

**Car Evaluation**

**Course Name:** Data Warehousing and Data Mining

**Section-** D

**Course Instructor-** Akinul Islam Jony

**Submitted By-**

|  |  |
| --- | --- |
| **NAME** | **ID** |
| Md Rakib Hasan | 20-43540-1 |
| TASRIFUL HAQUE | 18-38661-3 |
| SHAMIUR RAHMAN SHAKIR | 20-43543-1 |
| IFFAT ZAMAN FARIHA NIRJONA | 18-39225-3 |

1. **INTRODUCTION**

Understanding the idea in making a decision on a choice in getting a car is basic to everybody particularly the first time buyer or anyone who are inexperienced in how the car business functions. Generally we need a car as a methods for transportation however as we include fun into it and we tend to forget that we shouldn’t underestimate. Classifying a good car from a better than average to a terrible one are normally being finished physically with the assistance of car sales representative who guides us to purchase this along these lines or from the conclusion of our family and companions who had past experienced with vehicle inconveniences

In present times it is continuously the car sales representative who encourages us to purchase this car or not. We may or probably won’t know it consciously however we are basically ignoring the factors that would help us financially, comfortably, and safety in a long run.

3 popular classification models were used in our project. - Naïve Bayes, K-Nearest Neighbor (KNN) and Decision Tree. These are some popular classification algorithms that can be used for classifications.

**Naïve Bayes Algorithm:** It is a classification technique based on Bayes’ Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

**K-Nearest Neighbor (KNN):** K Nearest Neighbor algorithm falls under the Supervised Learning category and is used for classification (most commonly) and regression. It is a versatile algorithm also used for imputing missing values and resampling datasets.

**Decision Tree Algorithm:** The goal of this algorithm is to create a model that predicts the value of a target variable, for which the decision tree uses the tree representation to solve the problem in which the leaf node corresponds to a class label and attributes are represented on the internal node of the tree.

* 1. **Project Objective**

The objective of this report is especially to determine the decision making, identifying the car variables like car price value with other various variable to decide between a good acceptable cars from the unaccepted values from the target value.

* 1. **Description**

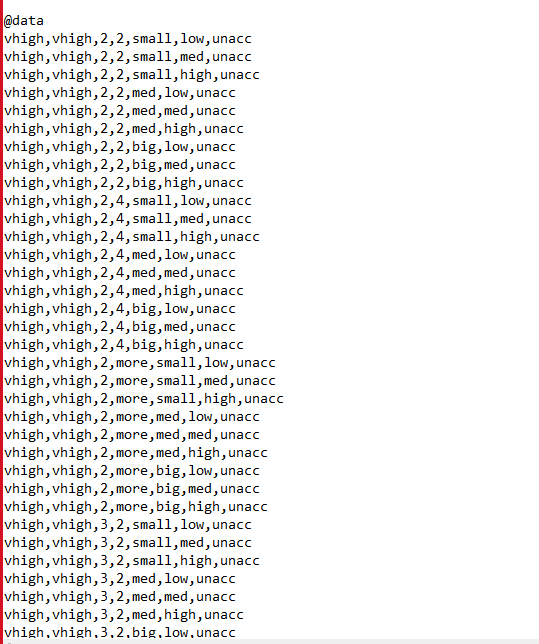
When an individual considers buying a car, there are many aspects that could affect his/her decision on which type of car he/she is interested in. Factors such as the prize, the regular maintenance, comfort, and safety are few vital issues that we need to consider. The car evaluation dataset is very important structural information that someone should look at for the features and classification of each car could be helpful in our decision making. Because of its underlying concept structure, this project is highly informative particularly useful for testing practical training and structure discovery methods.

In this report we process the data, exploring the variables relationship between the attributes and we model the data from different classification models, those are K nearest neighbor, naïve bayes and Decision trees in terms of their best set of parameter for each case and performance on car evaluation data set.

* 1. **Outcome**

From this car evaluation project we can know how will be the car acceptability, overall price, buying price ,maint price of the maintenance, technical characteristics. Anyone can know everything about a car configuration which is so much helpful for buying a car. They also can get knowledge about which car people want to buy. Anyone may learn about a car's general condition, which can assist them in purchasing a high-quality automobile. Before purchasing a car, this project might assist a person determine whether it is beneficial or bad for transportation. Our project could help people for business purposes or buying a car for personal use.

