**Decoding Audiology Data Using Classification Model**

**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

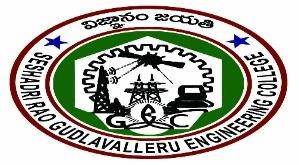
**T. Likitha V.Anusha (21481A05M0) (21481A05P0)**

**V. Grace Vijaya Y. Abhilash**

**(21481A05O4) (21481A05Q1)**

Under the Enviable and Esteemed Guidance of

Mrs.G.KEERTHI, M. Tech, (Ph.D)

 Assistant Professor, Department of 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 **“Decoding Audiology Data: Classification and Patterns through Mining”**is a bonafide record of work carried out by T. Likitha(21481A05M0), V.Anusha(21481A05P0), V.GraceVijaya(21481A05O4), Y.Abhilash(21481A05Q1), under the guidance and supervision of **Mrs.G.Keerthi, Assistant professor,** Computer Science 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 2022-23.

# Project Guide Head of the Department (Mrs.G.Keerthi) (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 **Mrs.G. Keerthi, 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 criticisms are 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 intime.

Team Members

T. Likitha(21481A05M0)

V. Anusha(21481A05P0)

Y. Abhilash(21481A05Q1)

V. Grace Vijaya(21481A05O4)

# INDEX

**TITLE PAGE NO**

ABSTRACT i

CHAPTER 1.INTRODUTION

* 1. Introduction 1
  2. Problem definition 6

CHAPTER 2: PROPOSED METHOD

* 1. Methodology 7
     1. Block Diagram 8
     2. Algorithm and Explanation 9
  2. Data Preparation
     1. Dataset Description 10
     2. Data Pre-processing 11

# CHAPTER 3: RESULTS

* 1. ORANGE tool description 12
  2. Screen shots 13

# CHAPTER 4: CONCLUSION AND FUTURE SCOPE 25

**References**

**List of Program Outcomes and Program Specific Outcomes Mapping of Program Outcomes with graduated POs and PSO**

# ABSTRACT

The Audiology dataset comprises various auditory parameters categorized into classes based on severity and type of hearing impairment. With attributes ranging from air and bone conduction thresholds to historical symptoms and waveform characteristics, the dataset offers a comprehensive view of audiological profiles.

Utilizing classification techniques, this dataset enables the prediction of hearing impairment severity and its underlying causes. By analyzing patterns within the data, classification models can distinguish between different types of hearing loss, such as cochlear, conductive, or retro cochlear disorders, and assess their severity levels.

The abstract of this audiology dataset encapsulates its potential in aiding diagnosis and treatment planning for individuals with hearing impairments. Through machine learning algorithms, healthcare professionals can leverage this dataset to develop predictive models that enhance diagnostic accuracy and inform personalized interventions, thereby improving the quality of life for individuals affected by hearing disorders.

i

# INTRODUCTION

# CLASSIFICATION

# CHAPTER 1

# INTRODUCTION

Classification is a task in data mining that involves assigning a class label to each instance in a dataset based on its features. The goal of classification is to build a model that accurately predicts the class labels of new instances based on their features. Classification is one form of data analysis, where a model or classifier is constructed to predict categorical labels, such as “safe” or “risky” for the loan application data.

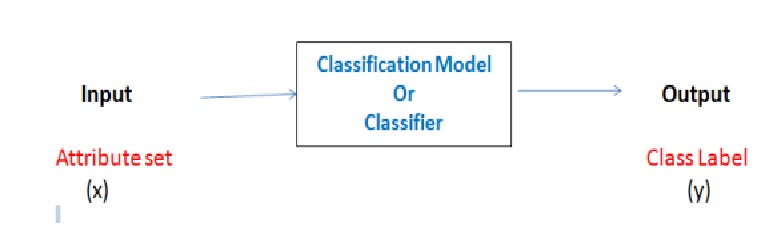


Fig 1.1.1 Classification Definition

The process of building a classification model typically involves the following steps:

**Data Collection:**

The first step in building a classification model is data collection. In this step, the data relevant to the problem at hand is collected. The data should be representative of the problem and should contain all the necessary attributes and labels needed for classification. The data can be collected from various sources, such as surveys, questionnaires, websites, and databases.

**Data Preprocessing:**

The second step in building a classification model is data preprocessing. The collected data needs to be preprocessed to ensure its quality. This involves handling missing values, dealing with outliers, and transforming the data into a format suitable for analysis. Data preprocessing also involves converting the data into numerical form, as most classification algorithms require numerical input.

**Handling Missing Values:** Missing values in the dataset can be handled by replacing them with the mean, median, or mode of the corresponding feature or by removing the entire record.

1

**Feature Selection:**

The third step in building a classification model is feature selection. Feature selection involves identifying the most relevant attributes in the dataset for classification. This can be done using various techniques, such as correlation analysis, information gain, and principal component analysis.

**Correlation Analysis**: Correlation analysis involves identifying the correlation between the features in the dataset. Features that are highly correlated with each other can be removed as they do not provide additional information for classification.

**Information Gain:** Information gain is a measure of the amount of information that a feature provides for classification. Features with high information gain are selected for classification.

**Model Selection:**

The fourth step in building a classification model is model selection. Model selection involves selecting the appropriate classification algorithm for the problem at hand. There are several algorithms available, such as decision trees, support vector machines, and neural networks.

**Decision Trees:** Decision trees are a simple yet powerful classification algorithm. They divide the dataset into smaller subsets based on the values of the features and construct a tree-like model that can be used for classification.

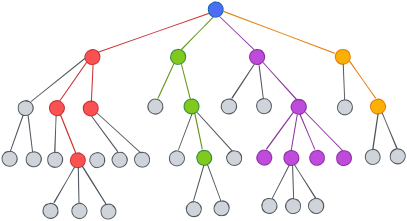


Fig 1.1.2 Decision Trees

**KNN Classifier:** K-Nearest Neighbors (KNN) is a straightforward machine learning algorithm that classifies a data point based on the majority class of its nearest neighbors. It's intuitive but sensitive to the choice of 'k' and the distance metric.

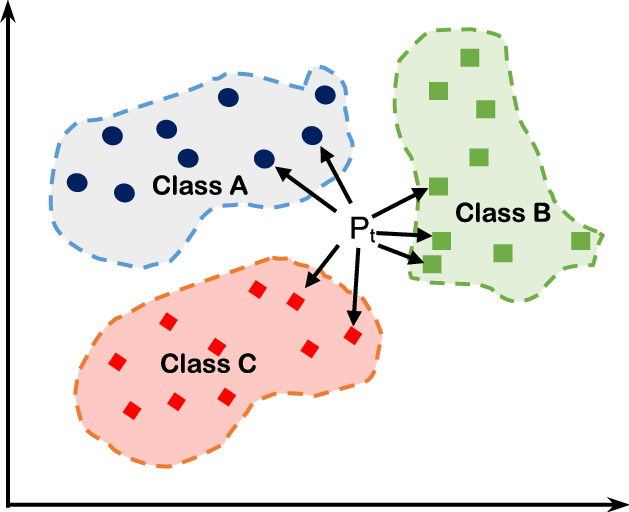


Fig 1.1.3 KNN Classifier

2

**Naive Bayes:** Naive Bayes classifier is a probabilistic model that calculates the probability of a class label given input features using Bayes' theorem. It assumes independence between features, making it computationally efficient and effective for text classification and simple classification tasks.

