STATISTICAL STUDY OF DIFFERENT

DISEASE SYMPTOMS

**Project submitted in partial fulfilment of the requirement**

**For the degree of**

**Master of Computer Application By**

Sukesh kumar Dalai (2261020032)

[pal2001suman@gmail.com](mailto:pal2001suman@gmail.com)

Under the guidance

Dr. Saswati Mahapatra

Assistant Professor

Department of Computer Application



Institute of Technical Education and Research

Siksha ‘O’ Anusandhan Deemed to be University

2023

**Title-** **STATISTICAL STUDY OF DIFFERENT DISEASE SYMPTOMS**

**Abstract:**

Disease symptom and patient profile dataset involves summarizing the dataset at a higher level to capture key information while maintaining privacy and confidentiality. Each disease in the dataset is assigned a unique identifier or label, such as "Disease Common Cold", "Disease diabetes”, “Disease Liver Cancer" etc. The specific details of the diseases, including their names, causes, and treatments, are not included in the abstraction. Symptoms associated with each disease are grouped into categories. For example, symptoms like fever, cough, and sore throat can be grouped under the category "Respiratory Symptoms." This abstraction provides a higher-level view of the symptoms without revealing individual symptoms or their detailed descriptions.

Instead of providing specific details about each patient, the patient profiles can be abstracted to capture demographic information and relevant characteristics. This may include age ranges (e.g., 19-40 years), gender distribution (e.g., male, female).The dataset can be aggregated to provide summary statistics and frequencies for each disease, symptom category, and patient profile. This abstraction helps identify patterns and trends without disclosing sensitive information about individual patients.

**Submitted By: Guided By:**

**Sukesh kumar Dalai (2261020032) Dr. Saswati Mohapatra**

**Dept. of CA**

**INTRODUCTION:**

Disease Prediction using Machine Learning is a system which predicts the disease based on the information provided by the user. It also predicts the disease of the patient or the user based on the information or the symptoms he/she enter into the system and provides the accurate results based on that information. If the patient is not much serious and the user just wants to know the type of disease, he/she has been through. It is a system which provides the user the tips and tricks to maintain the health system of the user and it provides a way to find out the disease using this prediction. Now a day's health industry plays major role in curing the diseases of the patients so this is also some kind of help for the health industry to tell the user and also it is useful for the user in he/she doesn't want to go to the hospital or any other clinics, so just by entering the symptoms and all other useful information the user can get to know the disease he/she is suffering from and the health industry can also get benefit from this system by just asking the symptoms from the user and entering in the system and in just few seconds they can tell exact and up to some extent the accurate diseases. This DPUML is previously done by many other organizations but our intention is to make it different and beneficial for the users who are using this system.

This Disease Prediction Using Machine Learning is completely done with the help of Machine Learning and Python Programming language and also using the dataset that available previously by the hospitals using that we will predict the disease. Now a day's doctors are adopting many scientific technologies and methodology for both identification and diagnosing not only common disease, but also many fatal diseases. The successful treatment is always attributed by right and accurate diagnosis. Doctors may sometimes fail to take accurate decisions while diagnosing the disease of a patient, therefore disease prediction systems