1. **Data Set** 
   1. **Real Data Set:**



* 1. **Source:**

Site : UCI <http://archive.ics.uci.edu/ml/datasets/Car+Evaluation>

**2.3 Description:**  The model evaluates cars according to the following concept structure:

CAR car acceptability PRICE overall price buying buying price maint price of the maintenance

TECH technical characteristics COMFORT comfort doors number of doors persons capacity in terms of persons to carry lug\_boot the size of luggage boot safety estimated safety of the car

Input attributes are printed in lowercase. Besides the target concept (CAR), the model includes three intermediate concepts: PRICE, TECH, COMFORT. Every concept is in the original model related to its lower-level descendants by a set of examples. The Car Evaluation Database contains examples with the structural information removed, i.e., directly relates CAR to the six input attributes: buying, maint, doors, persons, lug\_boot, safety. Because of known underlying concept structure, this database may be particularly useful for testing constructive induction and structure discovery methods.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Attribute** | **DataType** | **Missing values** | **Categories** | **Target** | **Entropy** |
|  | buying | nominal | 0 | 4 |  | 1.39 |
|  | maint | nominal | 0 | 4 |  | 1.39 |
|  | doors | nominal | 0 | 4 |  | 1.39 |
|  | persons | nominal | 0 | 3 |  | 1.1 |
|  | lug\_boot | nominal | 0 | 3 |  | 1.1 |
|  | safety | nominal | 0 | 3 |  | 1.1 |

1. **Number of Instances:** 1728
2. **Number of Attributes:** 7
3. **Missing Attribute Values**: None

**4.Class Values:** unacc, acc, good, vgood

**5. Attributes:**

buying: vhigh, high, med, low. maint: vhigh, high, med, low.

doors: 2, 3, 4, 5more. persons: 2, 4, more.

lug\_boot: small, med, big. safety: low, med, high.