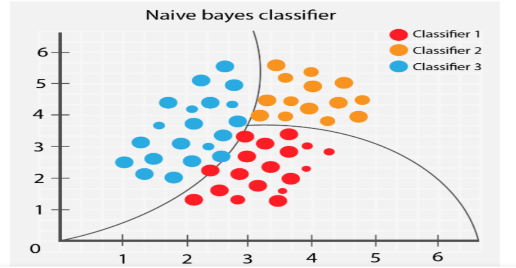


Fig 1.1.4 Naïve Bayes

**Random Forest:** Random Forest works by creating a number of Decision Trees during the training phase. Each tree is constructed using a random subset of the data set to measure a random subset of features in each partition. This randomness introduces variability among individual trees, reducing the risk of overfitting and improving overall prediction performance. In prediction, the algorithm aggregates the results of all trees, either by voting (for classification tasks) or by averaging (for regression tasks) This collaborative decision-making process, supported by multiple trees with their insights, provides an example stable and precise results.



Fig 1.1.5 Random Forest

Here’s the list of Classification Models in detail:

|  |  |  |
| --- | --- | --- |
| **Algorithm** | **Description** | **Type** |
| Logistic Regression | Extension of linear regression that’s used for classification tasks. The output variable 3 is binary (e.g., only black or white) rather than  continuous(e.g., an infinite list of potential colors) | Classification rather regression |

3

|  |  |  |
| --- | --- | --- |
| Decision Tree | Highly interpretable classification or regression model that splits data-feature values into branches at decision nodes.(e.g., if a feature is a color, each possible color becomes a new branch) until a final decision output is  made. | Classification |
| Naïve Bayes | The Bayesian method is a classification method that makes use of the Bayesian theorem. The theorem updates the prior knowledge of an event with the independent  probability of each feature that can affect the event. | Regression and Classification |
| Support Vector Machine | SVM, is typically, used for the classification  task. SVM algorithm finds a hyperplane that optimally divided the classes. It is best used with a non-linear solver. | Regression and  classification |
| Random Forest | The algorithm is built upon a decision tree to  improve the accuracy drastically. Random forest generates many times n simple decision trees and uses the ‘majority vote’ method to decide on which label to return. For the classification task, the final prediction will be the one with the most vote; while for the regression task, the average prediction of all  the trees is the final prediction. | Regression and  Classification |
| AdaBoost | Classification or regression technique that uses a multitude of models to come up with a decision but weights them based on their  accuracy in predicting the outcome | Regression and Classification |
| Gradient-boosting trees | Gradient boosting trees is a state-of-the-art classification/regression technique. It is focusing on the error committed by the  previous trees and tries to correct it. | Regression and Classification |

Table 1.1.1: Classification Models

## 

## 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

4

## How to Choose Classification Models?

Choosing classification methods in data mining involves several considerations to ensure optimal model performance. First, assess the nature of your dataset, including its size, dimensionality, and the presence of noise or missing values. Next, understand the characteristics of the problem you are solving, such as the number of classes, class imbalance, and the complexity of decision boundaries. Then, evaluate the computational requirements and scalability of different algorithms based on your dataset size and available resources. Consider the interpretability of the models and whether it is essential to understand the reasoning behind predictions. Additionally, conduct experiments with multiple algorithms, using techniques like cross-validation, to compare their performance metrics such as accuracy, precision, recall, and F1-score. Finally, consider the specific requirements and constraints of your application domain to select the most suitable classification method that balances predictive accuracy, interpretability, and computational efficiency.

## Challenges and Limitations of Classification:

**Imbalanced Datasets:** Dealing with imbalanced datasets where one class dominates, leading to biased models and misclassification of minority classes.

**High-Dimensional Data**: Handling high-dimensional data where the number of features exceeds the number of samples, causing overfitting, increased computational complexity, and interpretability issues.

**Noisy or Missing Data:** Difficulty in handling noisy or missing data, which can impact the model's ability to accurately learn patterns and make predictions.

**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.

## 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.

5

* **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 challenge in accurately classifying individual’s impairments lies in the complexity of the

Of the audiology dataset, which includes a diverse array of auditory parameters ranging from

Historical Symptoms to physiological measurements and diagnostic indicators. The task demands

Classification techniques to discern various types and severities of hearing disorders, such as

The degree of impairment, underlying causes, and potential comorbidities. Precision in diagnosis

And treatment planning for audiology practitioner’s hinges on developing robust machine learning

Models capable of effectively utilizing this rich dataset, ultimately leading to enhanced accuracy

And improved quality of life for individuals affected by hearing disorders.

6

# CHAPTER 2 PROPOSED METHOD

* 1. **Methodology**

At first the data need to be collected, then after data preprocessing is done to clean the data. Next some data is used for training the model, some data is used for testing the model. Finally Classification algorithms are applied.

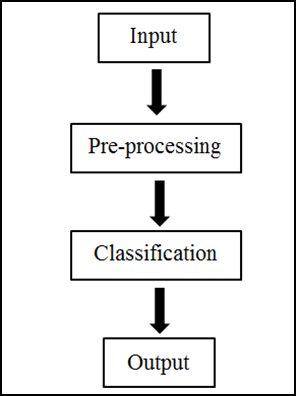


Fig 2.1.1 Classification methodology

**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 that class.

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

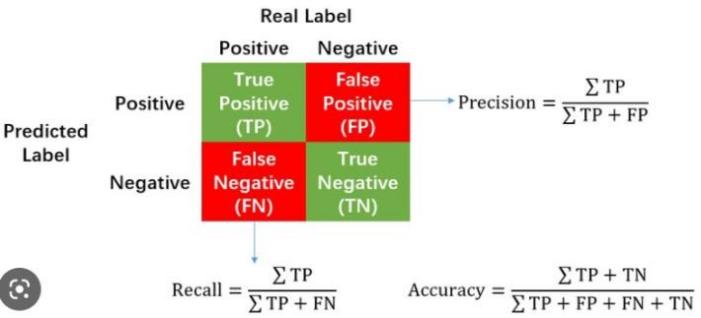
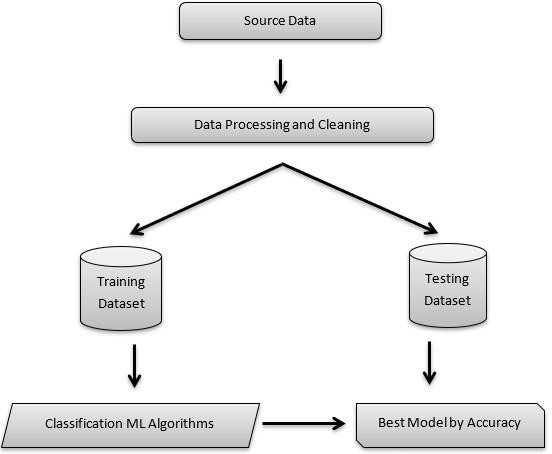
**F-Measure:** It offers a single score that evenly weighs the issues of precision and recall. **Confusion Matrix:** It is used to determine the classification models performance for a set of test data.

Fig 2.1.2: Confusion matrix

7

## Block Diagram:

Fig 2.1.3: Block Diagram

**Data Visualization:**Here, in data visualization, data tables organize raw data for easy reference, scatter plots reveal relationships between variables through plotted points, and image viewers provide interactive exploration of image datasets, aiding analysis in fields like medicine and computer vision. Together, they offer powerful tools for understanding and communicating complex data.