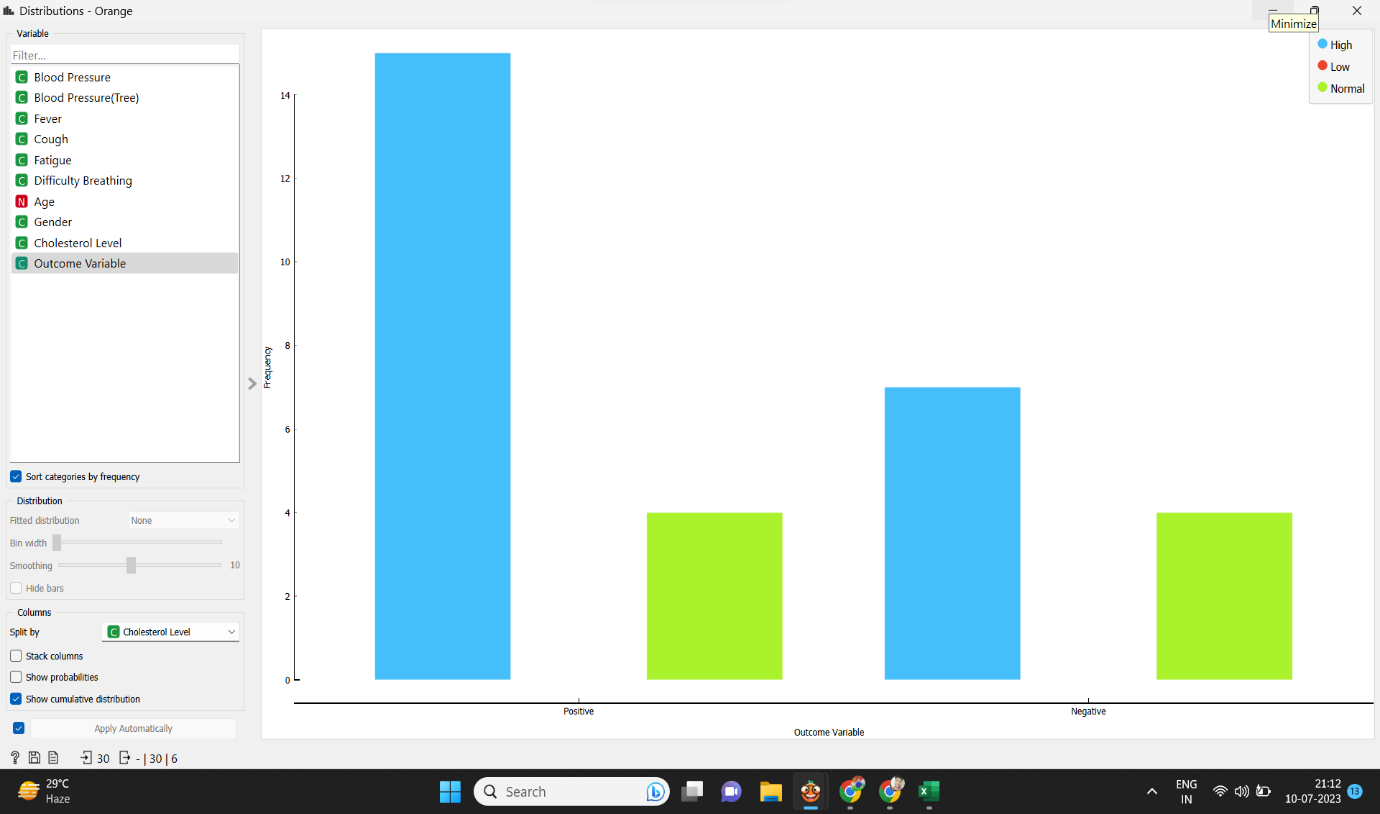


Fig. 1- Distribution Graph of Training Dataset

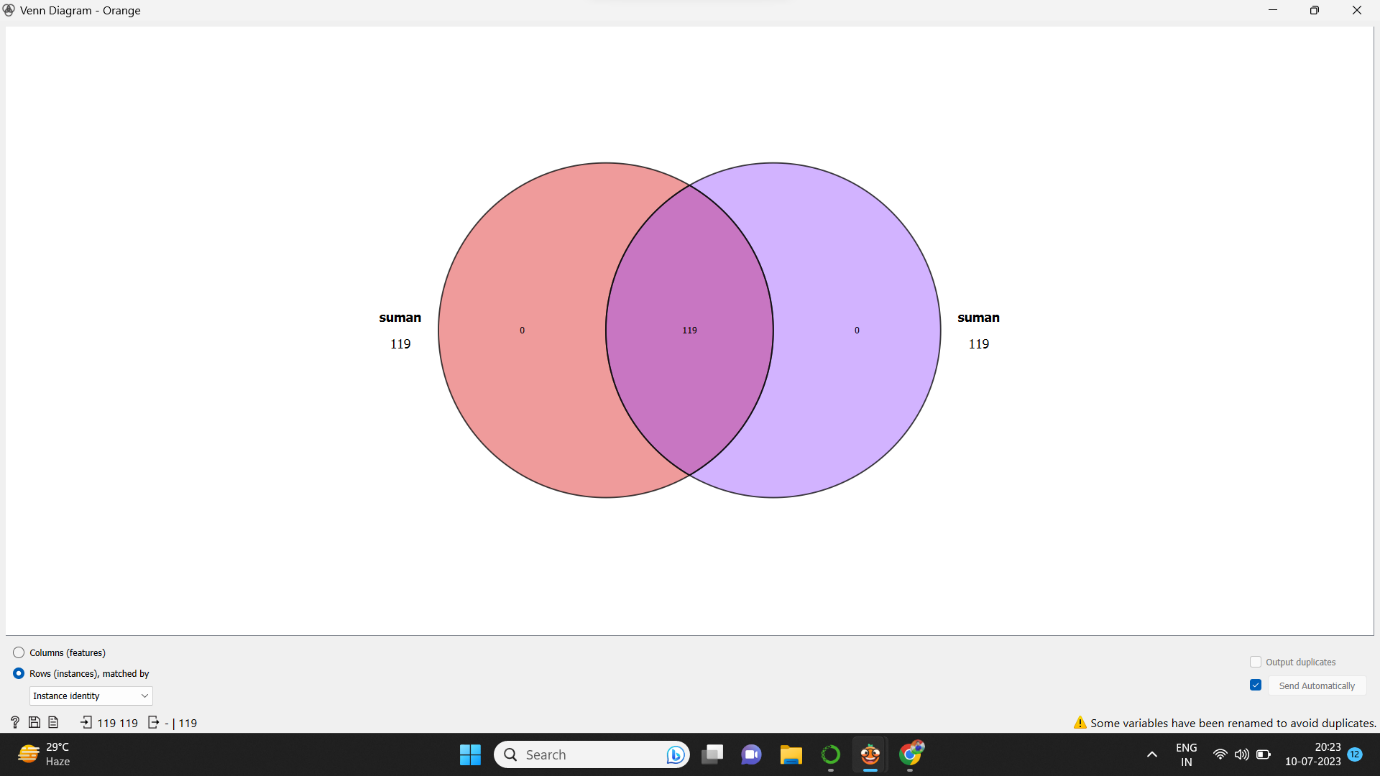


Fig. 2- Venn diagram of Training Dataset

**BACKGROUND STUDY:**

A disease symptoms and patient profile dataset provides a valuable resource for studying the relationship between symptoms exhibited by patients and their corresponding profiles. Here are some key aspects to consider when conducting a background study on such datasets:

Data Collection: Understand the methodology and sources of data collection for the dataset. The data may be collected through various means such as electronic health records, surveys, clinical trials, or real-time monitoring systems. Assess the reliability and representativeness of the data sources to ensure the dataset's quality.

Disease Classification: Familiarize yourself with the classification system used to categorize diseases in the dataset. This could be based on international disease coding systems like ICD-10 (International Classification of Diseases, 10th edition) or a custom classification scheme specific to the dataset. Understanding the disease classification helps in interpreting and comparing the findings across studies.

Symptom Identification: Gain insights into how symptoms are identified and recorded within the dataset. This may involve reviewing medical records, patient-reported data, or clinical assessments. Consider the level of granularity in symptom descriptions and whether they capture a comprehensive range of symptoms associated with each disease.