**Model Development :**

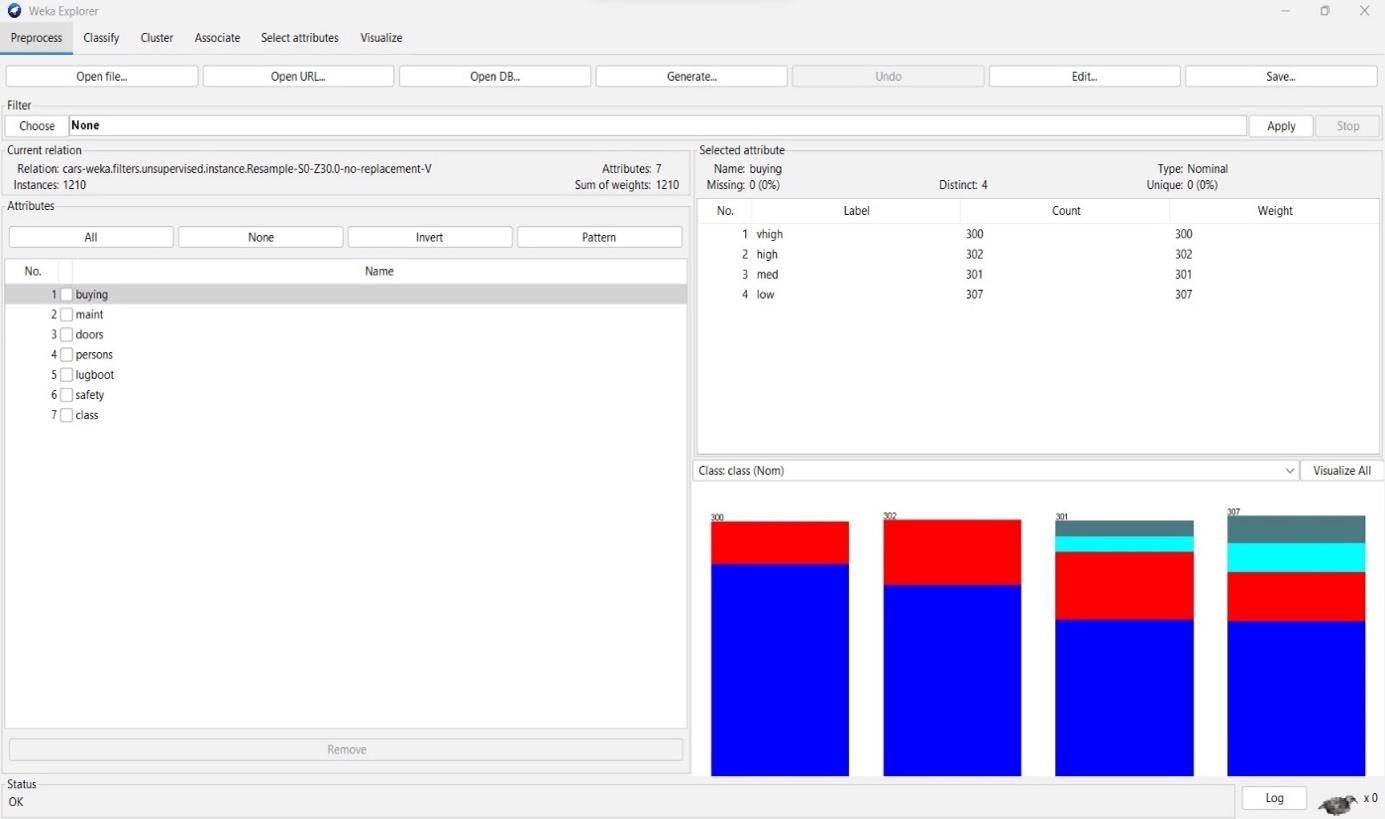


Fig: Attributes of the Dataset.

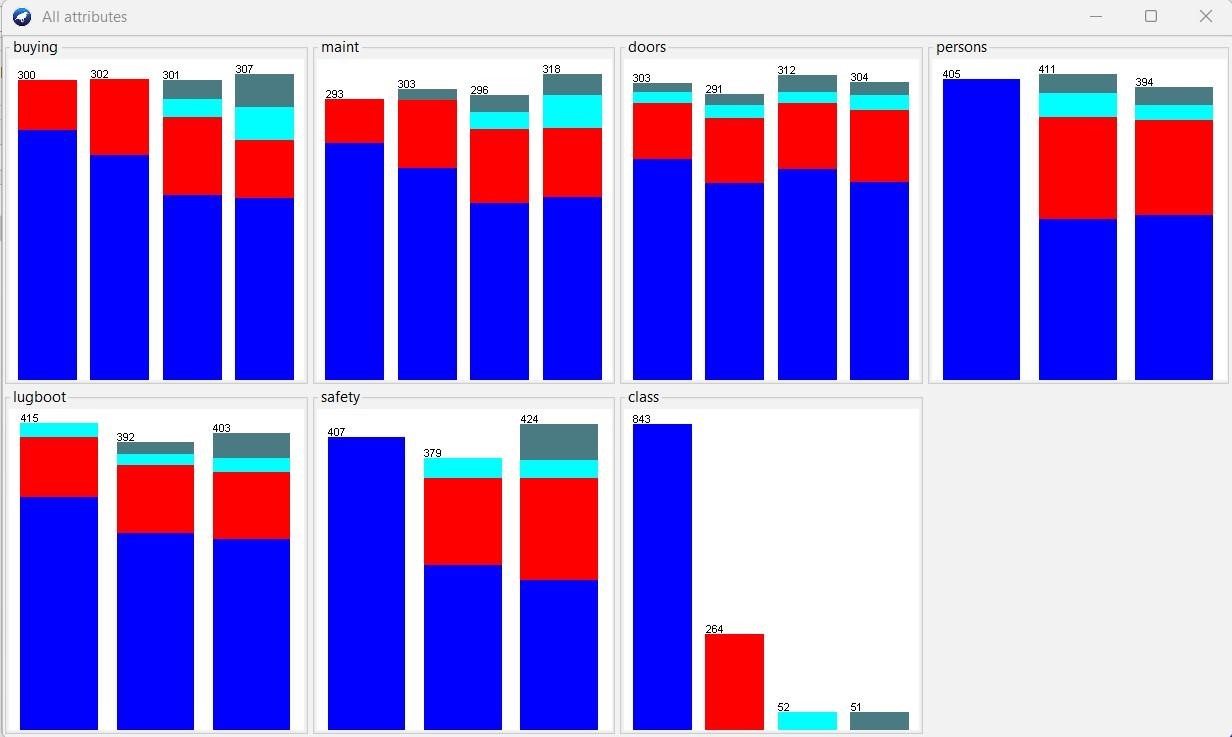


Fig 2 : Graph.