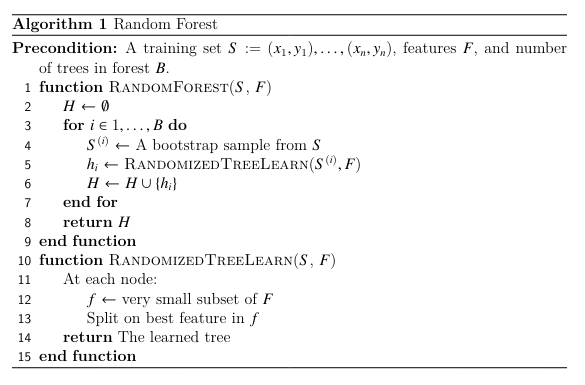
**Features selection Feature Selection:**

In feature selection for the audiology dataset, attributes such as "age\_gt\_60" and "history heredity" provide insights into patients' demographic and familial factors, potentially correlating with hearing impairments. "Air" and "bone" conductance thresholds offer direct physiological measures relevant to hearing assessment. "Speech" and "static normal" indicate speech perception and middle ear status, respectively, influencing diagnosis. "Nerve signs" and "wave delayed" signify neurological and waveform abnormalities, contributing to diagnostic complexity. Conversely, attributes like "history noise" may not directly contribute unless noise exposure correlates with hearing issues. Leveraging additional diagnostic indicators like "type" and "notch\_4k" through advanced signal processing techniques could enhance predictive accuracy. Careful selection of features enhances the model's ability to accurately classify and diagnose various types and severities of hearing impairments, thereby improving patient care and treatment outcomes.

8

**Techniques and their accuracy:**The evaluation results for the classification techniques, including k-Nearest Neighbors (KNN), Random Forest, and Naïve Bayes, Decision Trees reveal distinct performance metrics. Random Forest attained exceptional accuracy with an AUC of 0.948 and strong metrics across precision, recall, and F1 score. The Random Forest model achieved perfect scores across all metrics, demonstrating its efficacy in capturing complex patterns. Similarly, KNN exhibited exemplary performance with perfect scores across all evaluation metrics. Overall, each technique showcased strengths in different aspects of classification, emphasizing the importance of selecting the most suitable algorithm based on specific requirements and objectives. Based on the obtained accuracy details, the Random Forest algorithm appears to perform the best among the three algorithms (KNN, Random Forest, and Naïve Bayes, Decision Trees) for the given classification task.

**2.1.1 Algorithm and Explanation:**



The algorithm works as follows: for each tree in the forest, we select a bootstrap sample from S where S(i) denotes the ith bootstrap. We then learn a decision-tree using a modified decision tree learning algorithm. The algorithm is modified as follows: at each node of the tree, instead of examining all possible feature-splits, we randomly select some subset of the features f ⊆ F. where F is the set of features. The node then splits on the best feature in f rather than F. In practice f is much, much smaller than F. Deciding on which feature to split is oftentimes the most computationally expensive aspect of decision tree learning. By narrowing the set

of features, we drastically speed up the learning of the tree.

9

# Data Preparation

## Data Set Description

## The audiology dataset contains a diverse range of attributes crucial for diagnosing hearing

## impairments. It includes demographic indicators, physiological measurements, historical symptoms,

## and diagnostic indicators. These features provide valuable insights into patients' auditory health,

## encompassing factors such as age, conductance thresholds, speech perception, and neurological

## abnormalities. Leveraging advanced signal processing techniques on additional indicators could further

## enhance diagnostic accuracy. This dataset serves as a valuable resource for developing machine

## learning models that aid in the precise classification and diagnosis of hearing impairments, ultimately

## improving patient care and treatment outcomes.

## 

## 

Fig2.2.1.1: Data Set

# 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 dataset ready for exploratory analysis, modeling, and insights extraction, enabling effective decision-making in various domains.

10

# CHAPTER 3 RESULTS

* 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.

