PREDICTIVE MODELING OF IRIS FLOWER SPECIES

Project Report

Submitted to the Faculty of Engineering of

JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY KAKINADA, KAKINADA

In partial fulfillment of the requirements for the award of the Degree of

BACHELOR OF TECHNOLOGY

In

COMPUTER SCIENCE AND ENGINEERING

By

M Neeraja
(21481A05D0)

Mohammad Aslam
(21481A05D2)

J T V R Narayana
(21481A0586)

K Hasrhith Ram
(21481A05A2)

Under the Enviable and Esteemed Guidance of

Dr. G. BHARATHI, M. Tech, Ph.D

Senior Scale Grade Assistant Professor, CSE

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

SESHADRI RAO GUDLAVALLERU ENGINEERING COLLEGE

(An Autonomous Institute with Permanent Affiliation to JNTUK, Kakinada)

SESHADRIRAO KNOWLEDGE VILLAGE

GUDLAVALLERU – 521356

ANDHRA PRADESH

2023-24

SESHADRI RAO GUDLAVALLERU ENGINEERING COLLEGE

(An Autonomous Institute with Permanent Affiliation to JNTUK, Kakinada)

SESHADRI RAO KNOWLEDGE VILLAGE, GUDLAVALLERU

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the project report entitled **"PREDICTIVE MODELING OF IRIS FLOWER SPECIES"** is a bonafide record of work carried out by M Neeraja (21481A05D0), Mohammad Aslam (21481A05D2), J T V R Narayana (21481A0586), K Harshith Ram (21481A05A2), under the guidance of Dr. G. BHARATHI**,** M.tech, phD, Senior Scale Grade Assistant Professor**,** ComputerScience and Engineering, in the partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering of Jawaharlal Nehru Technological University Kakinada, Kakinada during the academic year 2023-24.

Project Guide
(Dr. G. BHARATHI)

Head of the Department
(Dr. M. BABU RAO)

ACKNOWLEDGEMENT

The satisfaction that accompanies the successful completion of any task would be incomplete without the mention of people who made it possible and whose constant guidance and encouragements crown all the efforts with success.

We would like to express our deep sense of gratitude and sincere thanks to **Dr. G. BHARATHI, M.Tech, PhD, Senior Scale Grade Assistant Professor ,** Computer Science and Engineering, for her constant guidance, supervision and motivation in completing the project work.

We feel elated to express our floral gratitude and sincere thanks to **Dr. M. Babu Rao**, **Head of the Department**, Computer Science and Engineering for his encouragements all the way during analysis of the project. His annotations, insinuations and criticismsare the key behind the successful completion of the project work.

We would like to take this opportunity to thank our beloved principal **Dr. B. Karuna Kumar,** for providing a great support for us in completing our project and giving us the opportunity for doing project.

Our Special thanks to the faculty of our department and programmers of our computer lab. Finally, we thank our family members, non-teaching staff and our friends, who had directly or indirectly helped and supported us in completing our project in time.

Team Members:

Matta Neeraja (21481A05D0)

Mohammad Aslam (21481A05D2)

Jannu T V R Narayana (21481A0586)

Katragadda Harshith Ram(21481A05A2)

INDEX

List of Figures

ABSTRACT

This project investigates the classification of Iris flowers using various techniques. We are Understanding the intricate structures of Iris organs like sepals and petals contributes significantly to botanical research and applications in various fields. This study delves into the quantitative analysis of sepal and petal characteristics to unravel their diverse roles in flowers physiology, ecology, and evolution. Utilizing data gathered from extensive fieldwork and laboratory experiments, we investigate the morphological, anatomical, and physiological attributes of sepals and petals across diverse plant species.

The analysis encompasses parameters such as size, shape, color, texture, venation patterns, and biochemical composition, shedding light on the underlying mechanisms governing floral development and function. Through advanced statistical techniques and machine learning algorithms, patterns and correlations within the data are elucidated, providing valuable insights into plant adaptation, pollination strategies, and ecosystem dynamics.

This research explores the implications of environmental factors, genetic variations, and evolutionary pressures on sepal and petal traits, highlighting their adaptive significance and potential applications in agriculture, horticulture, and biotechnology. By integrating multidisciplinary approaches, including genetics, physiology, ecology, and bioinformatics, this study advances our understanding of plant biology and fosters innovations for sustainable management and conservation of plant diversity in a changing world.

Through comprehensive experimentation and analysis, this project provides valuable insights into the effectiveness of various machine learning techniques for Iris flower classification, contributing to the broader understanding of data mining methodologies in real-world applications.

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

In the enchanting world of flowering plants, sepals and petals are not just mere adornments; they are the exquisite guardians of reproductive success and ecological adaptation. Sepals and petals, collectively known as floral organs, play pivotal roles in plant reproduction, pollinator attraction, and environmental interaction. Understanding the intricate structures and functions of these floral components unveils the captivating mechanisms driving plant biology and ecological relationships.

Sepals, typically found in the outermost whorl of a flower, serve as protective structures during bud development and contribute to floral aesthetics. Often green and leaf-like in appearance, sepals shield delicate floral organs from physical damage, desiccation, and herbivory. Beyond their protective role, sepals also participate in floral development, helping to regulate the opening of flower buds and providing structural support.

Petals, nestled within the sepals, represent nature's palette of colors, shapes, and fragrances, enticing pollinators with their alluring beauty. Adorned with pigments, patterns, and textures, petals serve as beacons for pollinators, guiding them towards the flower's reproductive organs. Moreover, petals play a crucial role in reproductive isolation and speciation, as variations in petal morphology can influence the preferences of specific pollinators, leading to the evolution of floral diversity..

Fig 1.1 Iris setosa

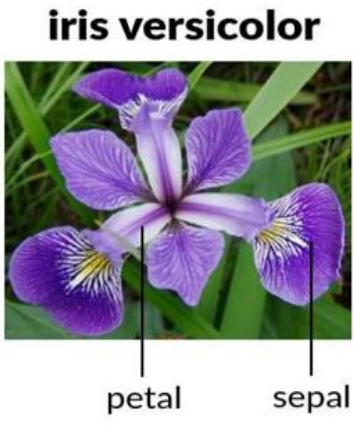


Fig 1.2 Iris versicolor

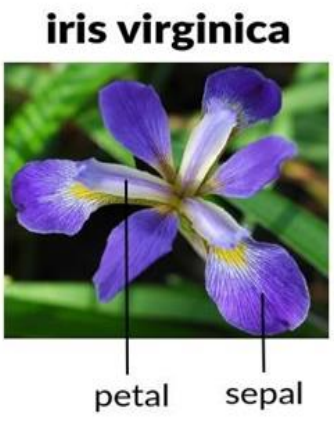


Fig 1.3 Iris virginica

CLASSIFICATION

Classification is a task in data mining that involves assigning a class label to each instance in a dataset based on its features. Using data mining techniques to classify flower sepals and petals involves extracting meaningful patterns and relationships from datasets containing information about their morphological characteristics. Here's a general outline of how this process could be approached:.

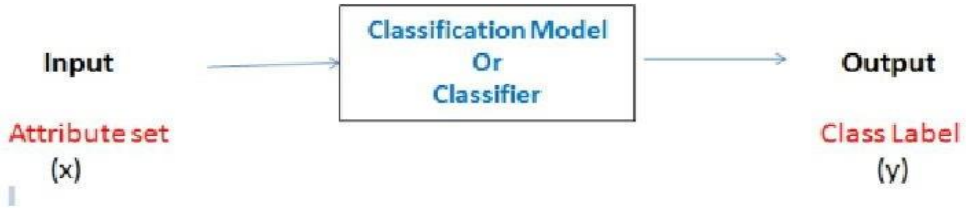


Fig 1.4 Classification Definition

classification of flowers typically involves using machine learning algorithms to categorize flowers based on their features. One of the most well-known examples is the classification of Iris flowers using the Iris dataset, introduced by Ronald Fisher in 1936.

Here's a step-by-step overview of how flower classification is typically carried out in data mining:

1. Data Collection
2. Data Preprocessing
3. Exploratory Data Analysis (EDA)
4. Feature Engineering
5. Model Selection
6. Model Training
7. Model Evaluation
8. Model Interpretation
9. Validation and Deployment

Data Collection:

Gather a comprehensive dataset containing measurements and features of sepals and petals from various flower species. This dataset should include attributes such as length, width, color, texture, shape, venation patterns, and any other relevant morphological traits.

Data Preprocessing:

Clean the dataset to handle missing values, outliers, and inconsistencies. Normalize or standardize the data to ensure that all features are on a similar scale. Additionally, consider encoding categorical variables and performing feature selection to reduce dimensionality and focus on the most informative attributes.

Exploratory Data Analysis (EDA):

Conduct exploratory data analysis to gain insights into the distribution and relationships within the dataset. Visualize the data using techniques such as scatter plots, histograms, and box plots to identify patterns and potential clusters.

Feature Engineering:

Extract or derive new features from the existing dataset that may enhance the classification task. This could involve calculating ratios, transformations, or combinations of existing features to capture additional information about sepals and petals.

Model Selection:

Choose appropriate data mining algorithms for classification based on the nature of the dataset and the objectives of the analysis. Common algorithms for classification tasks include decision trees, random forests, support vector machines (SVM), k-nearest neighbors (KNN), and neural networks.

Model Training:

Split the dataset into training and testing sets to train the classification model. Use techniques such as cross-validation to assess the performance of different algorithms and tune hyperparameters for optimal results.

Model Evaluation:

Evaluate the trained models using performance metrics such as accuracy, precision, recall, F1-score, and confusion matrix. Compare the performance of different algorithms to identify the most effective approach for classifying sepals and petals.

Model Interpretation:

Interpret the results of the classification model to understand the distinguishing characteristics and features that contribute to the classification of sepals and petals into different categories or species.

Validation and Deployment:

Validate the classification model using independent datasets or through real-world validation experiments. Once validated, deploy the model for practical applications such as species identification, biodiversity monitoring, or ecological research.

By leveraging data mining techniques, researchers can extract valuable insights from flower morphology data to classify sepals and petals accurately and efficiently, contributing to our understanding of plant diversity and evolutionary patterns.

Here's the list of Classification Models in detail:

1. Decision Trees:

Decision trees are versatile and interpretable models that recursively split the data based on feature thresholds. Each node in the tree represents a decision based on a feature, leading to a hierarchical structure. Decision trees are particularly useful for classification tasks involving categorical or ordinal data, making them suitable for classifying plant species based on discrete morphological features.

2. Random Forest:

Random Forest is an ensemble learning method that constructs multiple decision trees and combines their predictions through voting or averaging. It improves upon the performance of individual decision trees by reducing overfitting and increasing robustness. Random Forest can handle both classification and regression tasks and is suitable for datasets with a large number of features, making it applicable to complex morphological datasets of plant sepals and petals.

3. Support Vector Machines (SVM):

SVM is a powerful supervised learning algorithm that constructs a hyperplane to separate different classes in feature space. It works well for both linearly separable and nonlinearly separable data by using kernel functions to map the input features into higher-dimensional space .SVM is effective for classification tasks with high-dimensional feature spaces, making it suitable for plant sepal and petal classification based on multiple morphological attributes.

4. K-Nearest Neighbors (KNN):

KNN is a simple yet effective algorithm that classifies data points based on the majority class among their k nearest neighbors in feature space. It is a non-parametric method that does not require assumptions about the underlying data distribution. KNN is suitable for classification tasks where similar instances tend to belong to the same class, making it applicable to plant sepal and petal classification based on similarity in morphological features.

5. Neural Networks:

Neural networks, particularly deep learning architectures like convolutional neural networks (CNNs), can learn complex patterns from raw data. CNNs are well-suited for image-based classification tasks, where the input data consists of images of plant sepals and petals.