**K-Nearest Neighbour :**

The k-nearest neighbors (KNN) algorithm is a simple, supervised learning algorithm that can be used to solve both classification and regression problems.

**Advantages:**

* The algorithm is simple and easy to implement.
* There’s no need to build a model, tune several parameters, or make additional assumptions.

**Drawbacks:**

* High complexity for large dataset
* Need features scaling
* Sensitive to missing data

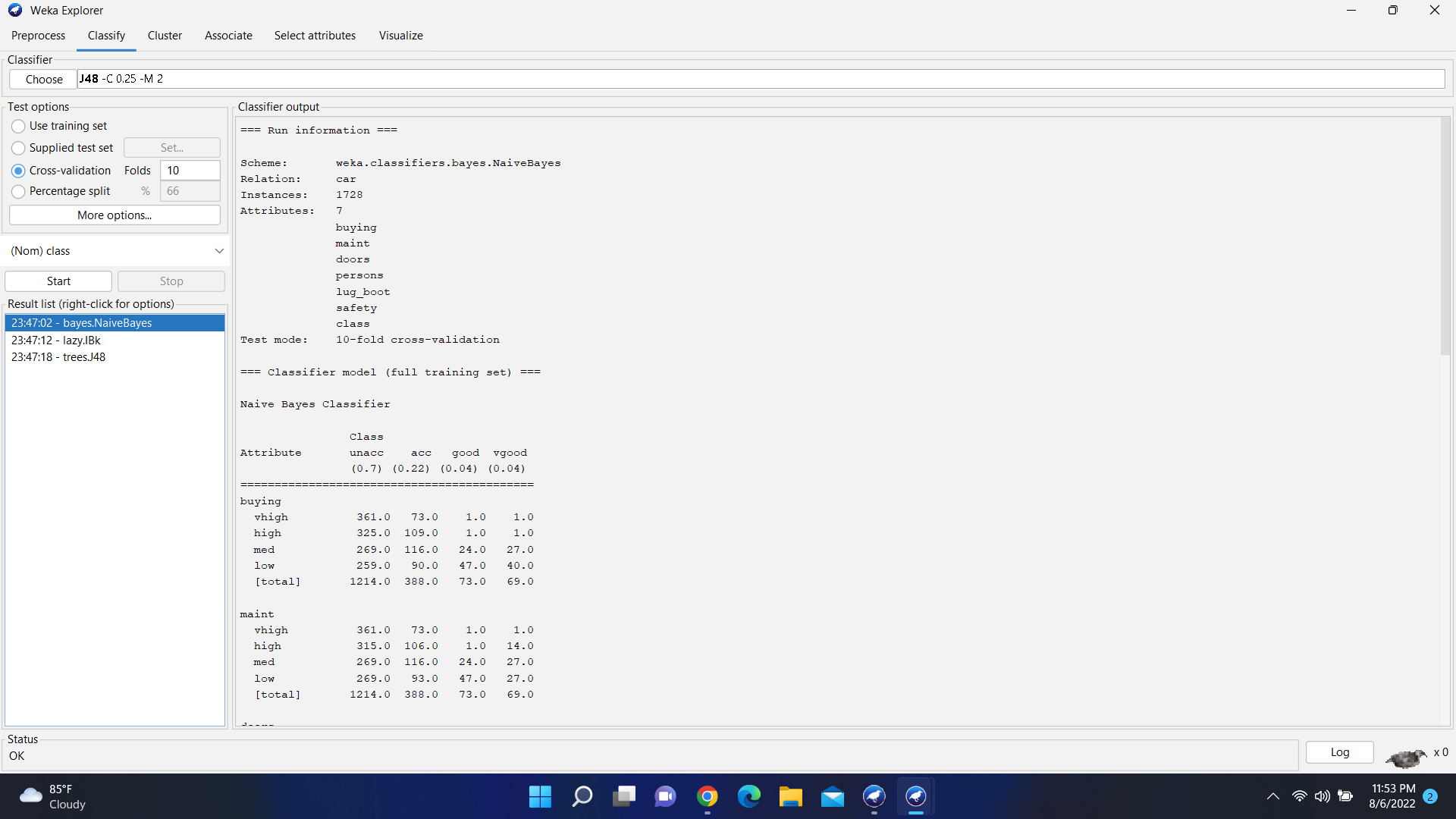
****

Fig 3 : Calculation of KNN.