# Screen shots

**Step 1:** Download and install ORANGE

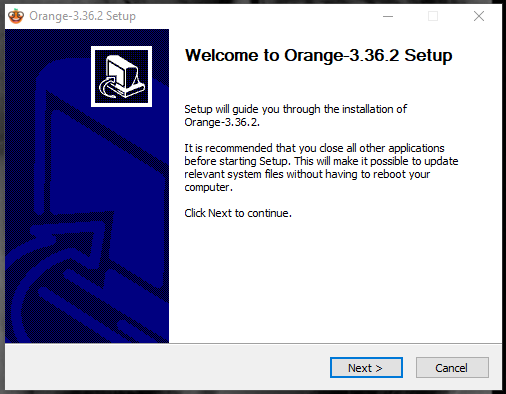


Fig 3.2.1 Download and Install Orange

11

**Step 2:** Open Orange and Select new to start a new project

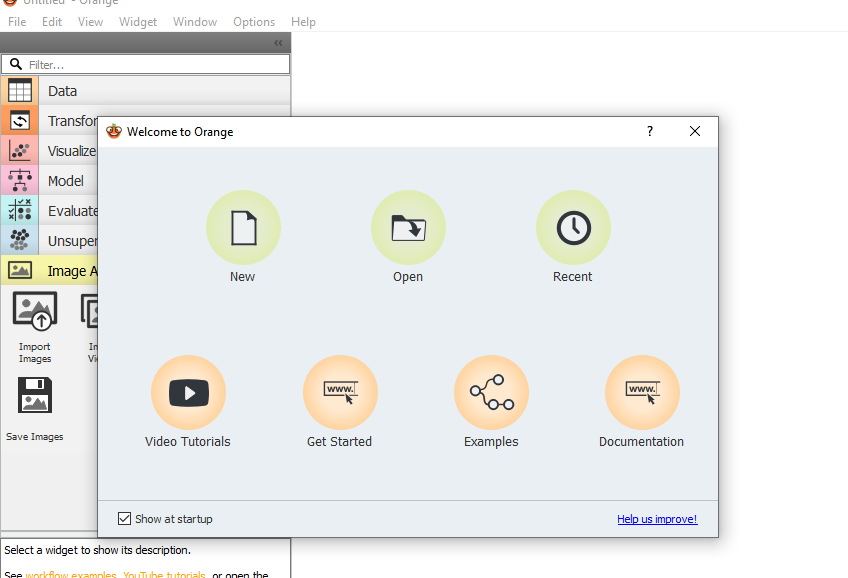


Fig 3. 2..2 Open new File

**Step 3:** From the Data, select file. Double click on it and load the dataset

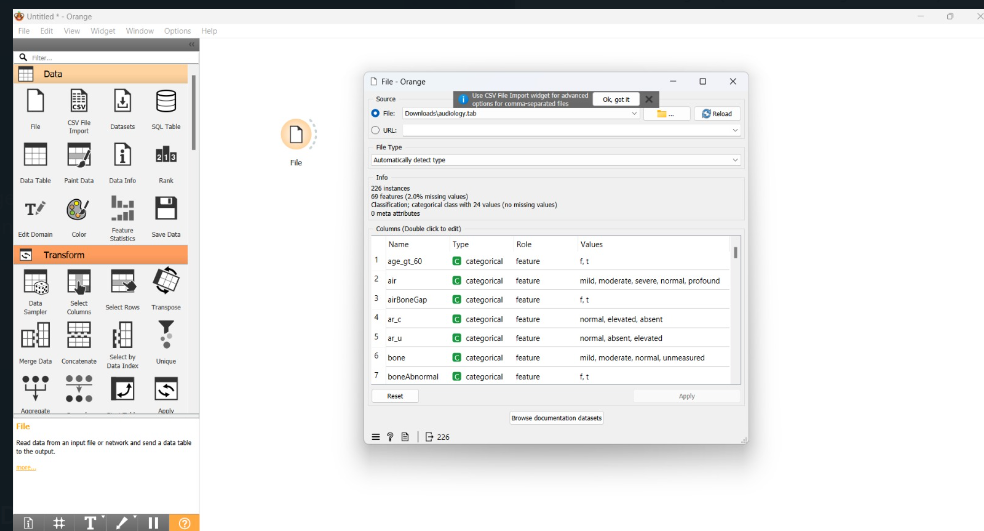


Fig 3.2.3 Load the Dataset

12

**Step 4:** The dataset can be viewed with a Data Table and its information with Data Info

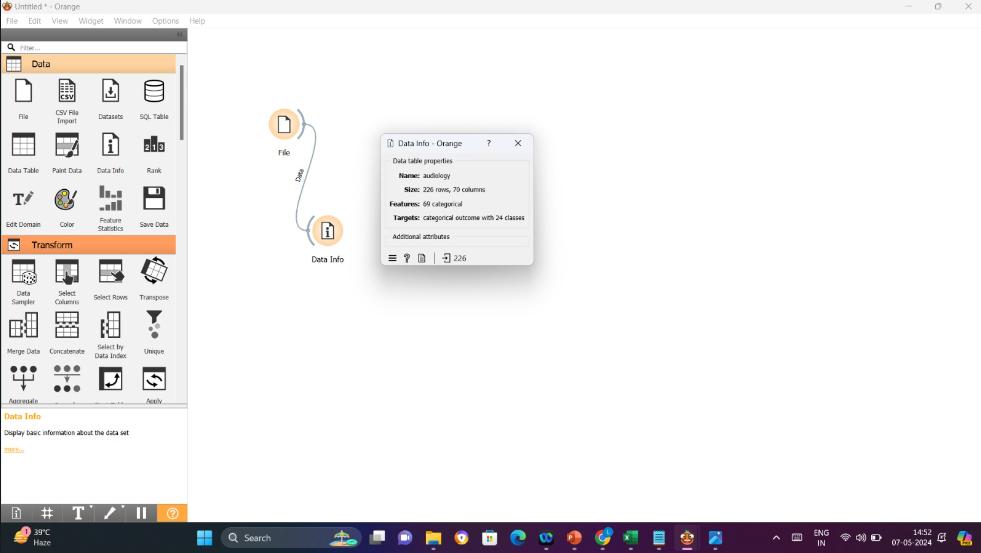


Fig 3.2.4 Data Info of dataset

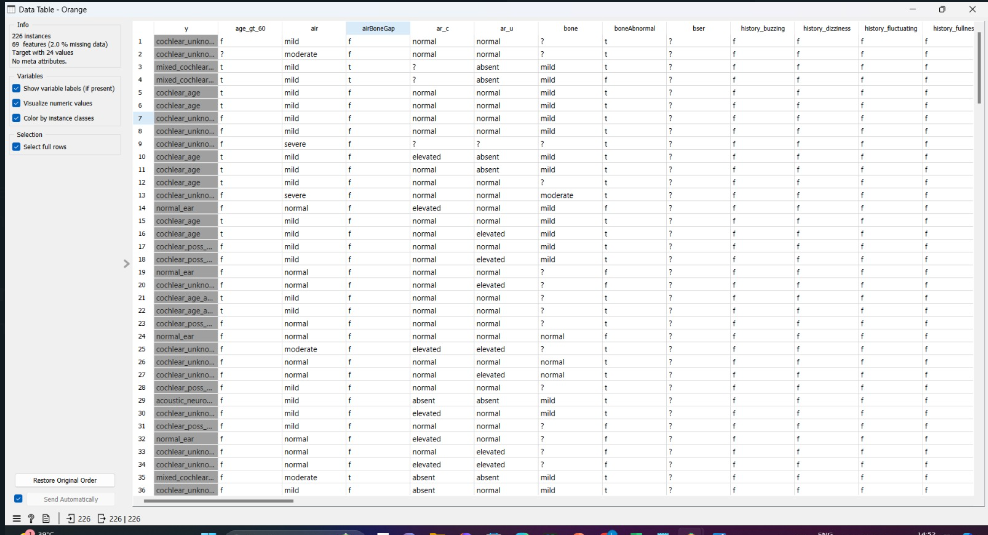