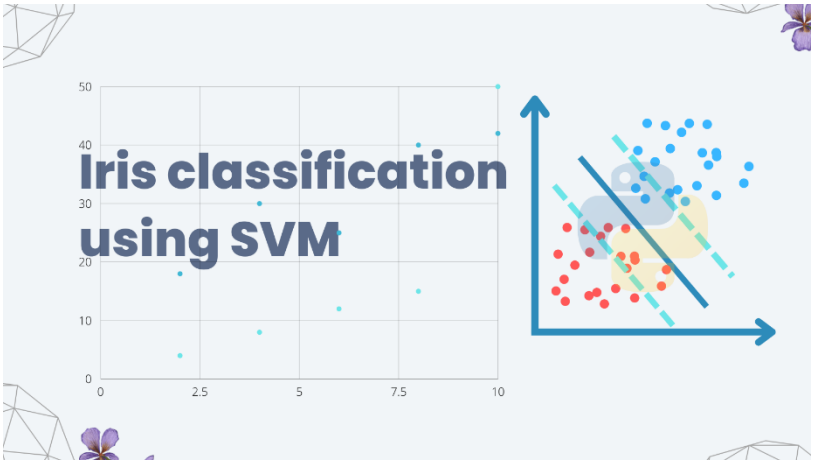


Fig 1.5 SVM Classification

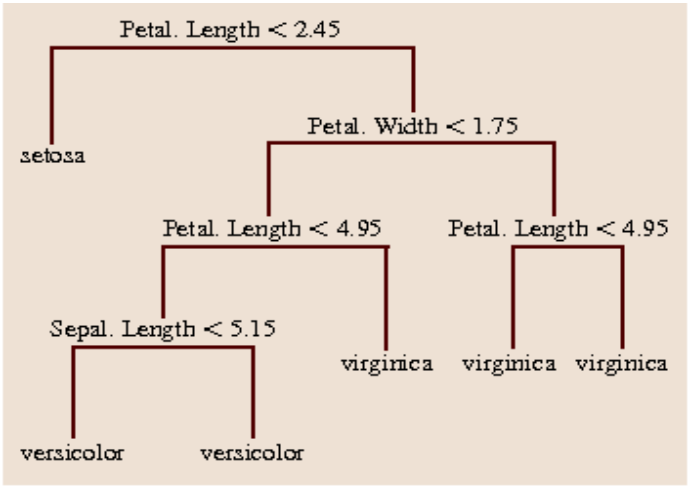


Fig 1.6 Decision trees

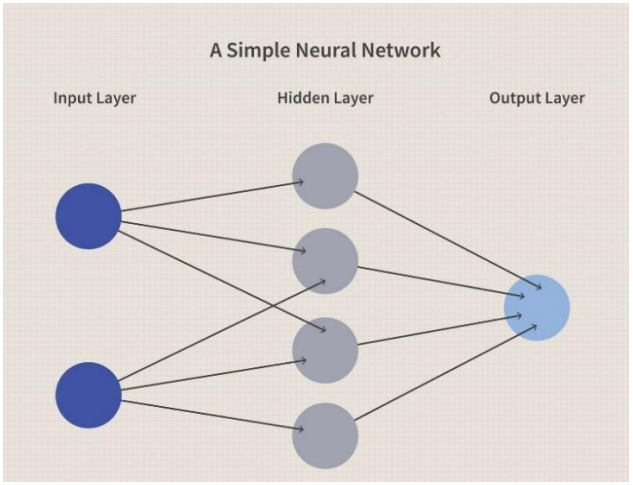


Fig 1.7 neural network

Model Training:

The fifth step in building a classification model is model training. Model training involves using the selected classification algorithm to learn the patterns in the data. The data is divided into a training set and a validation set. The model is trained using the training set, and its performance is evaluated on the validation set.

Model Evaluation:

The sixth step in building a classification model is model evaluation. Model evaluation involves assessing the performance of the trained model on a test set. This is done to ensure that the model generalizes well

Challenges and Limitations of Classification:

- **Overfitting**: One of the primary challenges in classification is overfitting, where a model learns to memorize the training data rather than generalize patterns
- **Data Quality and Quantity**: The quality and quantity of training data significantly impact the performance of classification models. Insufficient or noisy data can lead to inaccurate classifications.
- **Noisy or Missing Data:** Difficulty in handling noisy or missing data, which can impact themodel's ability to accurately learn patterns and make prediction.
- **Algorithm Selection:** Choosing the appropriate classification algorithm is challenging and depends on factors such as dataset size, class distribution, and problem complexity.
- **Interpretability**: Some classification models, particularly deep learning algorithms, lack interpretability, making it difficult to understand the reasoning behind their predictions.
- **Performance Degradation:** The performance of classification models may degrade over time due to concept drift, where the statistical properties of the data change, necessitating continuous model monitoring and adaptation.
- **Computational Resources**: Training and deploying complex classification models may require significant computational resources, including processing power, memory, and storage. This can be a limitation, particularly for resource-constrained environments or real-time applications where computational efficiency is crucial.

Applications of Classification:

- **Medical Diagnosis:** Classification is used to predict diseases based on symptoms, medical history, and diagnostic tests. For example, classifying whether a patient has a particular disease or not based on medical data.
- **Customer Segmentation:** Classification helps segment customers into different groups based on demographics, behavior, or purchasing patterns. This segmentation can be used for targeted marketing or personalized recommendations.
- **Credit Scoring:** Classification is employed in credit scoring to predict the creditworthiness of applicants based on factors such as income, credit history, and debt-to-income ratio.
- **Sentiment Analysis:** Classification is used to analyze text data and determine the sentiment expressed in reviews, social media posts, or customer feedback. It can classify text as positive, negative, or neutral.
- **Fraud Detection:** Classification is utilized in fraud detection to distinguish between legitimate and fraudulent transactions based on transactional data and behavioral patterns.
- **Image Recognition:** Classification is employed in image recognition to classify images into different categories or classes, such as identifying objects, animals, or people in images.
- **Spam Filtering:** Classification is used in email spam filtering to classify emails as either spam or legitimate based on their content and characteristics.
- **Predictive Maintenance:** Classification is utilized in predictive maintenance to predict equipment failures or malfunctions based on sensor data, usage patterns, and maintenance history.
- **Biometric Identification:** Classification is employed in biometric identification systems to classify individuals based on their unique biometric traits such as fingerprints, iris patterns, or facial features.

1.2 PROBLEM STATEMENT

The task is to develop a classification model for the Iris Flowers dataset. The dataset contains four features: sepal length, sepal width, petal length, and petal width, along with the class label specifying the species of Iris flower (Setosa, Versicolor, or Virginica). The goal is to build a model that can accurately classify Iris flowers based on their features.

Specific Objectives:

- Explore and preprocess the dataset: Analyze the distribution of features, handle missing values (if any), and perform any necessary feature scaling or normalization.

- Select appropriate machine learning algorithms: Experiment with various classification algorithms such as Decision Trees, Random Forests, Support Vector Machines (SVM), k-Nearest Neighbors (kNN), etc.

- Evaluate model performance: Use appropriate evaluation metrics such as accuracy, precision, recall, and F1-score to assess the performance of the models.

- Tune hyperparameters: Fine-tune the hyperparameters of the selected models to improve their performance.

- Validate the model: Validate the final model using techniques such as cross-validation to ensure its generalizability.

- Interpret the results: Analyze the feature importance and decision boundaries of the model to gain insights into the classification process.

- Deploy the model: Once satisfied with the performance, deploy the model for real-world applications such as automated species identification in botanical gardens or environmental monitoring projects.

By addressing these objectives, the aim is to develop a robust classification model that accurately identifies the species of Iris flowers based on their characteristics.

CHAPTER 2
PROPOSED METHOD

2.1 Methodology

The analysis of various data mining Classification algorithms, the algorithms that are used in this model are K nearest neighbors (KNN), Neural Network, and Decision Tree Classifiers which can be helpful for analyzing the new flowers.

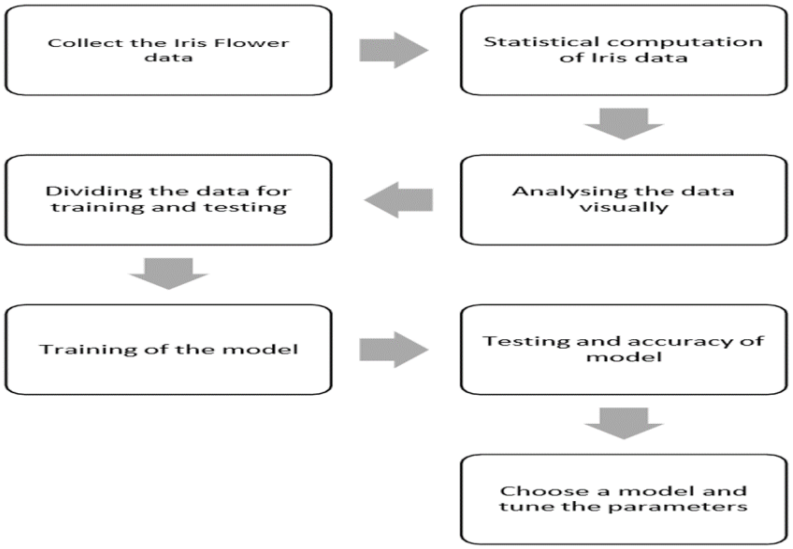


Fig 2.1 Methodology Block diagram

Classification Models*:* They include Naïve Bayes, Logistic Regression, Random Forest, Decision Tree, KNN, SVM.

Accuracy: Accuracy is Calculated and Compared and best one should be noticed.

Precision: It counts the number of predictions from the positive class that are actually in thatclass.

Recall: It calculates how many positive class predictions were made using all of the dataset'spositive examples.

F-Measure*:* It offers a single score that evenly weighs the issues of precision and recall.

Confusion Matrix*:* It is used to determine the classification models performance for a set oftest data.

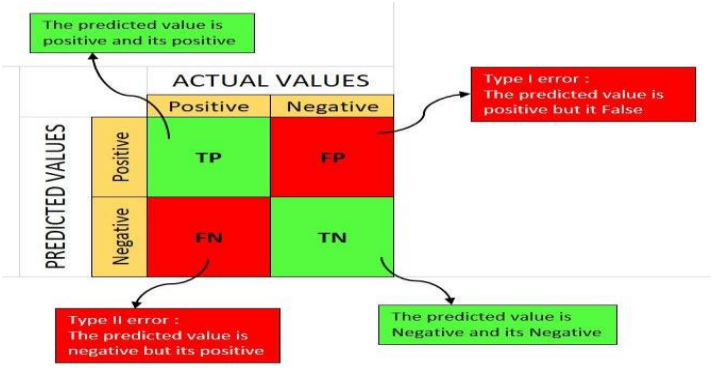


Fig 2.2 Confusion matrix

2.1.1 Block Diagram:

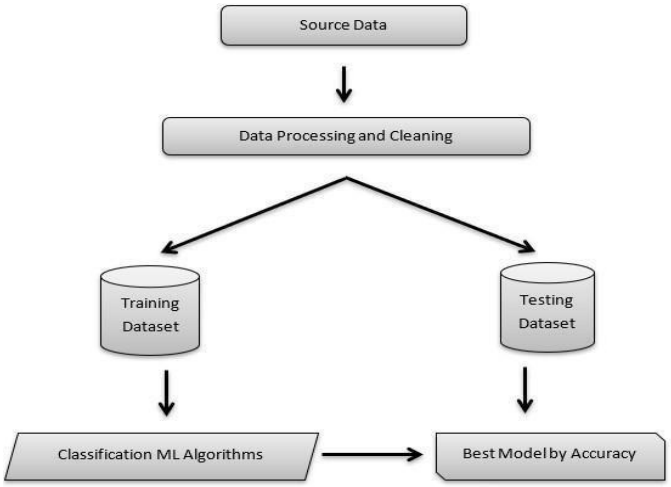


Fig 2.3 Block Diagram

Data Visualization*:* Here, Data visualization plays a crucial role in understanding the Iris dataset and extracting valuable insights about different species of iris flowers. Through scatter plots, box plots, histograms, and pair plots, we were able to visualize patterns, distributions, and relationships within the data. These visualizations not only enhance our understanding of the dataset but also facilitate better decision-making in model building and interpretation.