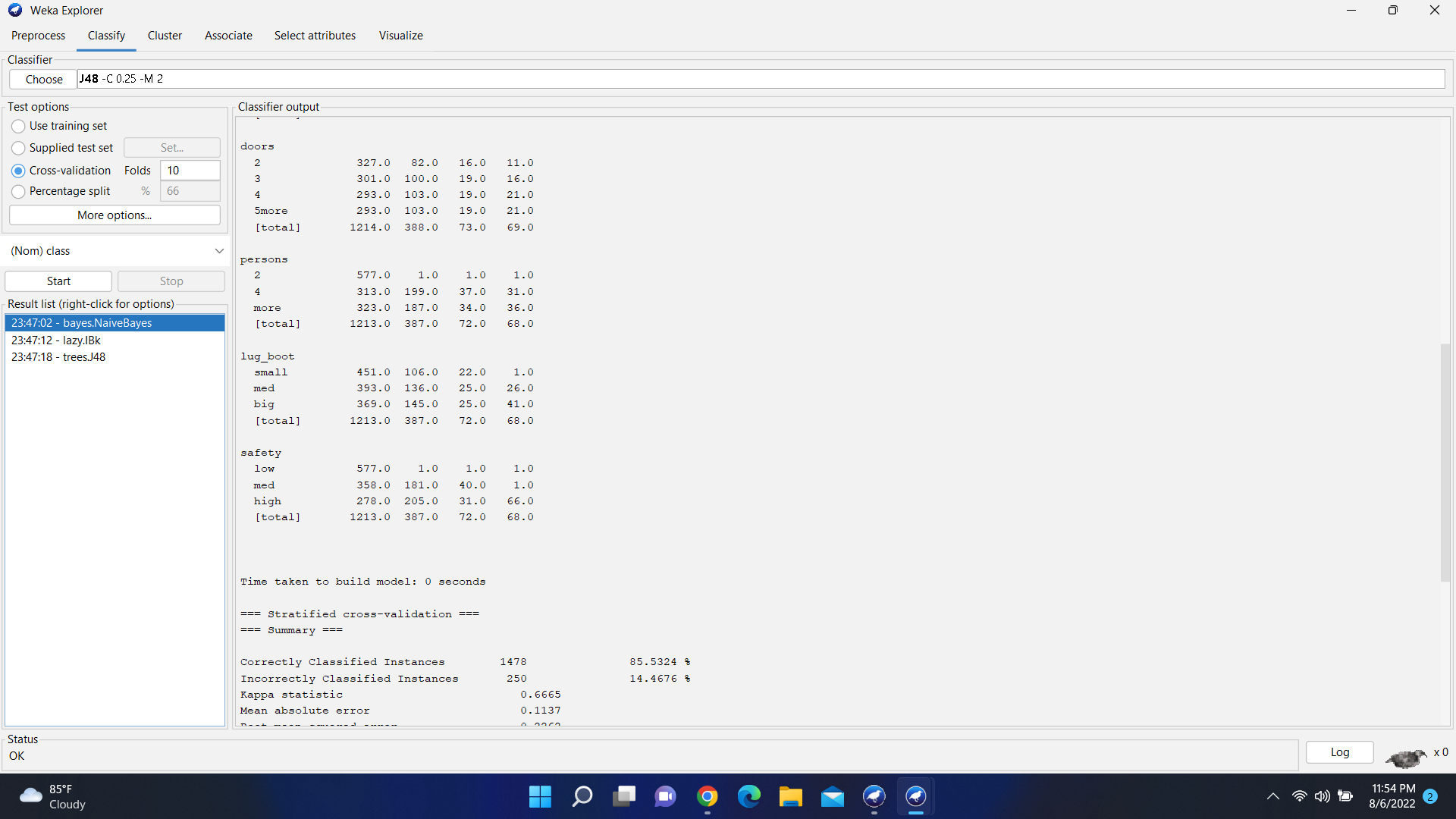


Fig-4: Calculation of KNN.