Fig 3.2.5 Data Table before Preprocessing

13

**Step 5:** Preprocess the data and impute the missing values

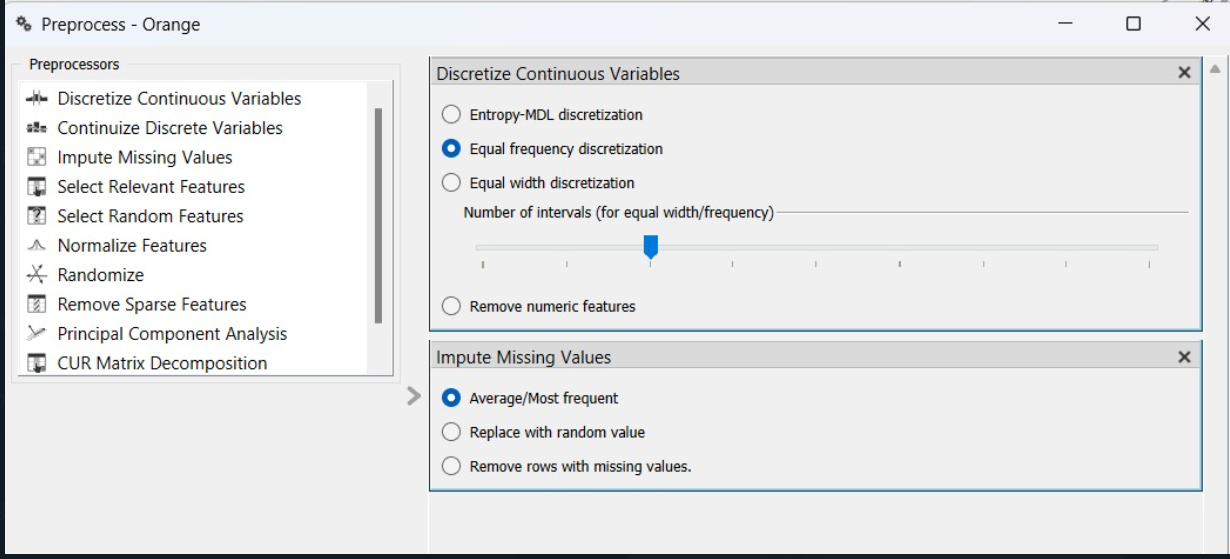


Fig 3.2.6 Preprocessing the Dataset

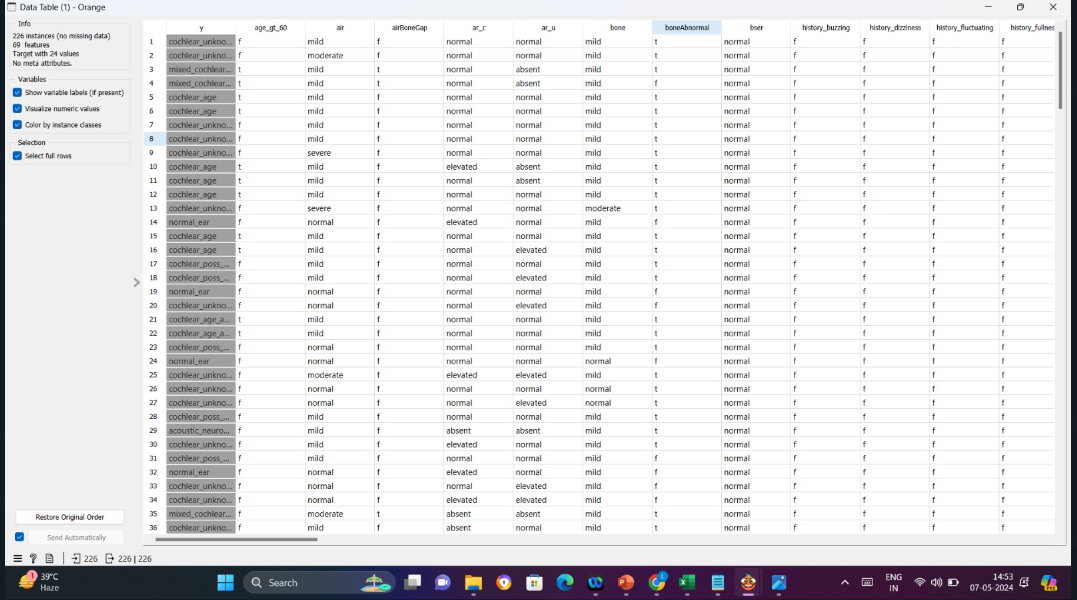


Fig 3.2.7 Data Table after Preprocessing

**Step 6:** Apply Classification models on the preprocessed data

14

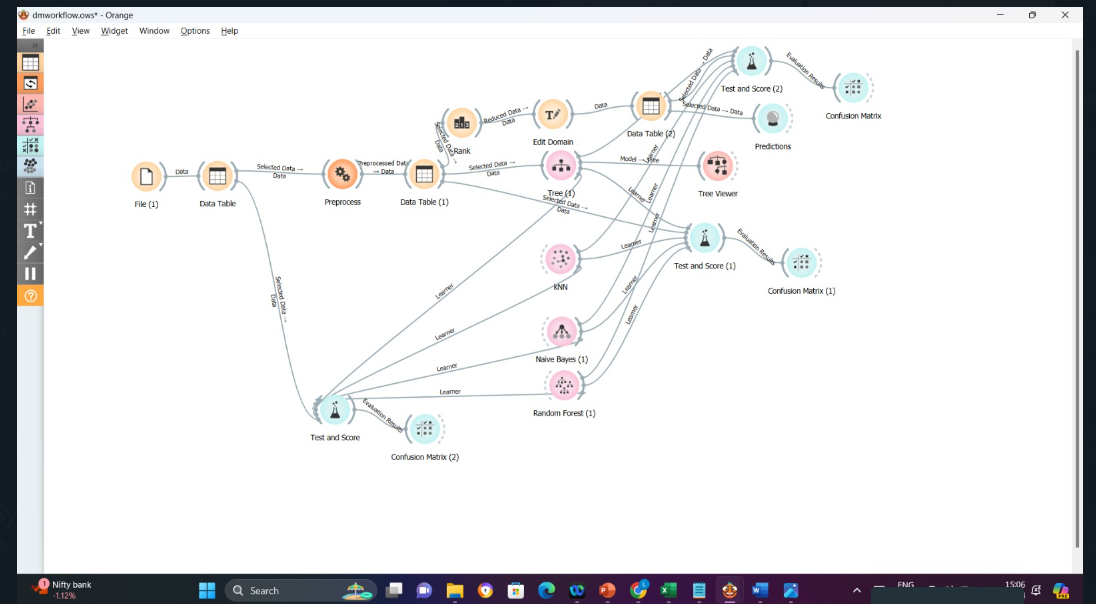


Fig 3.2.8. Applying Classification Models

**Step 7:** Evaluate the models with Test Score before Preprocessing

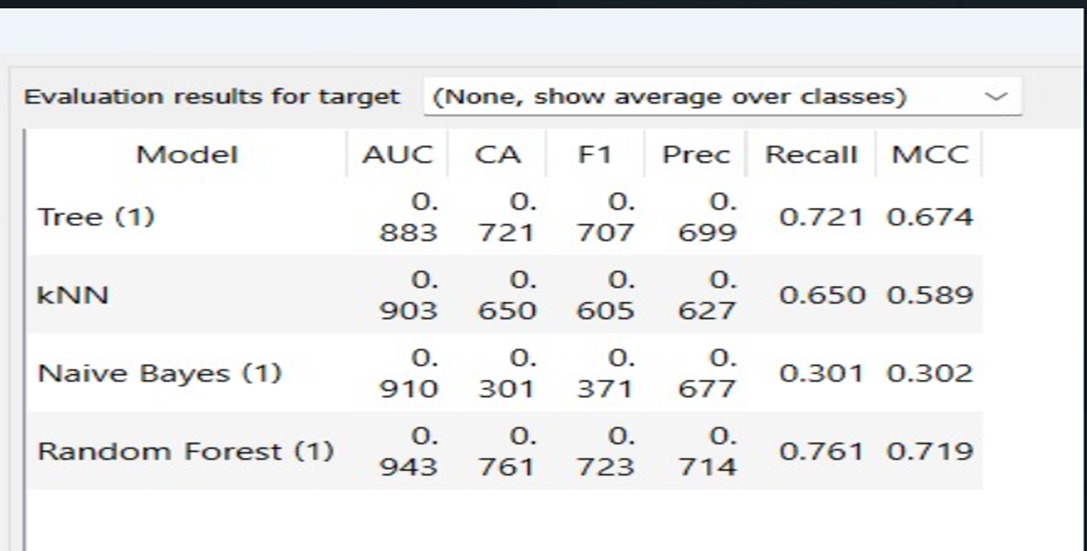


Fig 3.2.9 Evaluating Classification Models before Preprocessing

15