**OBJECTIVE:**

Currently, the scenario is if the patient is suffering from any symptoms then he/she must visit to the doctor or to the hospital to diagnose the disease. But, our main objective is to reduce such efforts taken by patients only to diagnose the disease. Many patients are losing their life only because of the late diagnosis of their disease. So our main aim is to reduce such deaths.

**METHOD:**

We have taken some manual input of data from the market survey and fetch the dataset to the orange tool in file. Applying different algorithms of Data Mining, we are going to perform Data mining task using orange tool widgets which are basically of four types i.e., Data input, Classification, Visualization, and Association widgets.

**DATASET:**

Here we have taken some instances and attributes.

• Number of instances = 119

• Number of attribute = 10

The attributes are:

|  |  |  |
| --- | --- | --- |
| **ATTRIBUTES** | **TYPE** | **RANGE** |
| Disease | Categorical |  |
| Fever | Categorical |  |
| Cough | Categorical |  |
| Fatigue | Categorical |  |
| Difficulty Breathing | Categorical |  |
| Gender | Categorical |  |
| Blood Pressure | Categorical |  |
| Cholesterol level | Categorical |  |
| Age | Numerical | 19-40 |
| Outcome Variable | Categorical |  |

**PROPOSED MODEL:**

In the proposed model we have discussed the methodologies in which we have compared the classification model. Using Orange data mining tool, the workflow diagram of the steps involved in the study are shown in fig. 3. We are going to explain each process involved in the proposed work from the workflow diagram in the following sub-section.