**Naive Bayes:**

Naïve Bayes is a method of classification that does not use rules, a decision tree or any other explicit representation of the classifier. It is a classification technique based on Bayes Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

**Advantages:**

* This algorithm is easy to implement.
* Find good results in most cases.
* It doesn’t require as much training data • It handles both continuous and discrete data
* This algorithm work quickly.
* Suitable for multiple prediction problem.

**Drawbacks:**

* All attributes being categorical.
* Assume that all veritable are independent.
* This algorithm faces zero frequency problem
* Estimation can be wrong in some cases.

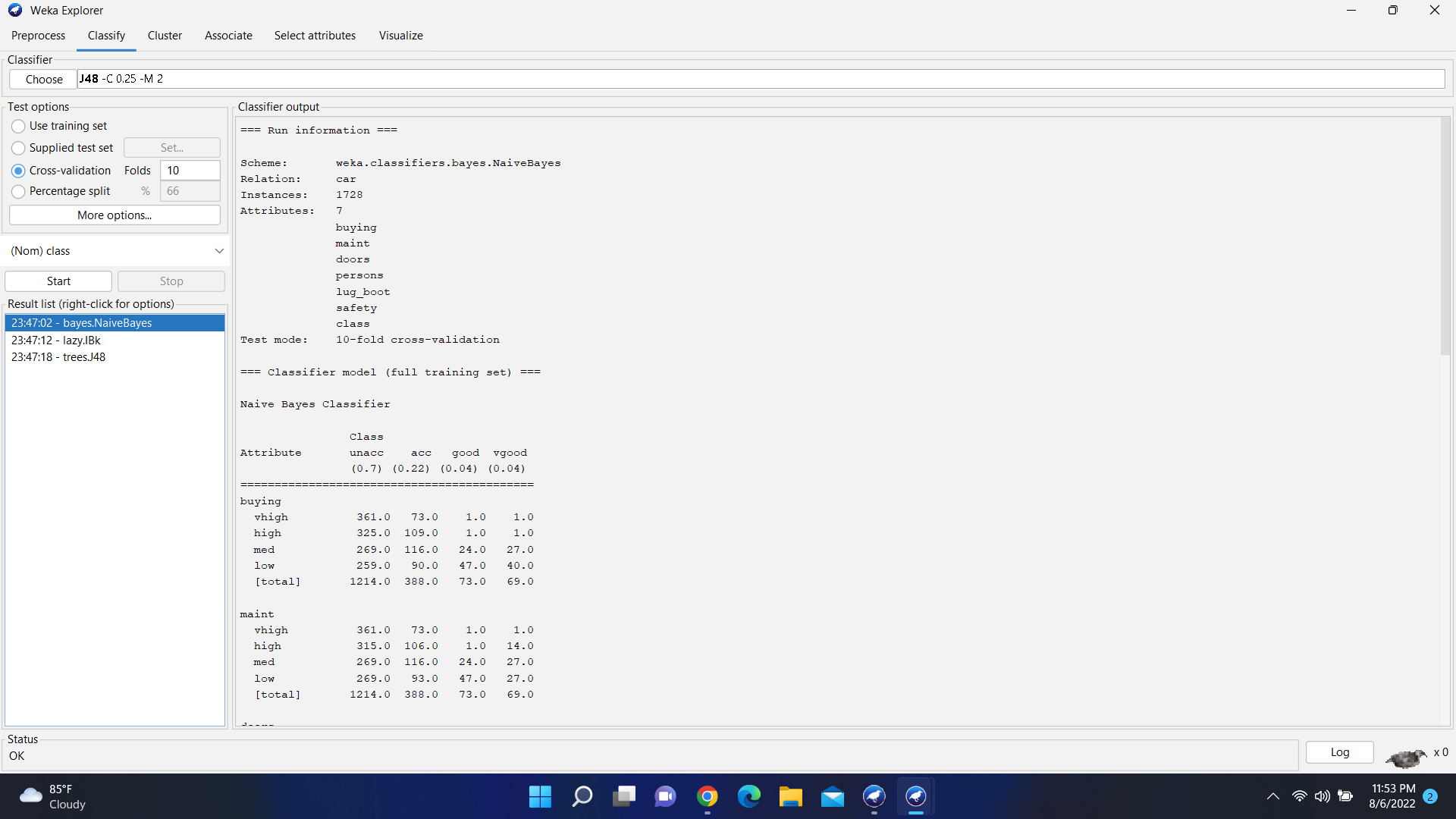


Fig: Calculation of Naïve bayes.