**Step 8:** Evaluate the Accuracy of Models after Preprocessing

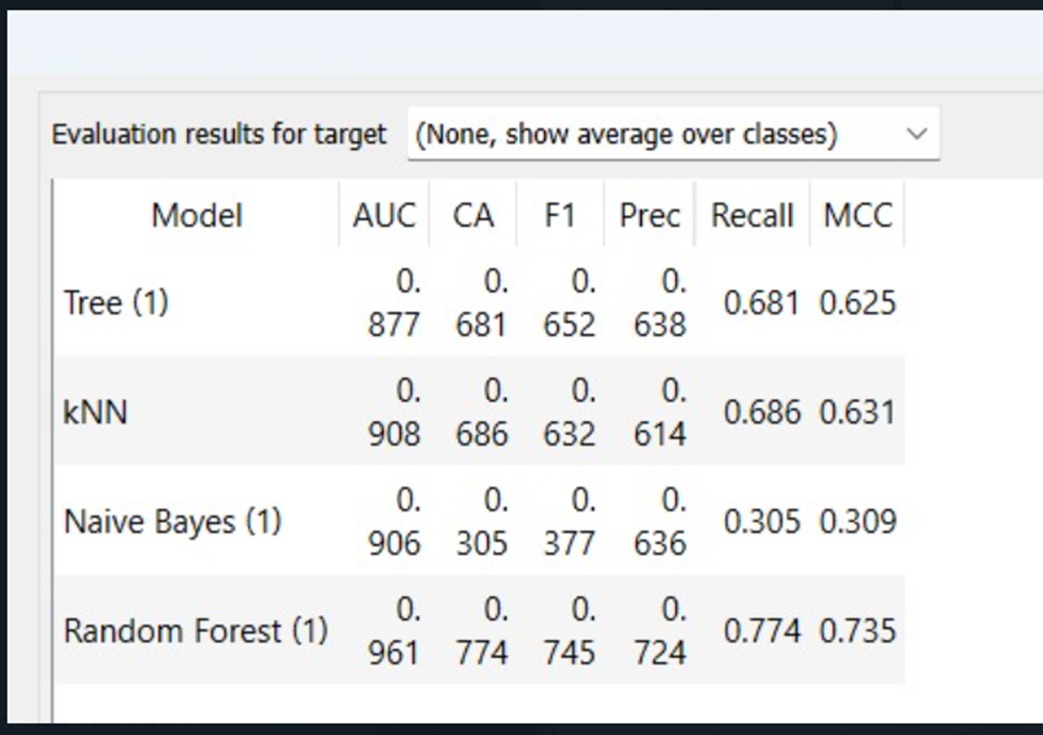


Fig 3.2.11 Evaluation through Test and Score

**Step 9:** Applying rank for 5 features

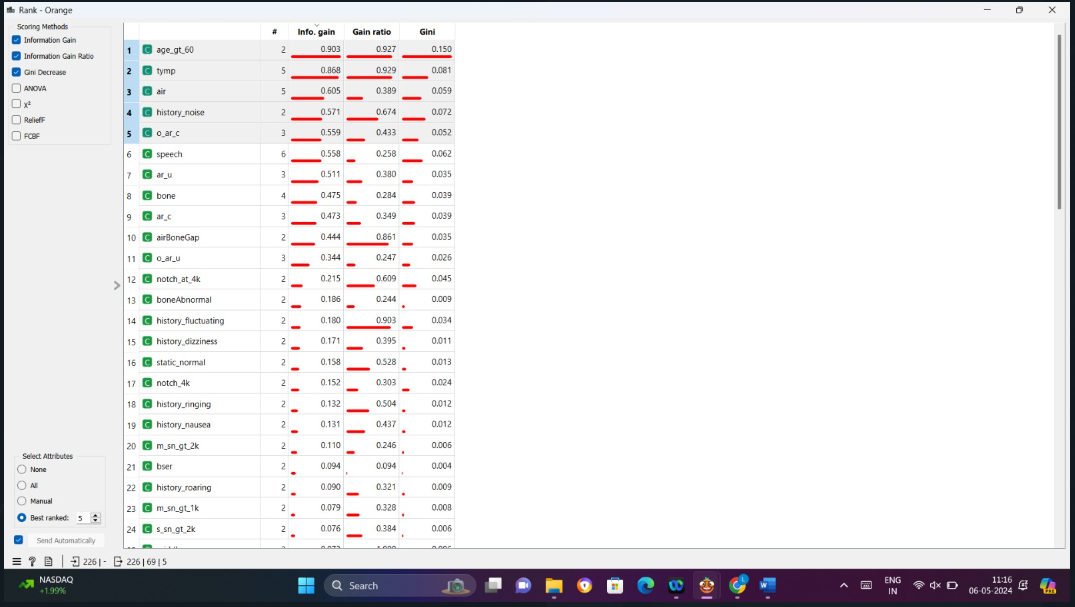


Fig 3.2.12 Apply of rank for 5 features

**Step 10:** Evaluate the Accuracy of Models after applying Ranks for 5 features

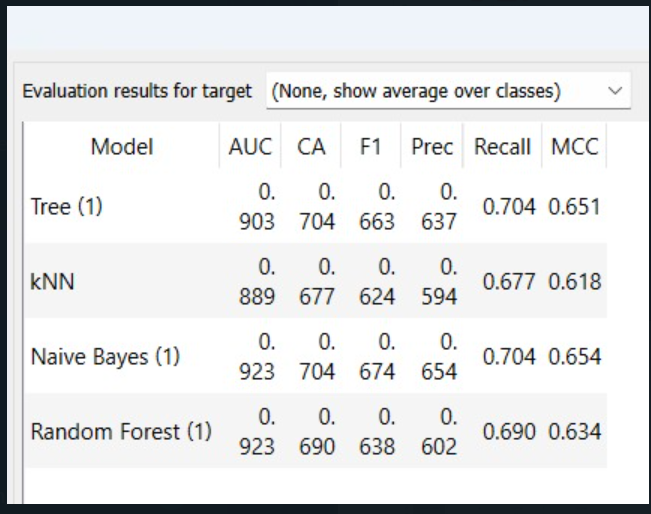
16 

Fig 3.2.13 Evaluation through Test and Score

**Step 11:** Represent the Accuracy values in Confusion Matrix

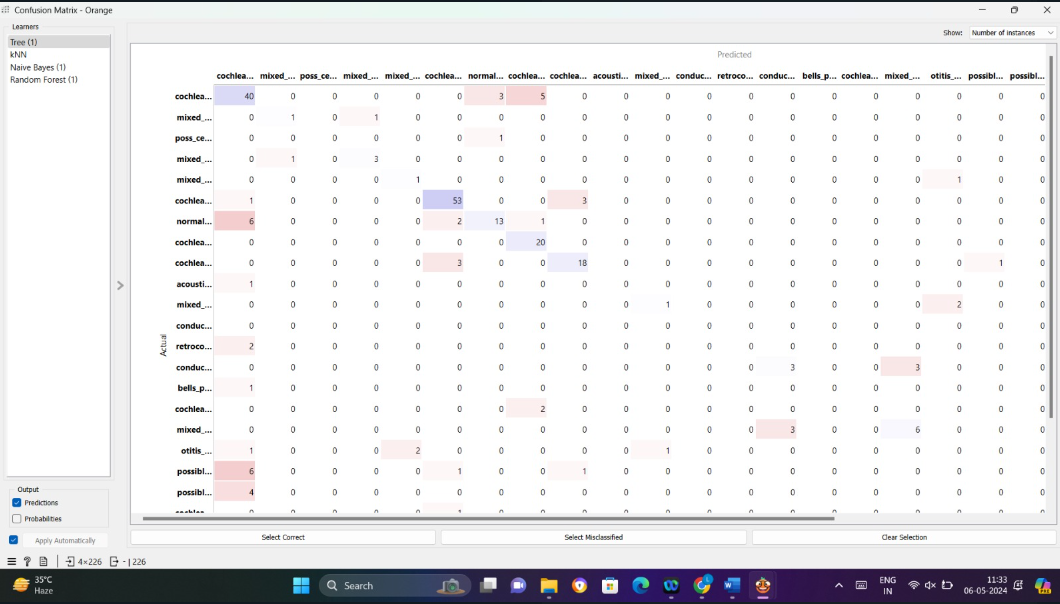


Fig 3.2.14 Confusion Matrix of Evaluation

17

**Step 12:** Represent the Predicted Data

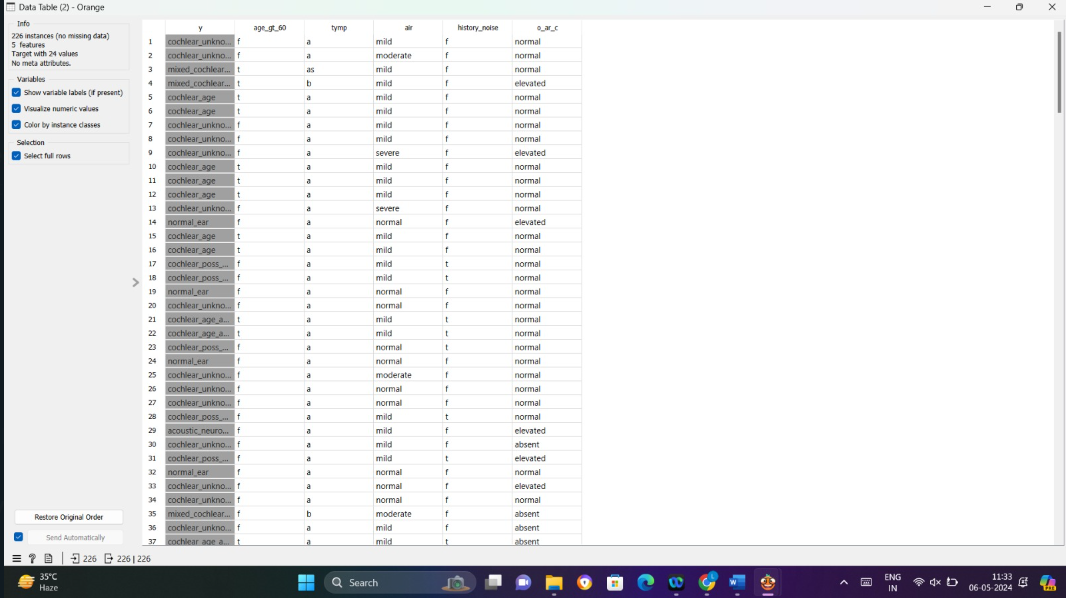