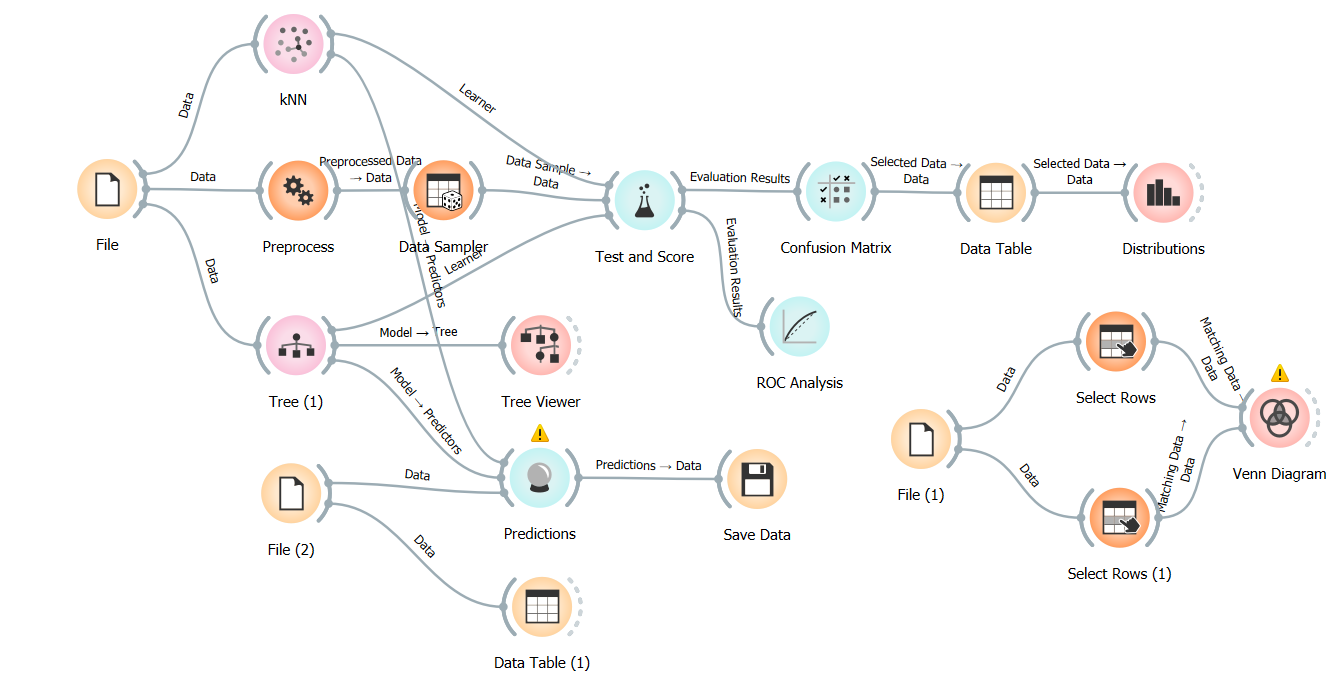
****

Fig. 3- Work flow diagram of the steps involved in the sleep disorder

**TRAINING DATASET:**

**File:**

We have collected a dataset of both categorical and numeric value and imported it into the workspace. The dataset has been uploaded to the orange tool through file widget. In this we have taken “Outcome” as target attribute (the main objective of our project) values as – positive or negative.

No of instances: 120

No of features: 10

**Data Table:**

We connect the file to the data table to visualize the data which are present in the training dataset.

**Algorithms:**

Decision trees are constructed in a top-down recursive divide and conquer manner and most of the algorithms of decision tree induction also follows the top-down approach which generates decision tree. 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 partition the data tuples into individual classes. Splitting attribute and possibly a splitting subset are consisting the criterion.

**NAIVE BAYES:**

Bayesian classifiers are statistical classifiers. They can predict class membership probabilities such as the probability that a given tuple belongs to a particular class. Naïve bayes classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attribute. It is made to simplify the computation involved with the data are given and it allows the representations of dependencies among subsets of attributes.

