating overlapping classes.

Decision Tree Classification

- A decision tree classifier was trained on the full dataset.
- The tree used petal length as the first split, perfectly separating Setosa.
- The model achieved ~96% accuracy.
- Its rule-based structure made it highly interpretable, offering human-readable insights such as:
- "If petal length ≤ 2.45 → Setosa"
- "Else if petal width ≤ 1.75 → Versicolor"
- "Else → Virginica"

- ## Glass Classification Dataset

Data Preprocessing

- Features included refractive index and elemental content (e.g., Na, Mg, Al).
- Dataset was normalized using feature scaling.

Clustering

- Performed K-Means clustering with k=6 (as per number of glass types).

- Repeated the clustering 10 times to analyze consistency.
- Constructed co-occurrence and co-consensus matrices.
- Plotted clusters using principal components for visualization.
- Co-consensus clusters revealed clearer separation than initial K-Means results.

Decision Tree

- Decision tree classifier applied for multiclass prediction of glass types.
- Tree used features like Mg and Ca to split data.
- Accuracy was lower due to high intra-class similarity.
- Nonetheless, the model generated interpretable classification rules.
- ## Car Evaluation Dataset

Preprocessing

- Categorical columns (buying, main, doors, persons, safety) encoded using one-hot.
- Dataset transformed into binary matrix and normalized.

Clustering

- K-Means applied with k=3, aligning with number of class labels.
- Scatter plots and PCA used for 2D visualization.
- Cluster-to-class confusion showed weak initial alignment.
- Co-occurrence and co-consensus analysis improved match for high and low safety classes.

Decision Tree

- Model classified cars into acceptability levels.
- Accuracy exceeded 90% with clear rules like:
- If safety = high and price = low → vgood
- If safety = low → unacc
- Tree structure easy to interpret, consistent with domain expectations.

- **Video Game Dataset**

Preprocessing

- Missing values in critic/user scores imputed.
- Categorical features like platform and genre were one-hot encoded.
- Normalization done on all numerical features.

Clustering

- K-Means clustering applied to identify groups based on score and genre.
- Ten random seed runs revealed variation in groupings.
- Co-occurrence matrix showed stable clusters for top-selling games.
- Co-consensus clustering grouped critically and commercially successful titles accurately.

Decision Tree

- Classifier trained to predict game success (e.g., hit or flop).
- Key features included critic score, platform, and genre.
- Tree generated clear decision paths.

- **Yeast Dataset**

Preprocessing

- Data was scaled using StandardScaler.
- Class imbalance handled via stratified sampling.
- Multi-class target (protein localization sites).

Clustering

- K-Means clustering performed with k=10 based on domain knowledge.
- Visual inspection showed overlapping clusters.
- Co-occurrence matrix revealed which points were consistently clustered.
- Co-consensus clustering improved consistency over runs.

Decision Tree

- Learned a moderately accurate classifier.
- Tree used nuclear localization signals and sequence motifs.
- Rules were helpful for protein classification despite moderate accuracy .
- **Plant Communication Dataset**

Preprocessing

- Cleaned and standardized features (e.g., light, pH, temperature).
- Target labels represented categories of plant signaling behavior.

Clustering

- K-Means clustering was run with different values of k, most stable around k=4.
- Scatter plots revealed meaningful groupings in behavior based on physiological traits.
- Co-occurrence matrix built over 10 K-Means runs to examine pairwise consistency.
- Co-consensus clustering produced clearer, reproducible groupings.

Decision Tree

- Classifier trained on the labeled signaling behavior.
- Model rules showed light intensity and salinity as key indicators.
- Achieved moderate to high accuracy depending on environmental variability.

- **Wine Quality Dataset**

Data Preprocessing

- Dataset included physicochemical properties of wines.
- Standardization of features was done.
- Quality scores (from 3 to 8) were treated as class labels for classification tasks.

Clustering

- Initial K-Means clustering with k=6, followed by performance comparison using a confusion matrix.
- Due to label overlap, clustering accuracy was moderate.
- K-Means was executed with 10 different seeds.
- Co-occurrence matrix captured consistency across seeds.
- Co-consensus clustering showed slight improvement in label alignment.
- Scatter plots of features like alcohol vs. pH visualized natural groupings.