Features selection*:* Feature selection is crucial in data mining tasks like classification, where you aim to identify patterns and relationships within data to make accurate predictions. For the Iris dataset, which is commonly used for classification tasks, feature selection helps in identifying the most informative features that contribute to distinguishing between different species of iris flowers.

Techniques and their accuracy*:* Based on the evaluation results, the Neural Network demonstrates the highest accuracy among the three techniques, achieving an AUC of 1.000, Classification Accuracy (CA) of 0.933, F1 Score of 0.934, Precision of 0.947, Recall of 0.933, and Matthews Correlation Coefficient (MCC) of 0.904. This indicates its robust performance in discriminating between classes and overall predictive power. K-Nearest Neighbors (KNN) follows with an AUC of 0.976, CA of 0.867, F1 Score of 0.864, Precision of 0.875, Recall of 0.867, and MCC of 0.793, demonstrating respectable but slightly lower accuracy compared to the Neural Network. Support Vector Machine (SVM) also performs well, with an AUC of 0.938, CA of 0.933, F1 Score of 0.934, Precision of 0.947, Recall of 0.933, and MCC of 0.904, showing comparable performance to the Neural Network. Overall, based on these metrics, the Neural Network emerges as the top-performing technique in terms of accuracy and predictive capability

.

2.1.1 Algorithm and Explanation:

Algorithm: Generate_decision_tree. Generate a decision tree from the training tuples of data partition, *D*.

Input:

- Data partition, *D*, which is a set of training tuples and their associated class labels;
- *attribute_list*, the set of candidate attributes;
- *Attribute_selection_method*, a procedure to determine the splitting criterion that "best" partitions the data tuples into individual classes. This criterion consists of a *splitting_attribute* and, possibly, either a *split-point* or *splitting subset*.

Output: A decision tree.

Method:

```
(1)  create a node N;
(2)  if tuples in D are all of the same class, C, then
(3)      return N as a leaf node labeled with the class C;
(4)  if attribute_list is empty then
(5)      return N as a leaf node labeled with the majority class in D; // majority voting
(6)  apply Attribute_selection_method(D, attribute_list) to find the "best" splitting_criterion;
(7)  label node N with splitting_criterion;
(8)  if splitting_attribute is discrete-valued and
         multiway splits allowed then // not restricted to binary trees
(9)      attribute_list ← attribute_list − splitting_attribute; // remove splitting_attribute
(10) for each outcome j of splitting_criterion
     // partition the tuples and grow subtrees for each partition
(11)     let Dj be the set of data tuples in D satisfying outcome j; // a partition
(12)     if Dj is empty then
(13)         attach a leaf labeled with the majority class in D to node N;
(14)     else attach the node returned by Generate_decision_tree(Dj, attribute_list) to node N;
     endfor
(15) return N;
```

Fig 2.4 Decision tree algorithm

Decision tree induction is the method of learning the decision trees from the training set. The training set consists of attributes and class labels. Applications of decision tree induction include astronomy, financial analysis, medical diagnosis, manufacturing, and production. A decision tree is a flowchart tree-like structure that is made from training set tuples. The dataset is broken down into smaller subsets and is present in the form of nodes of a tree. The tree structure has a root node, internal nodes or decision nodes, leaf node, and branches. The root node is the topmost node. It represents the best attribute selected for classification. Internal nodes of the decision nodes represent a test of an attribute of the dataset leaf node or terminal node which represents the classification or decision label. The branches show the outcome of the test performed.

2.2 Data Preparation

2.2.1 Data Set Description

The Iris dataset is readily available in many machine learning libraries and repositories and is commonly used for educational purposes, benchmarking algorithms, and exploring various machine learning techniques. These features are used to classify the iris flowers into the three species based on their measurements. The dataset is often used for supervised learning tasks, where the goal is to train a model to accurately predict the species of an iris flower based on its measurements. The dataset consists of 150 samples of iris flowers, each belonging to one of three species: Setosa, Versicolor, and Virginica. For each sample, four features were measured:

1.Sepal length (in centimeters)

2.Sepal width (in centimeters)

3.Petal length (in centimeters)

4.Petal width (in centimeters)

	iris	sepal length	sepal width	petal length	petal width
1	Iris-setosa	5.1	3.5	1.4	0.2
2	Iris-setosa	4.9	3.0	1.4	0.2
3	Iris-setosa	4.7	3.2	1.3	0.2
4	Iris-setosa	4.6	3.1	1.5	0.2
5	Iris-setosa	5.0	3.6	1.4	0.2
6	Iris-setosa	5.4	3.9	1.7	0.4
7	Iris-setosa	4.6	3.4	1.4	0.3
8	Iris-setosa	5.0	3.4	1.5	0.2
9	Iris-setosa	4.4	2.9	1.4	0.2
10	Iris-setosa	4.9	3.1	1.5	0.1
11	Iris-setosa	5.4	3.7	1.5	0.2
12	Iris-setosa	4.8	3.4	1.6	0.2
13	Iris-setosa	4.8	3.0	1.4	0.1
14	Iris-setosa	4.3	3.0	1.1	0.1
15	Iris-setosa	5.8	4.0	1.2	0.2
16	Iris-setosa	5.7	4.4	1.5	0.4
17	Iris-setosa	5.4	3.9	1.3	0.4
18	Iris-setosa	5.1	3.5	1.4	0.3
19	Iris-setosa	5.7	3.8	1.7	0.3
20	Iris-setosa	5.1	3.8	1.5	0.3
21	Iris-setosa	5.4	3.4	1.7	0.2
22	Iris-setosa	5.1	3.7	1.5	0.4
23	Iris-setosa	4.6	3.6	1.0	0.2
24	Iris-setosa	5.1	3.3	1.7	0.5

Fig 2.5 Iris Data Set

2.2.2 Data Pre-Processing

Data Validation/ Cleaning/Preparing Process:

In the data validation, cleaning, and preparation process using the Orange tool, the first step is to address missing values in the dataset. Utilizing the preprocessing module, we employ imputation techniques to handle these missing values effectively. By imputing missing values, such as those denoted by "?", with appropriate strategies like mean, median, or mode imputation, we ensure the completeness and integrity of the dataset. This step is crucial as missing data can adversely affect the performance and accuracy of downstream analysis and modeling tasks. Once missing values are imputed, the dataset undergoes further preprocessing steps, such as normalization or standardization, to ensure uniformity and comparability across features. Through these data preparation processes, we aim to create a clean and reliable datasetready for exploratory analysis, modeling, and insights extraction, enabling effective decision- making in various domains.

Data pre-processing is a critical step in data mining, particularly for plant data, as it ensures that the dataset is clean, consistent, and suitable for analysis. Here are the key steps involved in pre-processing plant data:

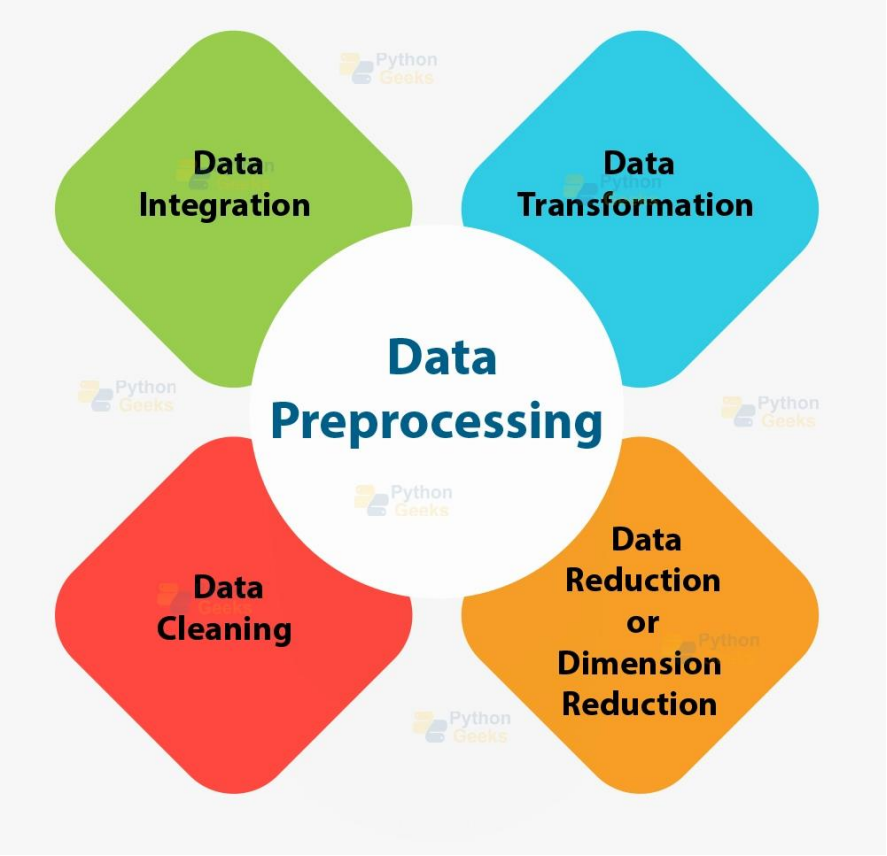


Fig 2.6 Data preprocessing

CHAPTER 3
RESULTS

3.1 ORANGE tool description:

Orange is an open-source data visualization and analysis tool designed for users seeking intuitive yet powerful solutions in machine learning and data mining. Its hallmark feature is a visual programming interface, facilitating the construction of data analysis workflows through interconnected components (widgets). With this approach, users can perform various tasks seamlessly, including data preprocessing, exploratory data analysis, predictive modeling, and visualization. Orange offers an array of preprocessing techniques, allowing users to handle missing values, scale features, encode categorical variables, and select relevant features effortlessly. Moreover, its extensive collection of visualization tools enables users to explore datasets visually, uncovering relationships, distributions, and patterns. Through integration with machine learning algorithms and ensemble learning methods, Orange empowers users to train models for classification,regression, clustering, and association rule mining. Model evaluation tools further aid in assessing model performance, ensuring robust and reliable results. With its blend of usability and versatility, Orange serves as a valuable asset for data scientists, researchers, and analysts across various domains, fostering innovation and insight discovery.