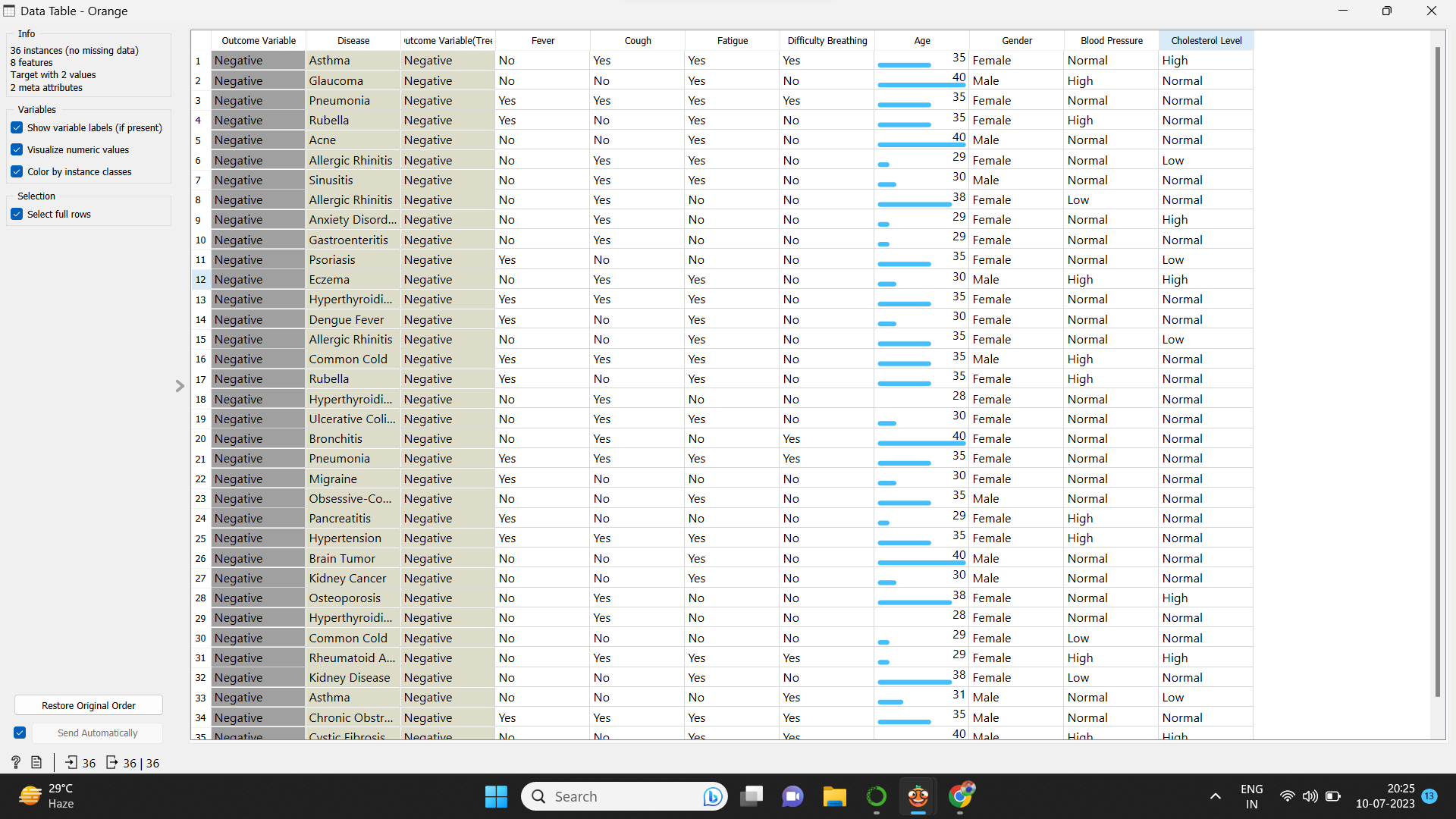


Fig-4: Data table of Training Dataset

**TEST DATASET**

File:

We have selected the test dataset from the training dataset of both categorical and numeric value and imported it into Orange tool through File widget.

No of instances: 8

No of features: 14

Data Table:

We connect the file to the data table to visualize the data which are present in the training dataset.