Decision Tree

- Decision Tree model trained to classify wine quality.
- Tree depth and rules indicated alcohol, volatile acidity, and sulphates were key features.
- Performance was moderate due to class imbalance and overlapping quality categories.
- ## Cancer Data (Denmark)

Data Preprocessing

- Cleaned missing values and standardized features using scaling.
- Exploratory Data Analysis was conducted to understand feature distributions.

Clustering

- K-Means applied with k=6 (reflecting benign and malignant classes).
- Scatter plots showed visually separable clusters.
- Clustering was run 10 times with different seed points to assess stability.
- A **Cluster-Class Matrix** was generated to map clusters to real labels.
- A **Co-Occurrence Matrix** was built based on how often each data point pair was clustered together across runs.
- Treated co-occurrence values as features and re-applied K-Means (called **Co-Consensus Clustering**), again using 10 seeds.

- Co-consensus clustering improved the consistency of cluster assignments.

Decision Tree

- A decision tree was trained to classify cancer status.
- Rules extracted from the tree made clinical interpretation easy (e.g., “if radius > X and concavity < Y → Malignant”).
- Model demonstrated high accuracy and clear interpretability.

6. Observations and Insights

- **Iris & Cancer datasets** had clear, well-separated clusters and high decision tree accuracy .
- **Wine, Glass, and Yeast datasets** showed overlapping classes. Clustering was noisy but improved using **co-consensus**, while decision tree accuracy was moderate .
- **Car Evaluation** and **Video Game datasets** required encoding but produced meaningful clusters and **interpretable rules** from decision trees .
- **Plant Communication dataset** showed stable clusters and key features like light and temperature influencing classification .
- **Co-occurrence and co-consensus methods** made clustering more reliable, especially when results varied with random seeds.
- **Decision trees consistently offered interpretable, rule-based models**, making them valuable even on complex datasets.

Tasks & Responsibilities:

- Develop clustering models such as K-Means, Co-occurrence Matrix, and Consensus Clustering to identify meaningful patterns and groupings within labelled datasets.

- Implement classification algorithms — primarily Decision Trees — to class labels and evaluate feature importance.
- Perform data pre-processing and feature selection, including handling missing values, normalizing data, and selecting relevant attributes for model training.
- Compare and evaluate multiple machine learning models based on performance metrics like accuracy, precision, recall, and F1-score to determine the most effective techniques.
- Maintain codebase at an industry-level standard, ensuring readability, modularity, and documentation for reproducibility.
- Generate visualizations (e.g., cluster plots, tree diagrams) and reports to communicate findings effectively to stakeholders.
- Collaborate with mentors and team members to incorporate feedback, refine models, and align project outcomes with organizational goals.

Conclusion

This internship project successfully demonstrated the application of clustering and classification techniques on labelled datasets, contributing to both technical learning and professional development. The use of **K-Means clustering**, combined with **Co-occurrence Matrix** and **Consensus Clustering**, allowed for meaningful grouping of data points and provided insights into underlying patterns that were not immediately visible. The **Consensus Clustering** approach enhanced the stability and reliability of clustering outcomes by reducing dependence on random initializations, especially in datasets where clear class boundaries were not easily defined.

The **Decision Tree classification** provided an interpretable model for predicting class labels and understanding the contribution of various features. The tree's ability to generate clear decision rules supported better analysis of feature importance and decision logic, offering transparency in model behavior.

From a personal and professional perspective, this internship provided valuable hands-on experience in solving real-world data science problems. It helped me strengthen my skills in machine learning, data pre-processing, model evaluation, and visualization. Beyond technical growth, the internship enhanced my communication skills, time management, and problem-solving abilities, preparing me for future roles in the corporate data science environment.

Overall, this project highlighted the power of integrating clustering and classification techniques to derive both exploratory and predictive insights, and underscored the importance of rigorous methodology in machine learning workflows. The experience gained during this internship at **Mahindra University** will serve as a solid foundation for my future career in data science.

References:

All Content used in this report is from

https://www.simplilearn.com/

https://www.wikipedia.org/

https://towardsdatascience.com/

https://www.expertsystem.com/

https://www.coursera.org/

https://www.edureka.co/

https://subhadipml.tech/

https://www.forbes.com/

https://medium.com/

https://www.google.com/

Book, I referred are

- Hands-on Machine Learning with Scikit-learn & TensorFlow By Aurelien Geron
- Python Machine Learning by Sebastian Raschk