3.2 Screen shots

ORANGE Tool: Demonstrate performing classification on data sets

- Download and install ORANGE

Fig 3.1 Download and Install Orange

- Open Orange and Select new to start a new project

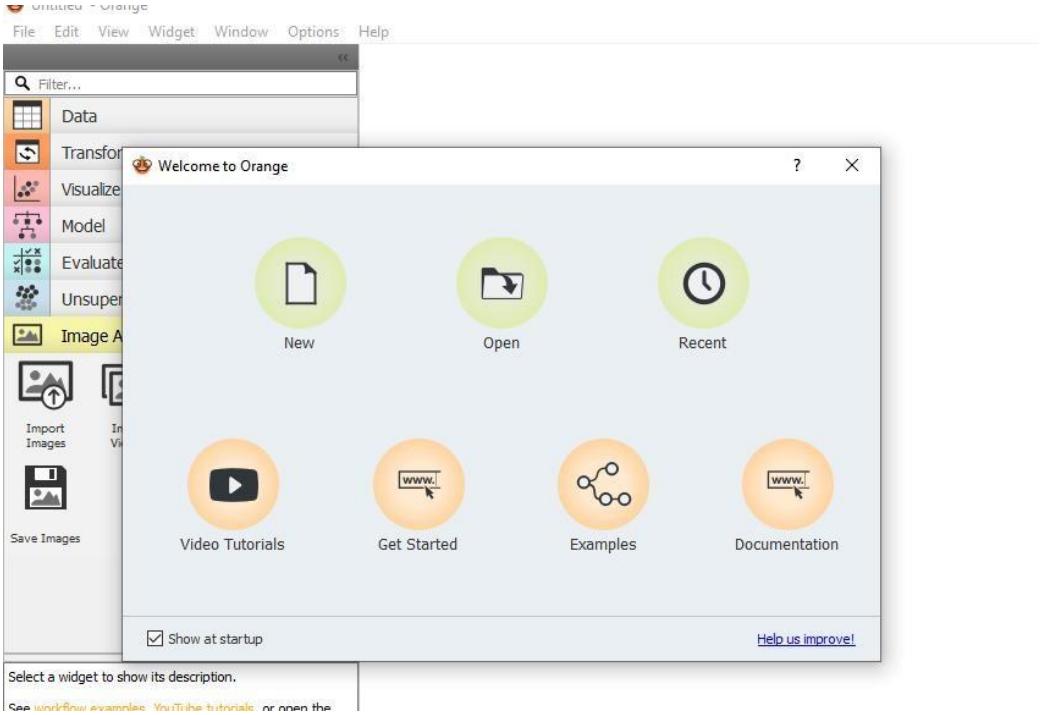


Fig 3.2 Open new File

- From the Data,select file.Double click on it and Load the dataset

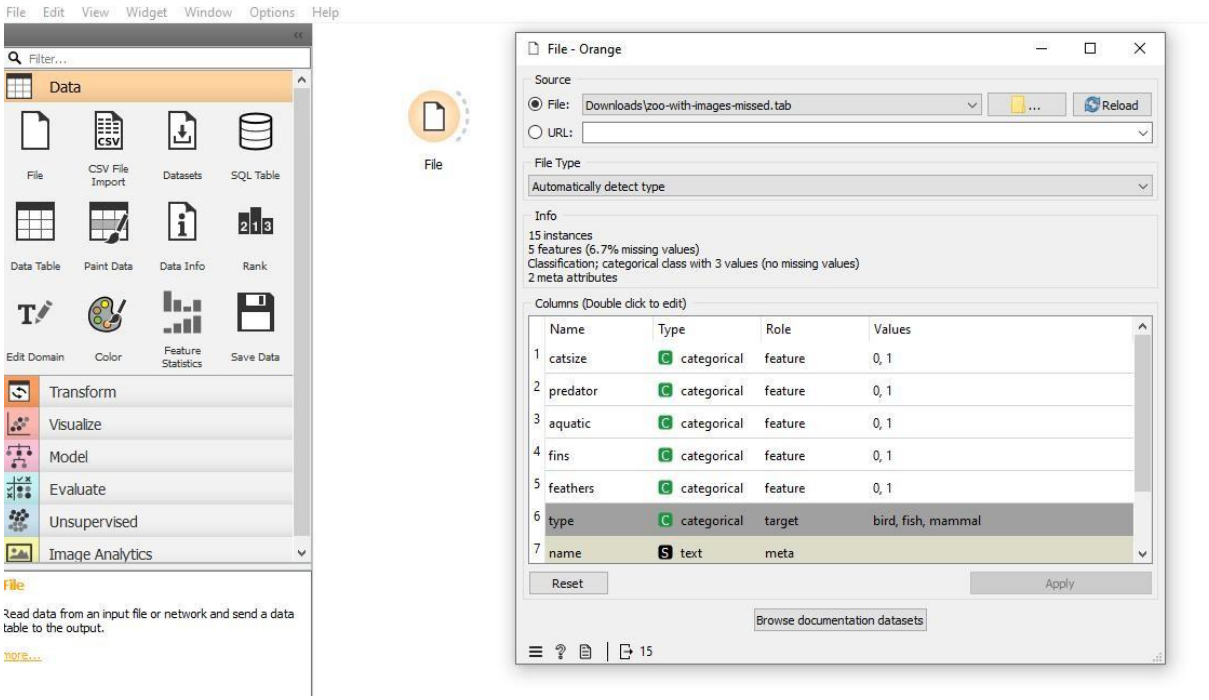


Fig 3.3 Load the Dataset

- The dataset can be viewed with a Data Table and its information with Data Info

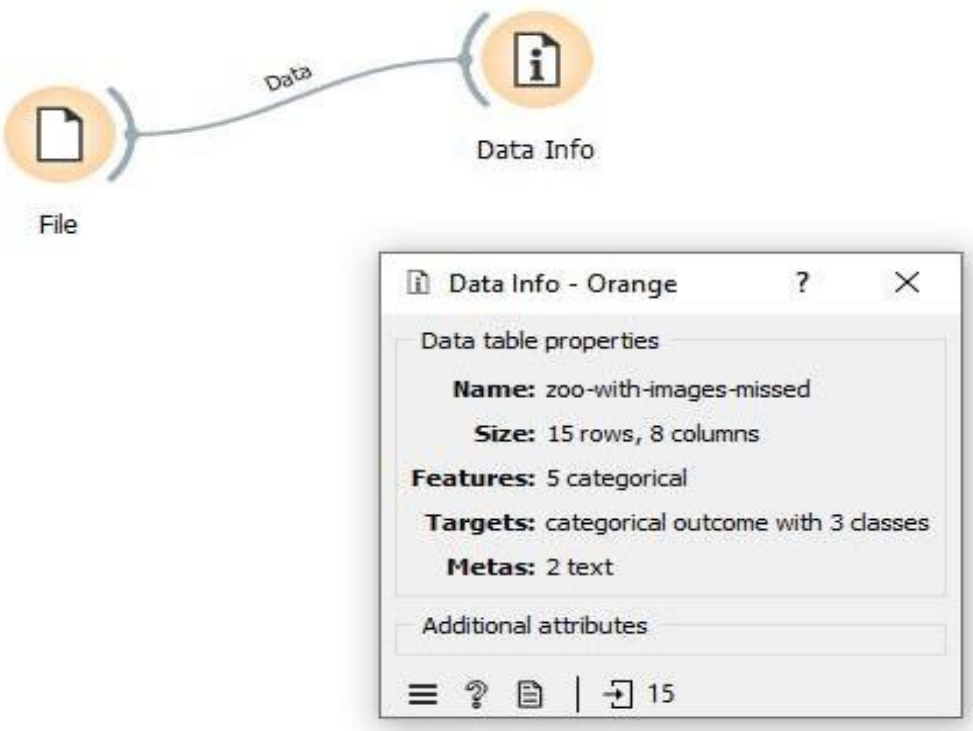


Fig 3.4 Data Info of datatset

	iris	sepal length	sepal width	petal length	petal width
1	Iris-setosa	5.1	3.5	1.4	0.2
2	Iris-setosa	4.9	3.0	1.4	0.2
3	Iris-setosa	4.7	3.2	1.3	0.2
4	Iris-setosa	4.6	3.1	1.5	0.2
5	Iris-setosa	5.0	3.6	1.4	0.2
6	Iris-setosa	5.4	3.9	1.7	0.4
7	Iris-setosa	4.6	3.4	1.4	0.3
8	Iris-setosa	5.0	3.4	1.5	0.2
9	Iris-setosa	4.4	2.9	1.4	0.2
10	Iris-setosa	4.9	3.1	1.5	0.1
11	Iris-setosa	5.4	3.7	1.5	0.2
12	Iris-setosa	4.8	3.4	1.6	0.2
13	Iris-setosa	4.8	3.0	1.4	0.1
14	Iris-setosa	4.3	3.0	1.1	0.1
15	Iris-setosa	5.8	4.0	1.2	0.2
16	Iris-setosa	5.7	4.4	1.5	0.4
17	Iris-setosa	5.4	3.9	1.3	0.4
18	Iris-setosa	5.1	3.5	1.4	0.3
19	Iris-setosa	5.7	3.8	1.7	0.3
20	Iris-setosa	5.1	3.8	1.5	0.3
21	Iris-setosa	5.4	3.4	1.7	0.2

Fig 3.5 Data Table

- Apply Classification models on the preprocessed data .We used KNN,Neural networks,Logistic Regression,Tree models forperforming classification on this dataset.

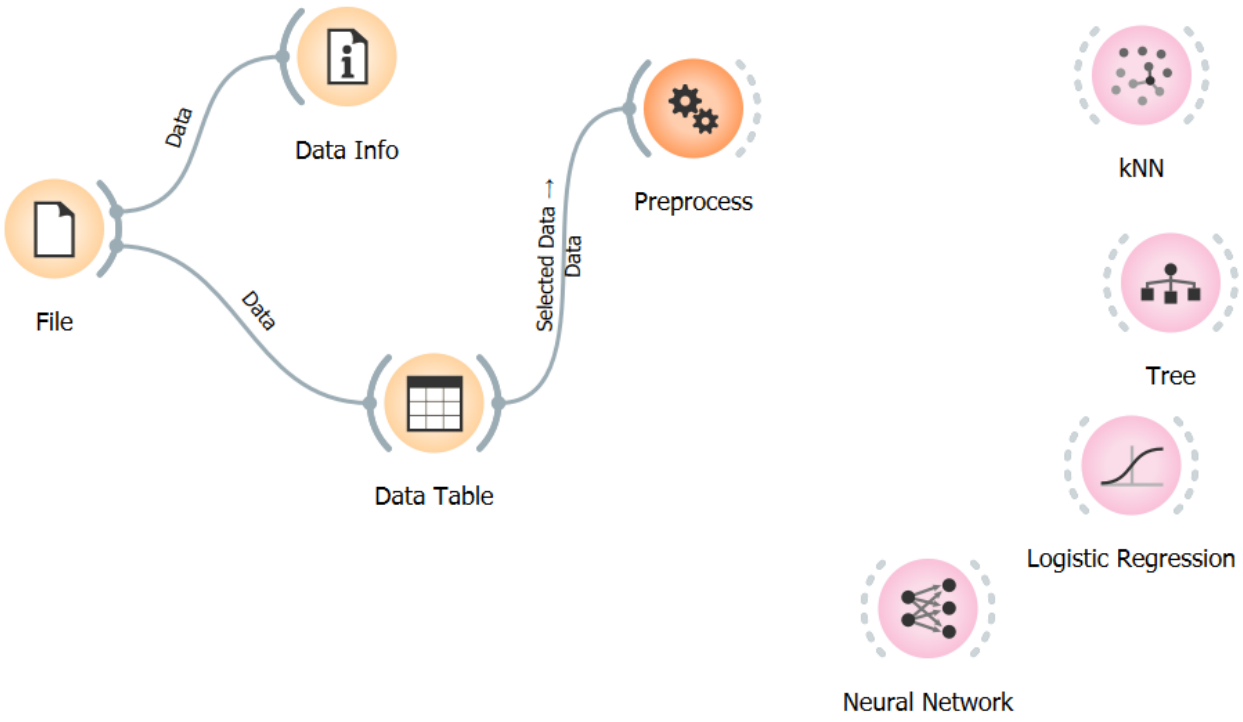


Fig 3.6 Applying Classification Models

- Evaluate the models with Test Score.