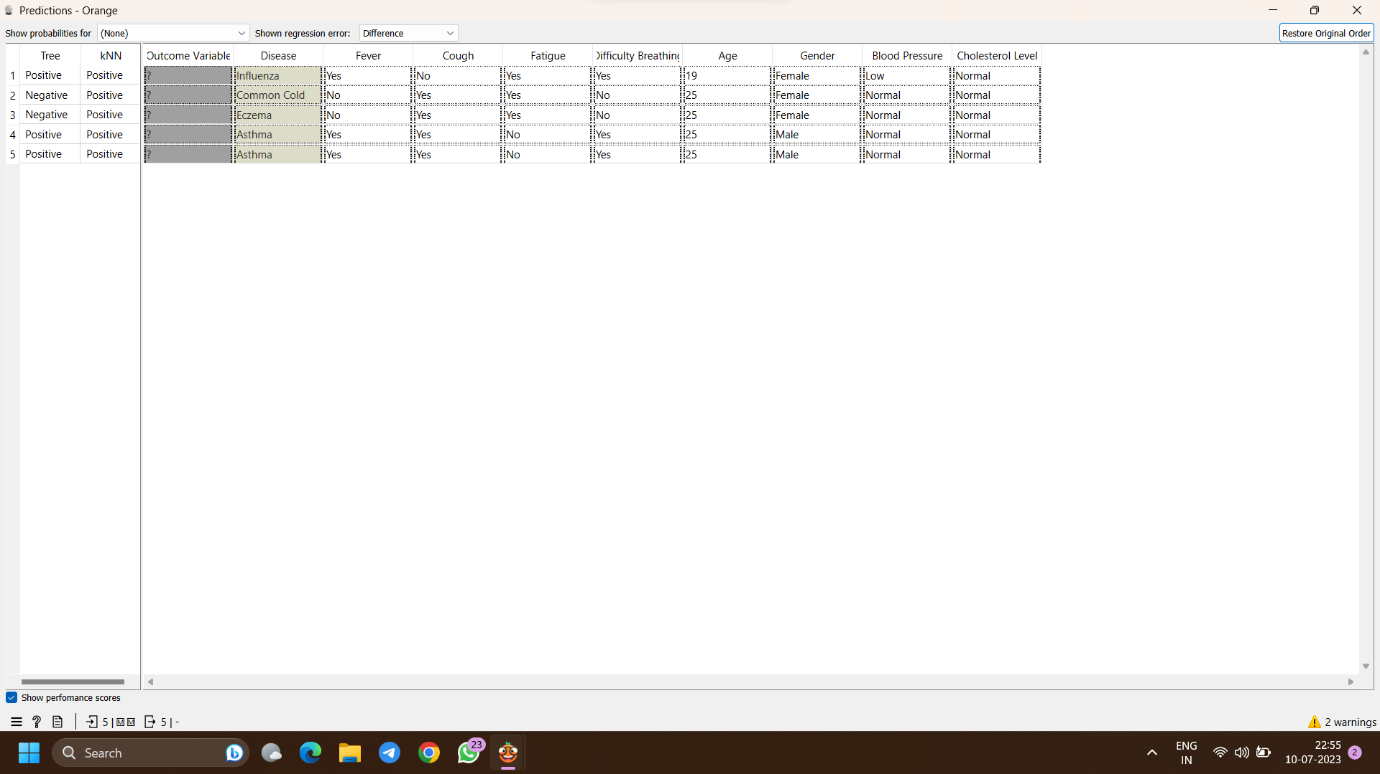


Fig-5: Data table of Training Dataset

**RESULT ANALYSIS:**

**TREE**:

In the particular experiment the tree is binary induced with minimum number of leaf instance to be 2, do not split subsets smaller than 2, limit the maximal tree depth 66, stop when majority reaches 66%., results are:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | AUC | CA | F1 | Precision | Recall |
| Tree | 0.954 | 0.867 | 0.866 | 0.867 | 0.867 |

**PREDICTION:**

We have connected the files containing the training dataset and test dataset to prediction widget to find out the result of prediction. We need to select the features which we want to be shown as output. We have connected the prediction widget to data table and confusion matrix. The prediction widgets show the probabilities and final decision of the proposed model in a data table if the model predicted the true class values correctly and the confusion matrix shows accuracy of the proposed model.

**TEST AND SCORE:**

Test and score widget takes these classifier model as input and gives a comparative value on the basis of CA, precision, recall, F-1 score and AUC.