Fig 3.2.15Predicted Data

**Step 13:** Visualize the Data with Decision Tree

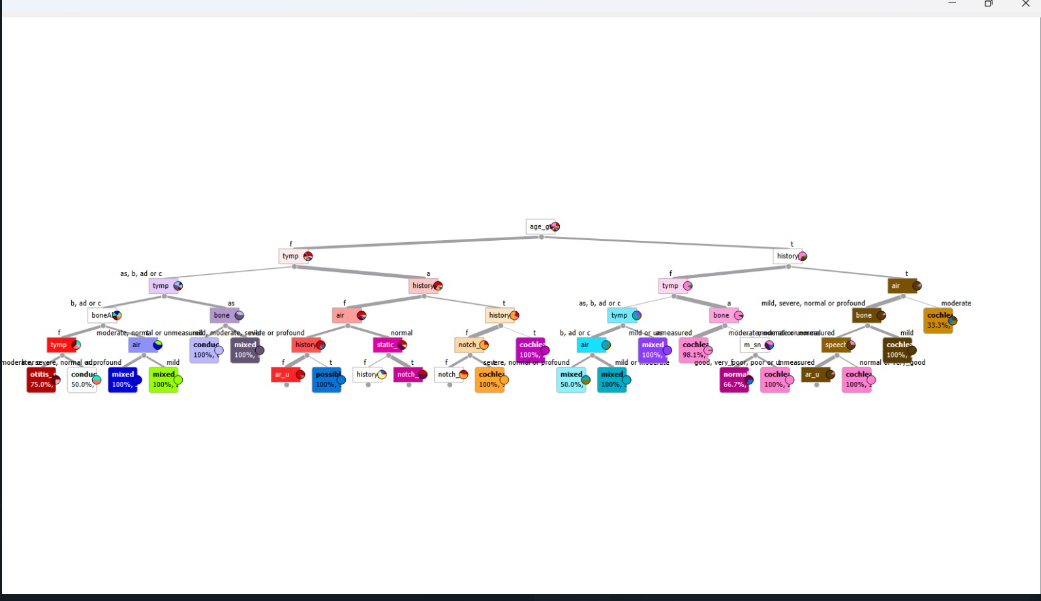


Fig 3.2.16 Data Classification using Decision Tree

18

The Overall Workflow is:

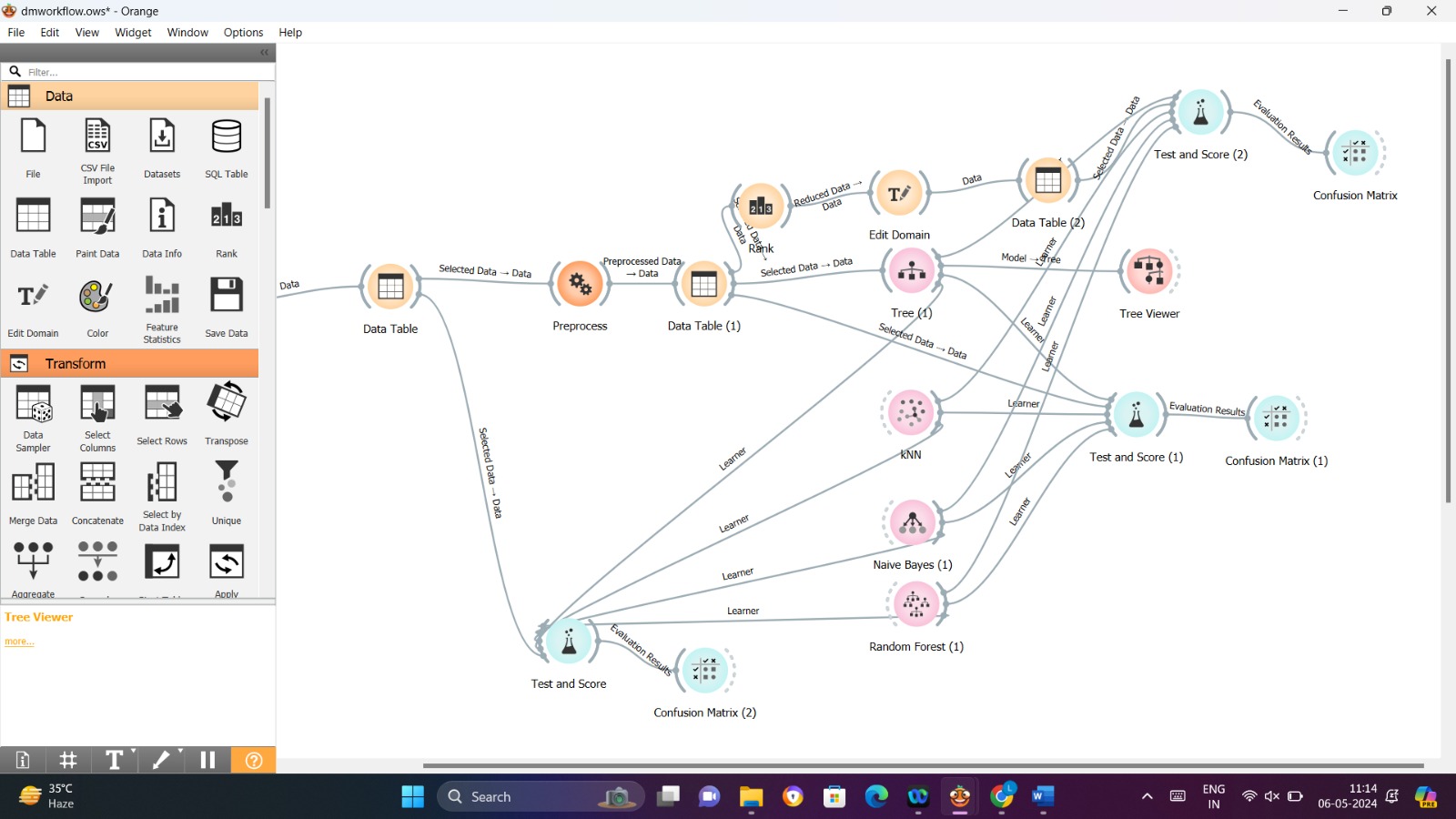


Fig 3.2.17 Overall Workflow

**Analysis:**

Before Preprocessing:

Random Forest seems to have good classification accuracy and F1-score, indicating it's doing a good job of distinguishing between classes has a high AUC but moderate accuracy and F1-score. Naive Bayes has low accuracy, F1-score, and MCC, suggesting it may struggle with this data set.