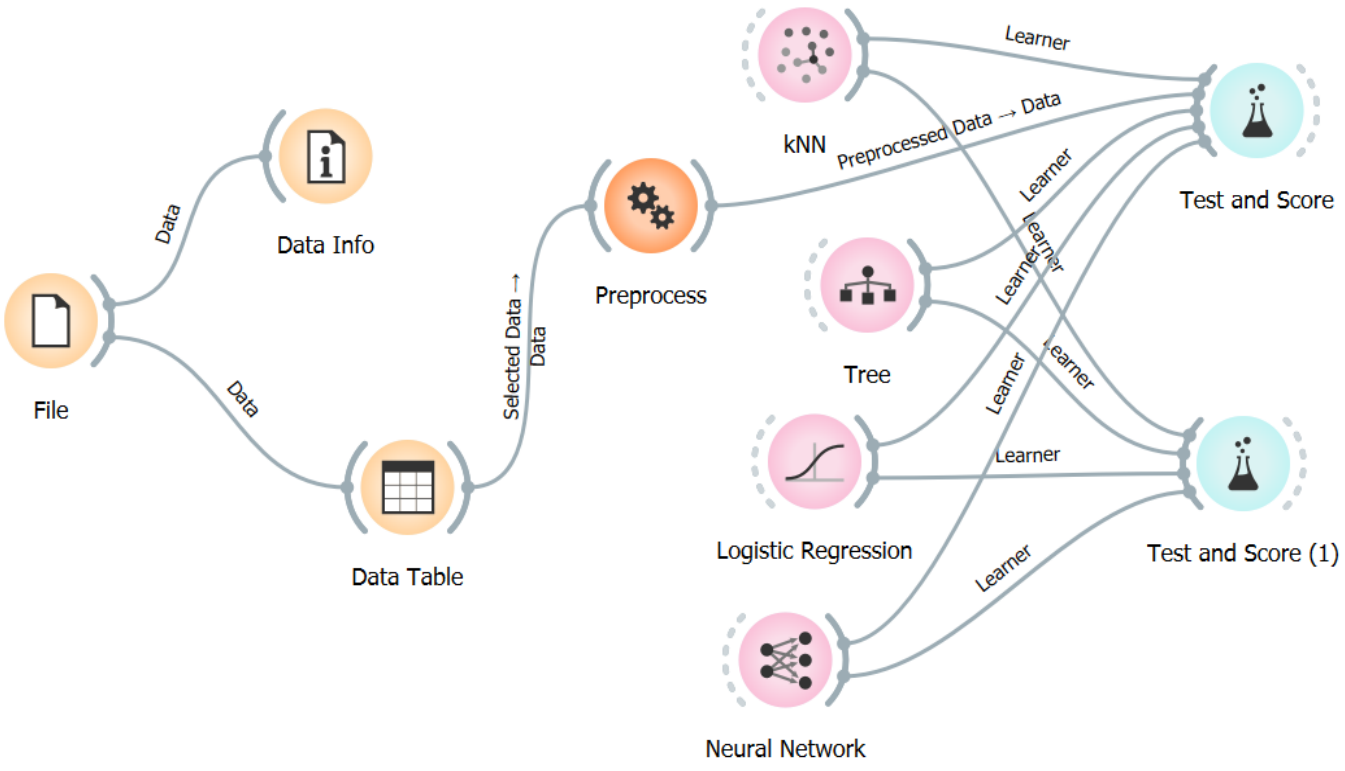


Fig 3.7 Evaluate test score

This is the evaluation metrics derived from the test and score before performing preprocessing techniques.

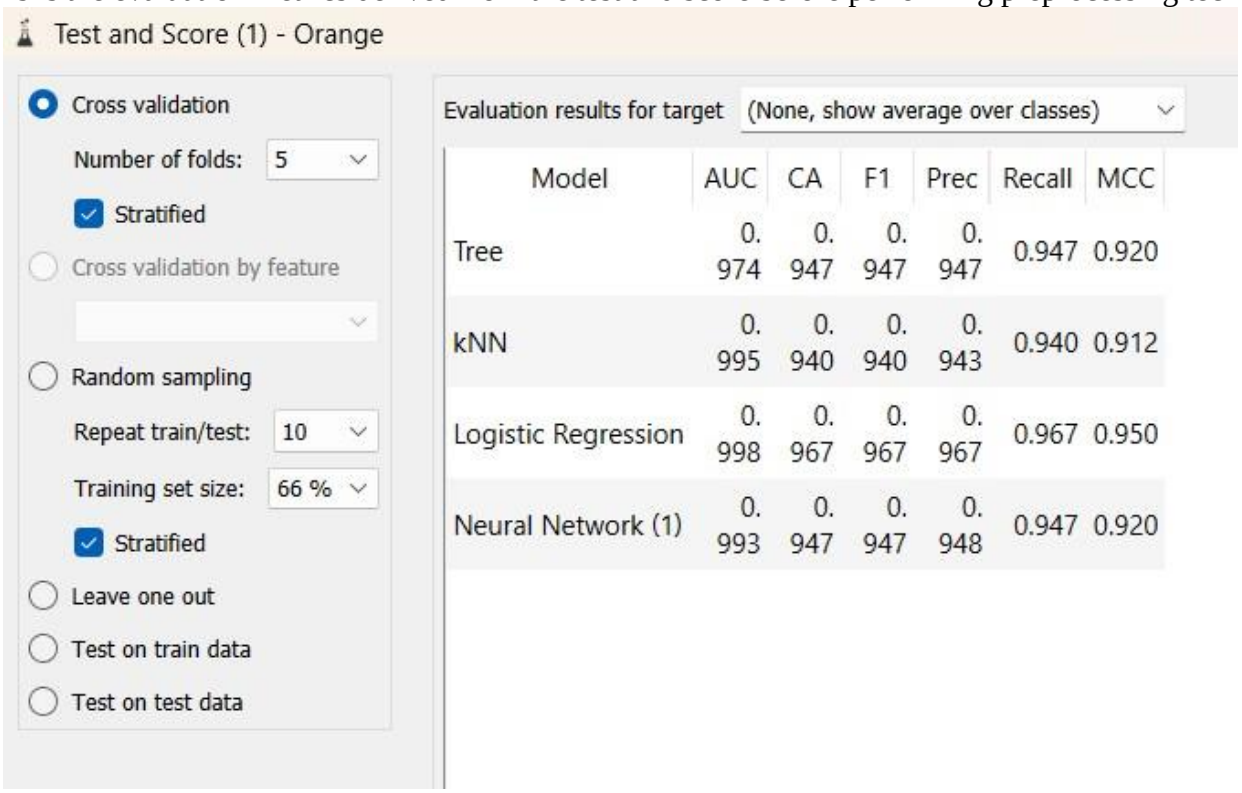
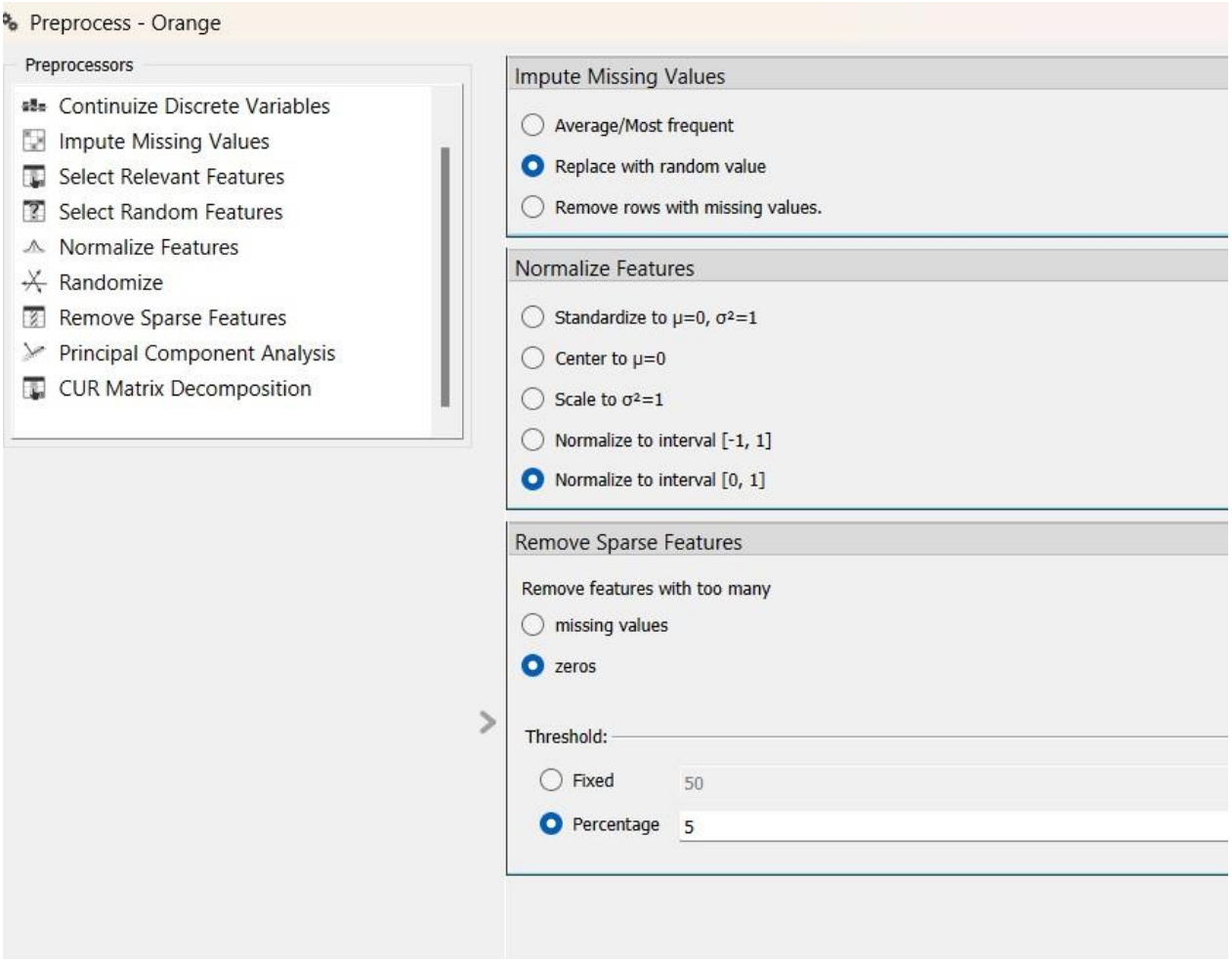
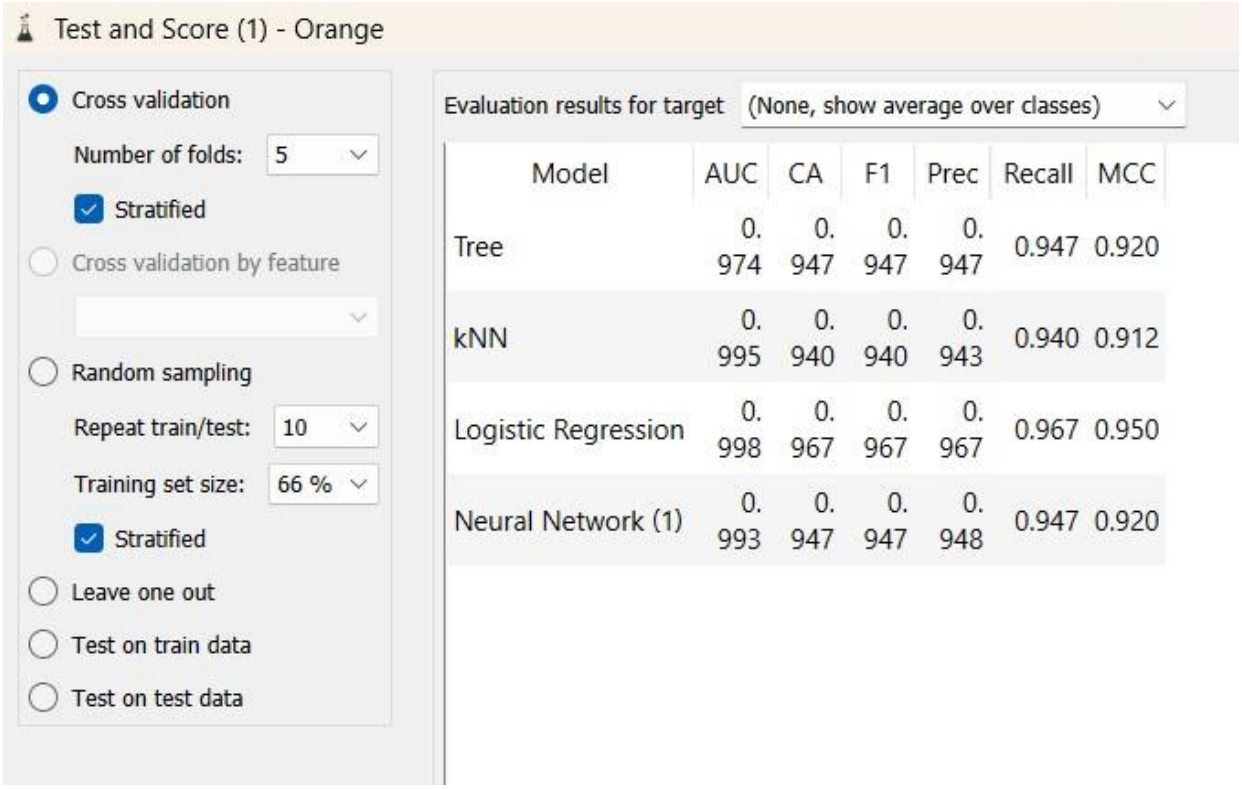