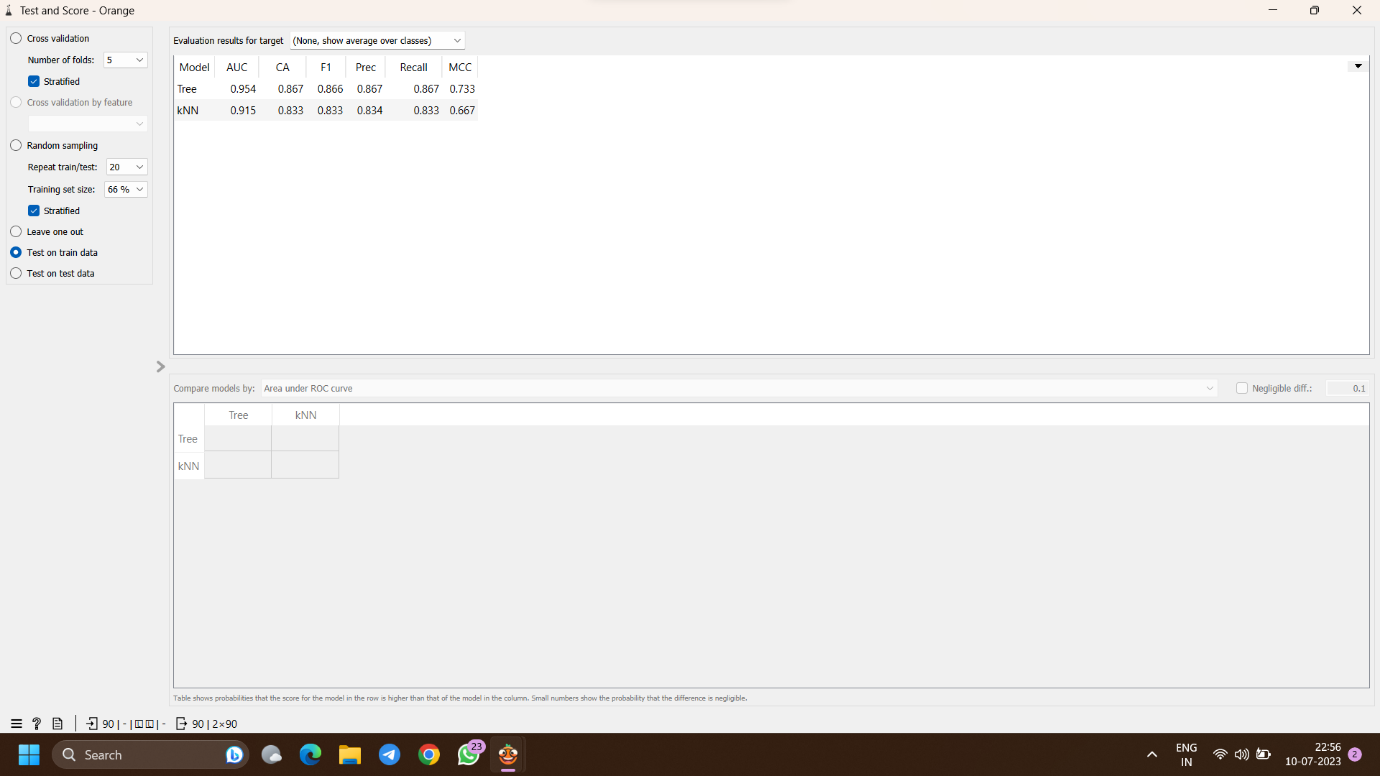
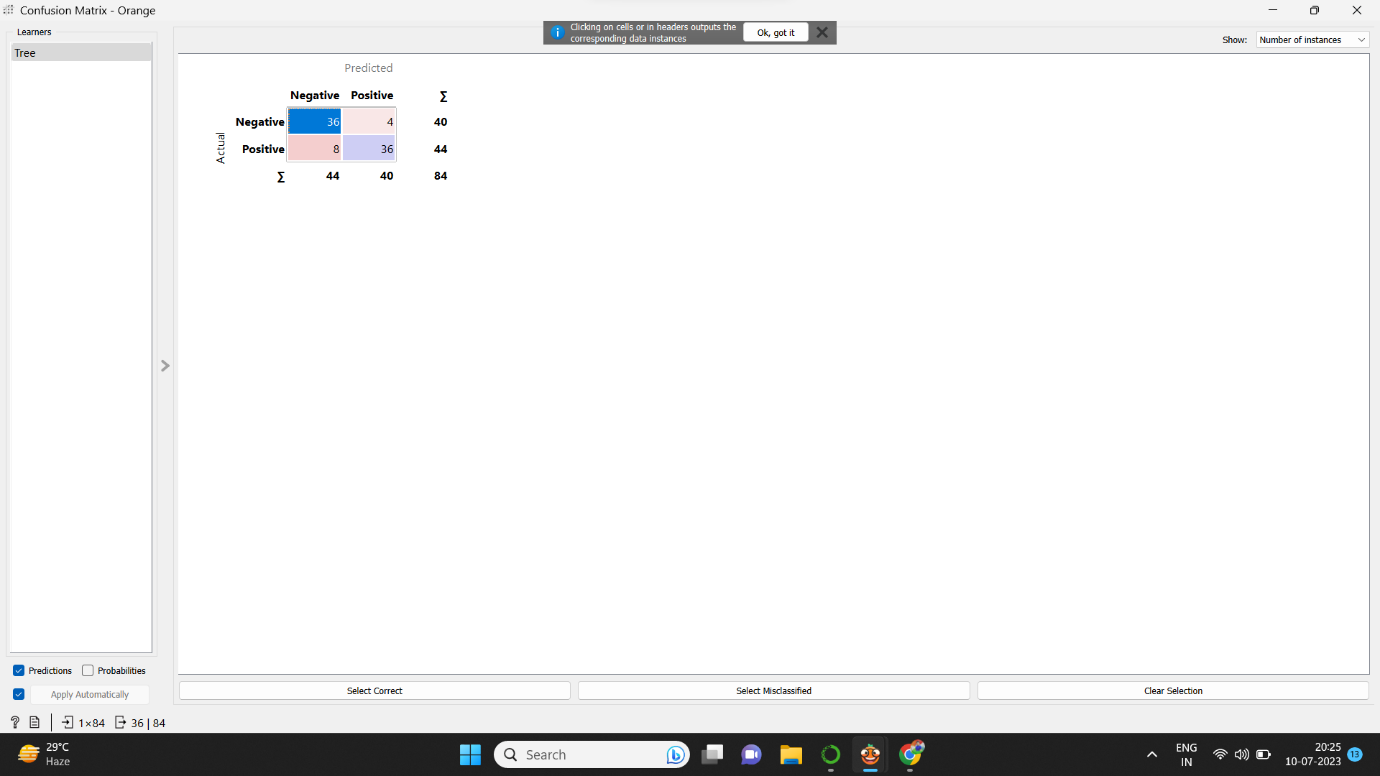