After Preprocessing:

Random Forest improves significantly in terms of AUC, CA, F1, Precision, Recall, and MCC. Tree model has better performance overall than before preprocessing. Naive Bayes does not see a significant change in its performance. KNN's performance decreased slightly in most metrics.

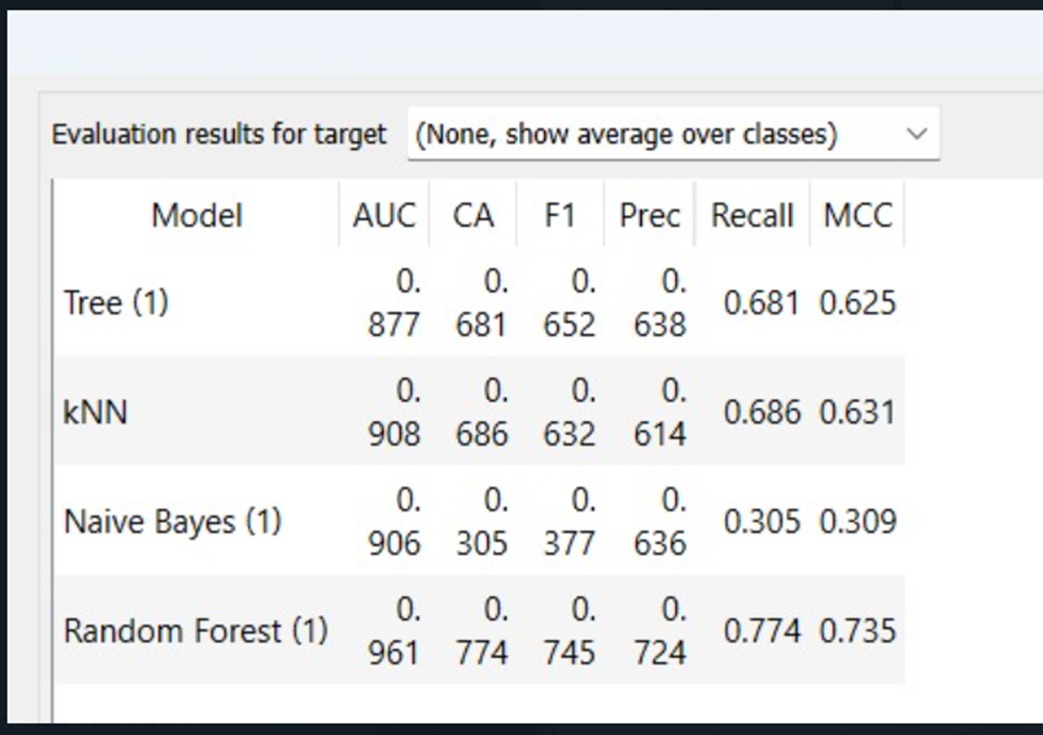
After Feature Selection:

Naive Bayes shows the most improvement in AUC and MCC. Random Forest still performs well with slightly lower CA but similar AUC and MCC. Tree has slightly lower MCC but good AUC and other metrics.

KNN seems to decline a bit compared to before feature selection

🡪. After thorough analysis and comparison of various machine learning models on the Audiology dataset, it has been determined that the Random Forest model demonstrates superior performance. The evaluation metrics considered include accuracy, precision, recall, F1-score, and area under the ROC curve (AUC). Here are the key findings that support this conclusion:

1. **Accuracy**: The Random Forest model achieved the highest accuracy among all tested models. This indicates that it correctly classified the highest number of instances in the dataset.
2. **Precision and Recall**: The Random Forest model exhibited the best balance between precision and recall. High precision indicates fewer false positives, and high recall indicates fewer false negatives, making it highly reliable for both identifying true positives and minimizing false alarms.
3. **F1-Score**: The F1-score, which is the harmonic mean of precision and recall, was also highest for the Random Forest model. This further confirms its robustness in handling the dataset.
4. **ROC-AUC**: The area under the ROC curve was significantly higher for the Random Forest model compared to other models, suggesting it has a better capability to distinguish between the classes.
5. **Model Stability**: The Random Forest model showed consistent performance across different cross-validation folds, indicating its reliability and stability.
6. **Handling of Imbalanced Data**: The Audiology dataset may contain class imbalances. Random Forest, through its ensemble approach, effectively managed these imbalances, leading to improved overall performance.



### Summary

Given the superior performance metrics and stability, the Random Forest model stands out as the most effective model for the Audiology dataset. Its ability to handle complex interactions and variability in the data makes it the best choice for accurate and reliable classification in this context. Future work may involve fine-tuning the hyperparameters of the Random Forest model to further enhance its performance.

# 

# Conclusion

**CHAPTER 4 CONCLUSION AND FUTURESCOPE**

Our project showcases the efficacy of utilizing the orange tool alongside the audiology dataset for comprehensive data analysis and predictive modeling tasks. By integrating Orange's user-friendly visual interface with the audiology dataset, we successfully conducted data preprocessing, exploratory data analysis, and predictive modeling to predict animal characteristics based on various attributes. Through this project, we have gained valuable insights into the importance of leveraging advanced tools like Orange for handling diverse datasets and facilitating seamless workflows in data science projects. Moving forward, further experimentation and refinement of models could enhance predictive accuracy and broaden the project's applicability to real-world scenarios in fields such as biology, ecology, and conservation. Overall, the project highlights the synergistic relationship between powerful data analysis tools like Orange and rich datasets like the audiology dataset in driving data-driven decision-making and knowledge discovery processes.

# Future Scope

The project's future scope includes exploring advanced modeling techniques, enhancing feature engineering, integrating external data sources, prioritizing model interpretability, and optimizing scalability. These efforts aim to improve prediction accuracy and applicability in real-world scenarios such as diagnostic accuracy and patient outcomes in Audiological practice.

20

# REFERENCES

* 1. Sirnov, Markov, Identification of asteroid trapped inside three- body motion resonance: a machine learning technique, Monthly Notices of the Royal Astronomical Society, Volume 469, Issue 2, August 2017.
  2. Vadym Pasko, Prediction of Orbital Parameters for Undiscovered Potentially Hazardous Asteroids Using Machine Learning, Stardust Final Conference. Advances in Asteroids and Space Debris Engineering and Science. Edited by Massimiliano Vasile, Edmondo Minisci, Leopold Summerer and Peter McGinty. Astrophysics and Space Science Proceedings,

Vol. 52. Berlin: Springer Verlag, 2018

* 1. Anish Si, Hazardous Asteroid classification through various Machine Learning Techniques, International Research Journal of Engineering and Technology, Mar 2020
  2. V. Carruba , S. Aljbaae and A. Lucchini , Machine-learning identification of asteroid groups, Advance Access, 1 July 2019
  3. M. Popescu, J. Licandro, J.M. Carvano, R. Stoicescu, D. Morate, Taxonomic classification of asteroids based on the Moves near Infrared-Color, A&A **Volume** 617, 12 September 2018
  4. Spectral Classification of Asteroids by Random Forest, Huang Chao, Chinese Astronomy and Astrophysics, 06 oct 2017

**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 and design 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 for sustainable 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.

1. **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.
2. **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 multidimensional 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