Fig 3.8 Evaluating Classification Models

- We can also perform various preprocessing techniques such as :
  Discretize Continuous Variables.
  Continuous Discrete Variables
  Select Random Features, etc

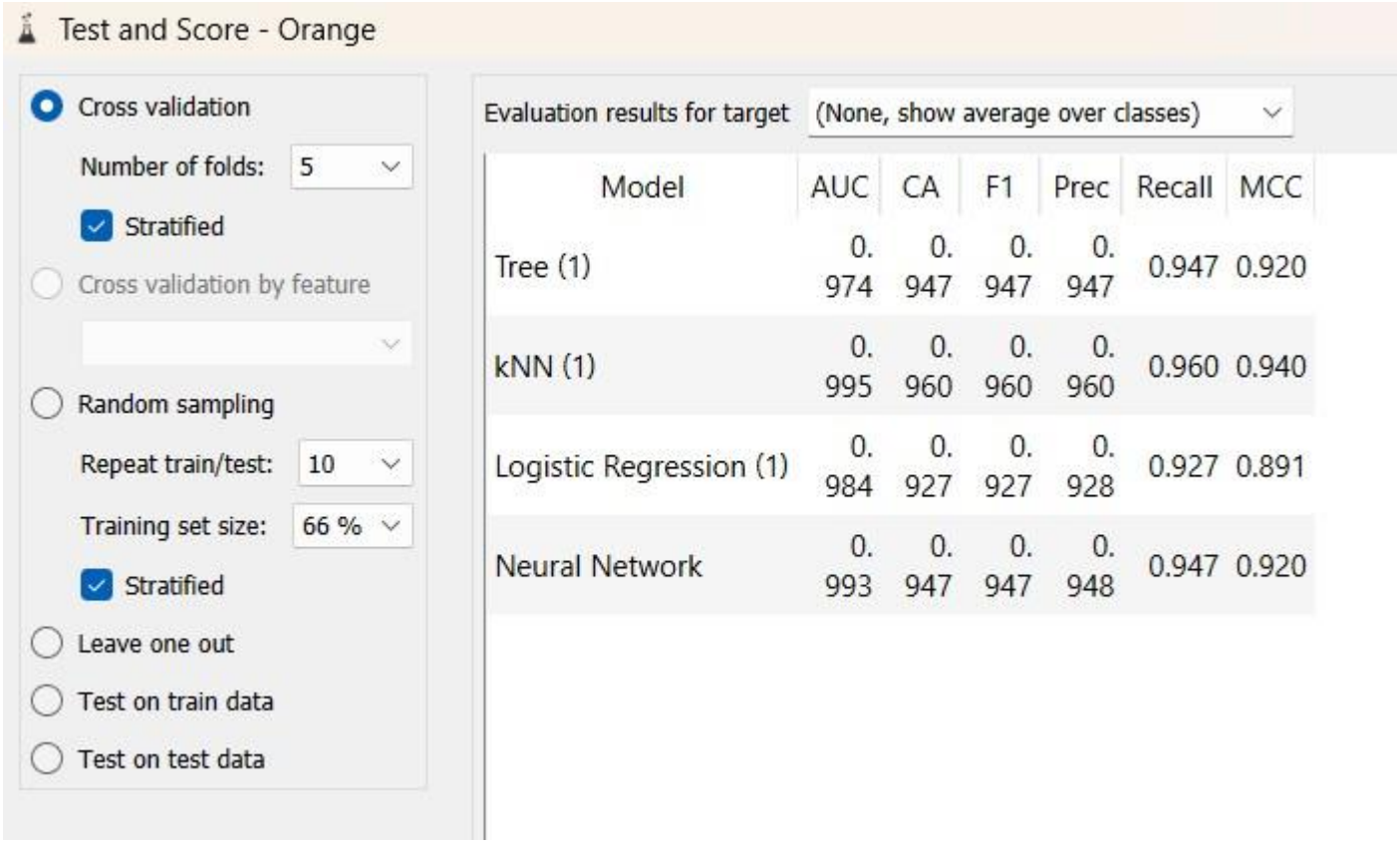


After the application of **Discretize continuous variables** .The evaluation is:



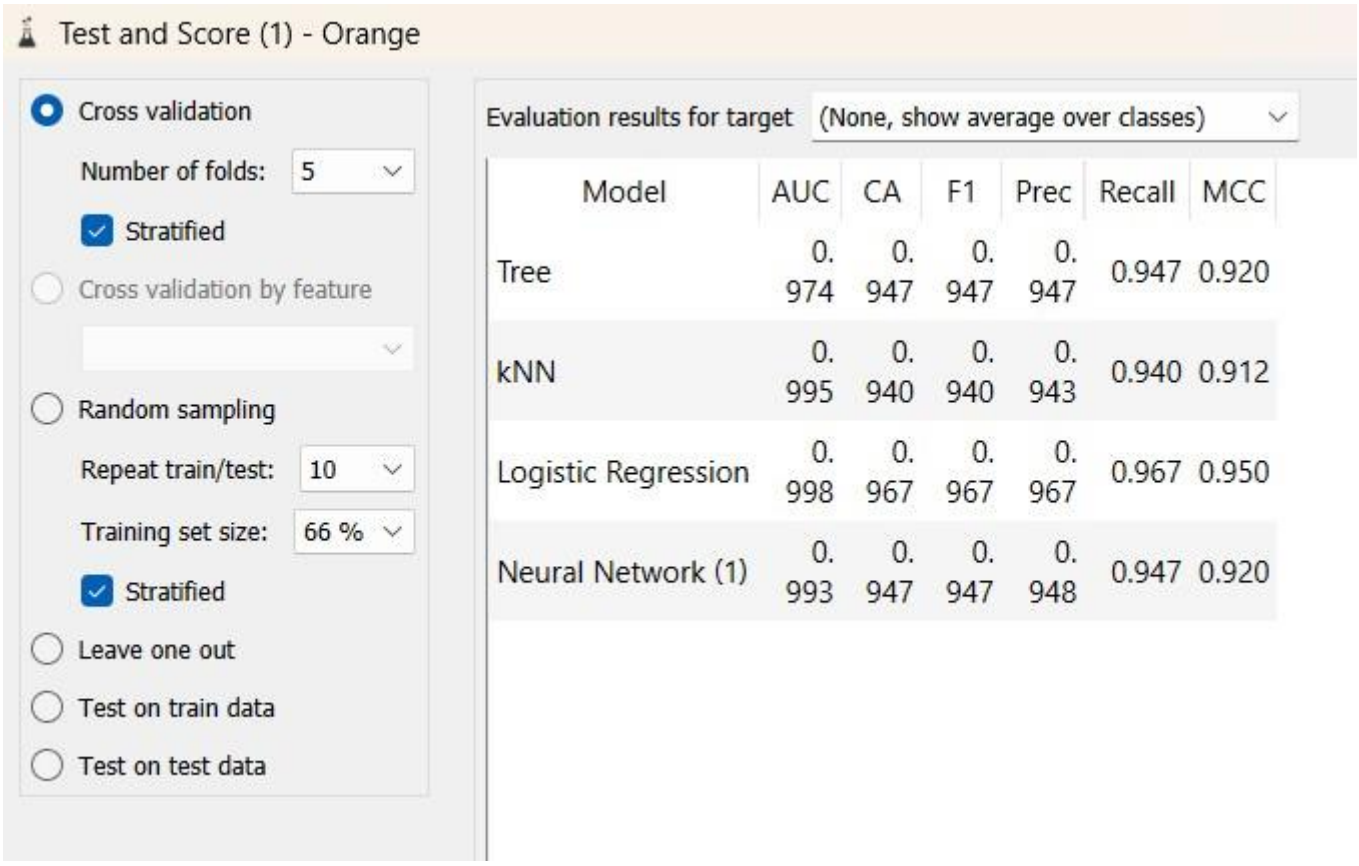
Model	AUC	CA	F1	Prec	Recall	MCC
Tree	0.974	0.947	0.947	0.947	0.947	0.920
kNN	0.995	0.940	0.940	0.943	0.940	0.912
Logistic Regression	0.998	0.967	0.967	0.967	0.967	0.950
Neural Network (1)	0.993	0.947	0.947	0.948	0.947	0.920

After the applying **Continuous Discrete Variables** The evaluation is:



Model	AUC	CA	F1	Prec	Recall	MCC
Tree (1)	0.974	0.947	0.947	0.947	0.947	0.920
kNN (1)	0.995	0.960	0.960	0.960	0.960	0.940
Logistic Regression (1)	0.984	0.927	0.927	0.928	0.927	0.891
Neural Network	0.993	0.947	0.947	0.948	0.947	0.920

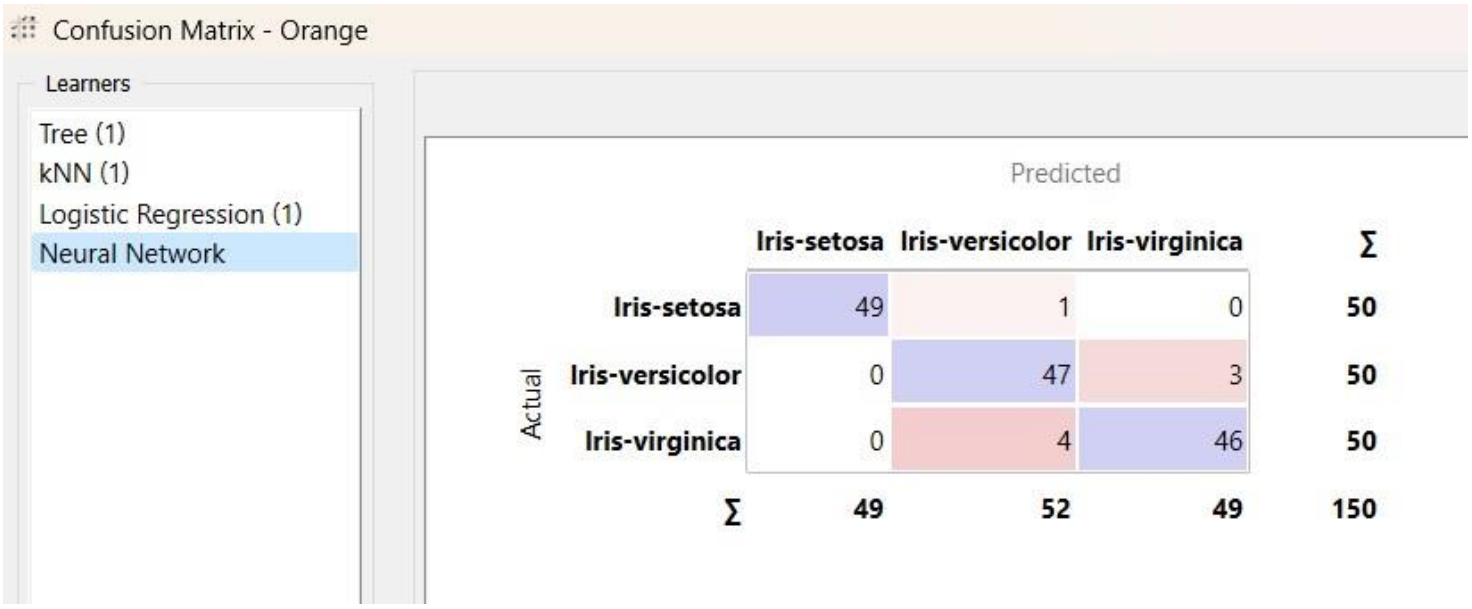
This is the evaluation metrics derived from the test and score after performing preprocessing techniques.