Fig-5: Result Analysis

**CONFUSION MATRIX:**

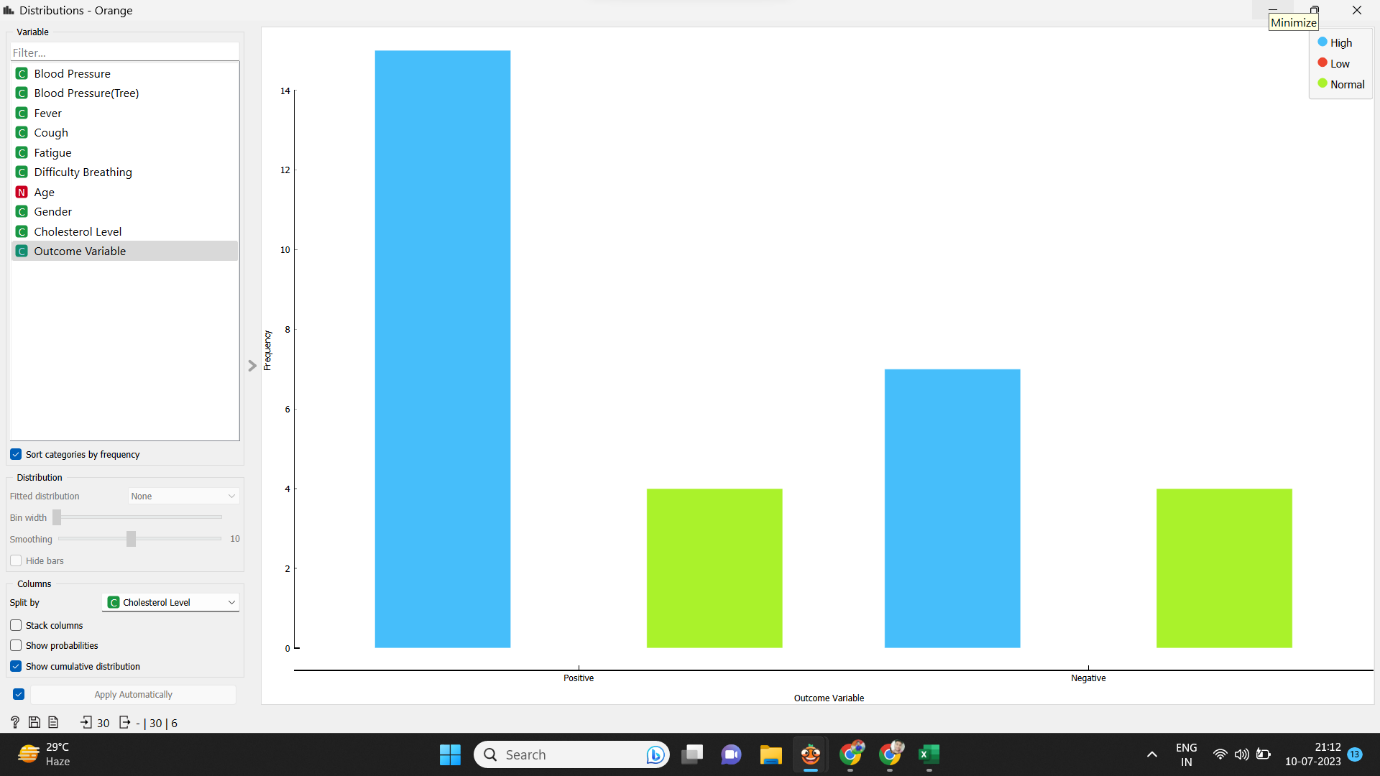
Test and score widget is connected to ‘confusion matrix’. It classifies all the classifying model used and gives Classification Accuracy (CA).

Fig. 6- Confusion matrix of tree



**DISTRIBUTION WIDGET:**

The values of distribution discrete and continuous attributes are displayed through Distribution widgets in Orange tool Distribution widget. Here the distribution widget shows which phone is easy affordable through a graph view.



## **Conclusion**

**Data mining techni**ques helps in finding the hidden knowledge in a group of disease data that can be used to analyse and predict the future behaviour of diseases. Classification is one the data mining techniques which assigned a class label to a set of unclassified cases. The main objective of this paper is to compare the data mining tools on the basis of their classification accuracy. According to the result of three data mining tools used in this paper, it has been observed that different data mining tools are furnishing different results on same data set with different classification algorithm. WEKA is showing best classification accuracy when compared to rapid miner and orange. In future, more disease dataset can be used for classification techniques and other data mining techniques such as clustering can be used to compare the performance of various data mining tools.

**REFERENCE:**

[**https://impact.dbmi.columbia.edu/~friedma/Projects/DiseaseSymptomKB/index.html**](https://impact.dbmi.columbia.edu/~friedma/Projects/DiseaseSymptomKB/index.html)

[**https://en.wikipedia.org/wiki/Sleep\_disorder**](https://en.wikipedia.org/wiki/Sleep_disorder)

[**https://www.kaggle.com/datasets/uom190346a/disease-symptoms-and-patient-profile-dataset**](https://www.kaggle.com/datasets/uom190346a/disease-symptoms-and-patient-profile-dataset)

[**https://www.researchgate.net/publication/289842192\_Orange\_Data\_mining\_fruitful\_and\_fun\_-\_A\_historical\_perspective**](https://www.researchgate.net/publication/289842192_Orange_Data_mining_fruitful_and_fun_-_A_historical_perspective)