Model	AUC	CA	F1	Prec	Recall	MCC
Tree	0.974	0.947	0.947	0.947	0.947	0.920
kNN	0.995	0.940	0.940	0.943	0.940	0.912
Logistic Regression	0.998	0.967	0.967	0.967	0.967	0.950
Neural Network (1)	0.993	0.947	0.947	0.948	0.947	0.920

Fig 3.9 Evaluation through Test and Score

- Represent the Accuracy values in Confusion Matrix



	Iris-setosa	Iris-versicolor	Iris-virginica	Σ
Iris-setosa	49	1	0	50
Iris-versicolor	0	47	3	50
Iris-virginica	0	4	46	50
Σ	49	52	49	150

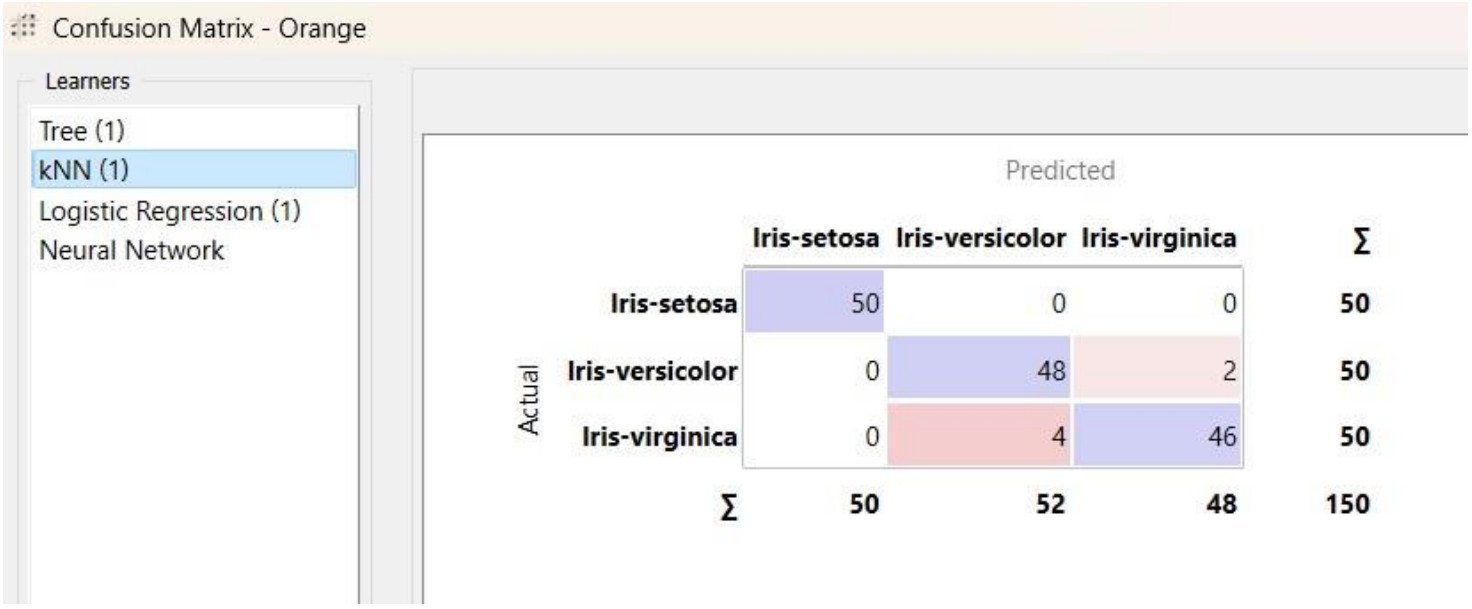


Fig 3.10 Mis-Classified with KNN Model

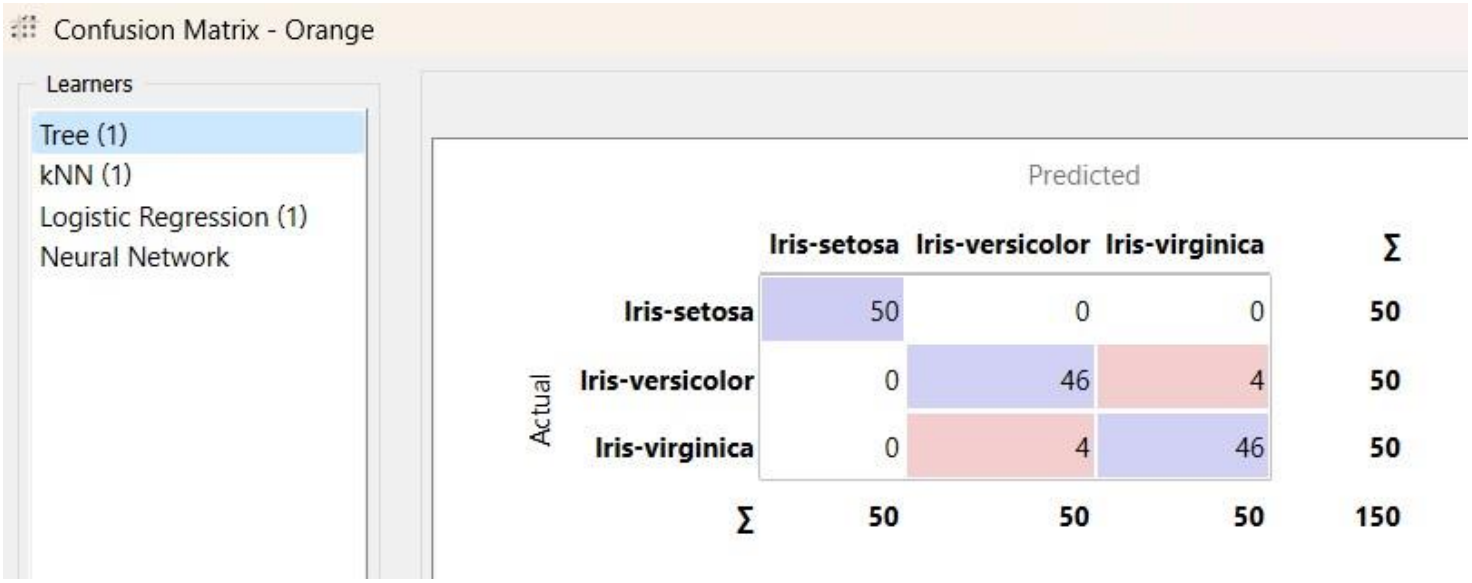
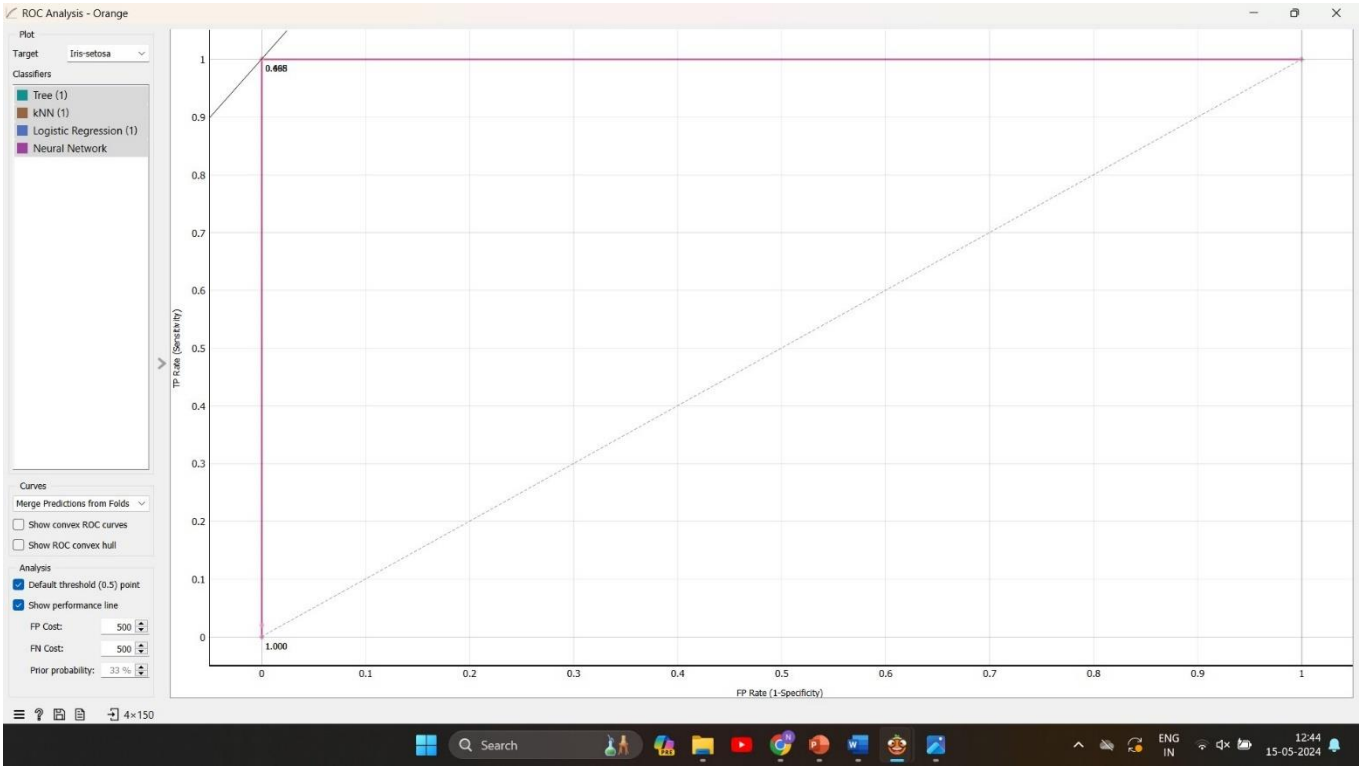


Fig 3.11 Confusion Matrix of Tree

- Visualize the output through Scatter plot



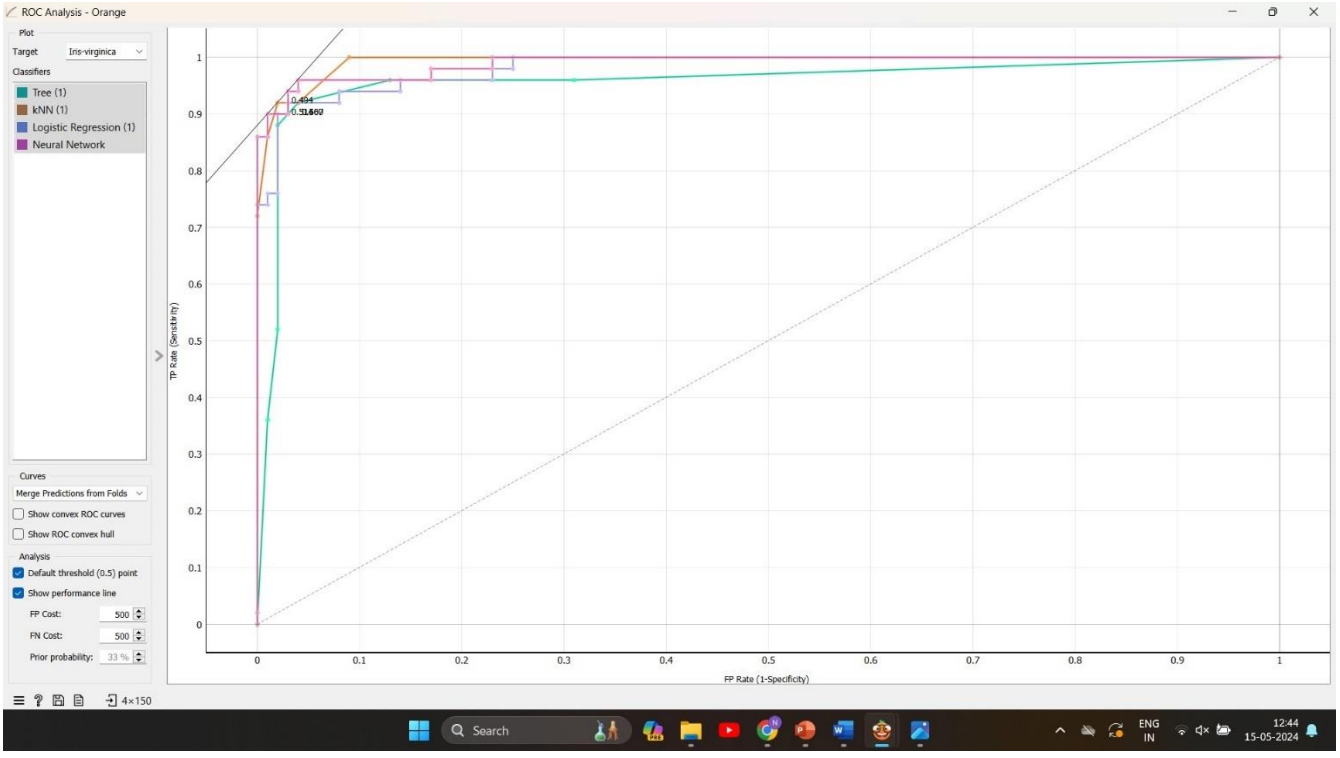


Fig 3.12 Scatter Plot

- The Overall Workflow is given below:

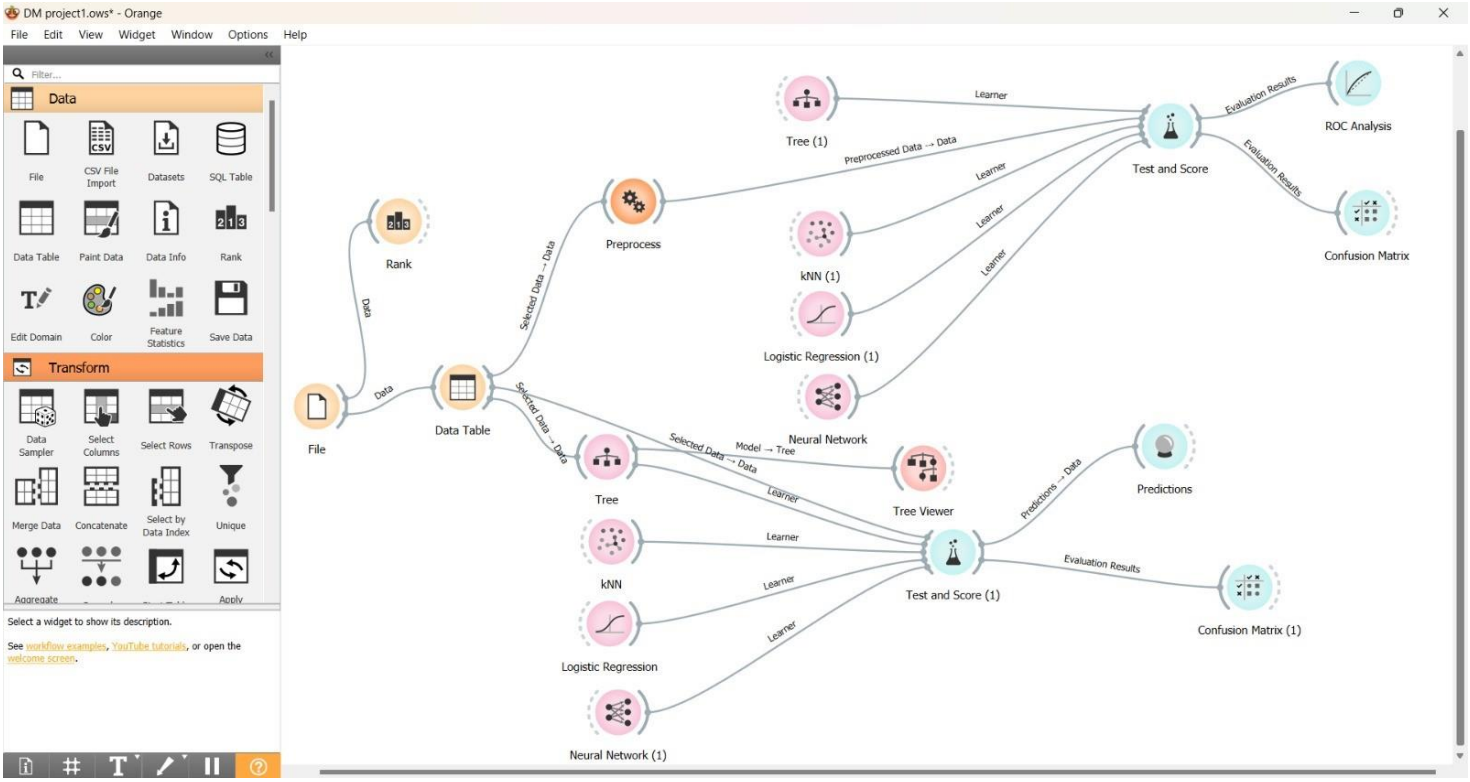


Fig 3.13 Overall Workflow

- The overall workflow diagram shows the final output which gives the analysis of total workflow from the file which contains the iris dataset and the data table which arranged the data in a order
- Then the data table is connected to preprocess which is later connected to test and score
- One data table connection goes through preprocess and another data table connection will go directly to test and score
- Then the connection will be made to the classification models and we have to note the difference between the outputs came from the preprocessed data and not preprocessed data
- Taking those differences to consideration we have to conclude which one gives us the better results and why
- Then the final confusion matrix will be taken into consideration for the conclusion

CHAPTER 4
CONCLUSION AND FUTURESCOPE

Conclusion

In this project, we explored the application of classification techniques to predict the species of the Iris flower using the well-known Iris dataset. Our primary goal was to accurately classify the Iris flowers into one of the three species: Setosa, Versicolor, and Virginica, based on four features - sepal length, sepal width, petal length, and petal width.

To achieve this, we implemented several classification algorithms, including:

1. Logistic Regression

2.K-Nearest Neighbors (KNN)

3. Neural Network

4. Decision Tree

The classification techniques applied in this project successfully predicted the species of Iris flowers with high accuracy. Random Forest and SVM emerged as the most effective classifiers for this dataset.

Furthermore, the Iris dataset proved to be an excellent candidate for demonstrating the effectiveness of various machine learning algorithms due to its balanced classes and well-separated feature distributions.

Overall, this project highlights the practicality and efficacy of machine learning techniques in solving classification problems and sets the stage for further exploration with more complex and larger datasets.

Future Scope:

The classification of Iris data using machine learning techniques offers a significant foundation for various future research and practical applications. One promising future scope lies in the enhancement of classification algorithms to achieve higher accuracy and efficiency in distinguishing between different Iris species. By incorporating advanced techniques such as deep learning, ensemble methods, and feature engineering, researchers can improve predictive performance and robustness.

Further, expanding this research to include a broader variety of plant species can contribute to biodiversity monitoring and conservation efforts. This can help in identifying and protecting endangered species by providing quick and reliable identification methods. Another significant direction is the application of these classification techniques in educational tools, where students can learn about plant biology and machine learning concurrently through interactive platforms.

REFERENCES

- Fisher, R. A. (1936). "The Use of Multiple Measurements in Taxonomic Problems". Annals of Eugenics. This paper is the original publication where the Iris dataset was first introduced.

- Pedregosa, F., et al.(2011). "Scikit-learn: Machine Learning in Python". Journal of Machine Learning Research, 12, pp. 2825-2830. This reference details the machine learning library used extensively in our project for implementing classification algorithms.

- Breiman, L. (2001). "Random Forests". Machine Learning, 45(1), pp. 5-32. This paper provides a comprehensive overview of the Random Forest algorithm, which was one of the top-performing classifiers in our project.

- Cortes, C. & Vapnik, V.(1995). "Support-vector networks". Machine Learning, 20(3), pp. 273-297. This foundational paper describes the SVM algorithm, another high-performing method in our analysis.

- Han, J., Kamber, M., & Pei, J. (2011). "Data Mining: Concepts and Techniques". Morgan Kaufmann. This textbook offers a detailed explanation of various classification techniques and their applications.

- Raschka, S. & Mirjalili, V.(2017). "Python Machine Learning". Packt Publishing. This book provides practical insights and examples on implementing machine learning algorithms using Python.

- Kuhn, M. & Johnson, K. (2013). "Applied Predictive Modeling". Springer. This reference covers advanced topics in predictive modeling, including model evaluation and feature engineering.

- 8. Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). "Data Mining: Practical Machine Learning Tools and Techniques". Morgan Kaufmann. This book is a useful resource for understanding the practical aspects of machine learning and data mining.

SESHADRI RAO GUDLAVALLERU ENGINEERING COLLEGE

(An Autonomous Institute with Permanent Affiliation to JNTUK, Kakinada)
Seshadri Rao Knowledge Village, Gudlavalleru

Department of Computer Science and Engineering

Program Outcomes (POs)

Engineering Graduates will be able to:

1. **Engineering knowledge**: Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.
2. **Problem analysis**: Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.
3. **Design/development of solutions**: Design solutions for complex engineering problems anddesign system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.
4. **Conduct investigations of complex problems**: Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions., component, or software to meet the desired needs.
5. **Modern tool usage**: Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations.
6. **The engineer and society**: Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice.
7. **Environment and sustainability**: Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need forsustainable development.
8. **Ethics**: Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.
9. **Individual and team work**: Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.
10. **Communication**: Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write

effective reports and design documentation, make effective presentations, and give and receive clear instructions.

11. **Project management and finance**: Demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.
12. **Life-long learning**: Recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.

Program Specific Outcomes (PSOs)

PSO1 : Design, develop, test and maintain reliable software systems and intelligent systems.

PSO2 : Design and develop web sites, web apps and mobile apps.

PROJECT PROFORMA

Classification of Project	Application	Product	Research	Review
	√			

Note: Tick Appropriate category

Data Mining Outcomes	
Course Outcome (CO1)	Describe fundamentals, and functionalities of data mining system and data preprocessing techniques.
Course Outcome (CO2)	Illustrate the major concepts and operations of multi dimensional data models.
Course Outcome (CO3)	Analyze the performance of association rule mining algorithms for finding frequent item sets from the large databases.
Course Outcome (CO4)	Apply classification algorithms to solve classification problems.
Course Outcome (CO5)	Use clustering methods to create clusters for the given data set.

Mapping Table

CS3509 : DATA MINING														
Course Outcomes	Program Outcomes and Program Specific Outcome													
	PO 1	PO 2	PO 3	PO 4	PO 5	PO 6	PO 7	PO 8	PO 9	PO 10	PO 11	PO 12	PSO 1	PSO 2
CO1	1	1										1		
CO2	1											1		
CO3	2	3	2									2	1	
CO4	2	2	3	2								2	2	
CO5	1	2	3	1								2	1	

Note: Map each Data Mining outcomes with POs and PSOs with either 1 or 2 or 3 based on level of mapping as follows:

1-Slightly (Low) mapped 2-Moderately (Medium) mapped 3-Substantially (High) mapped