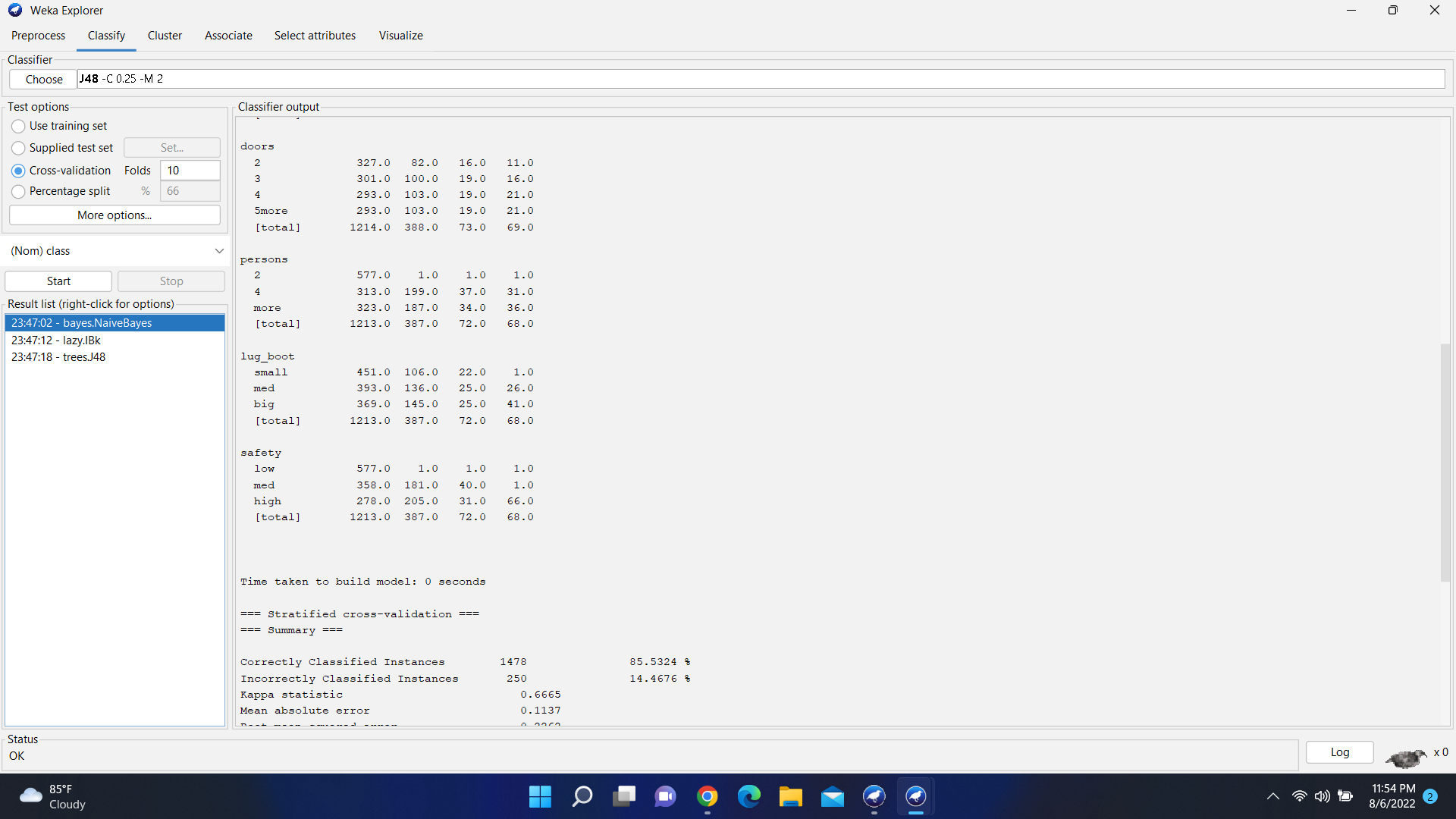


Fig: Calculation of Naïve bayes.

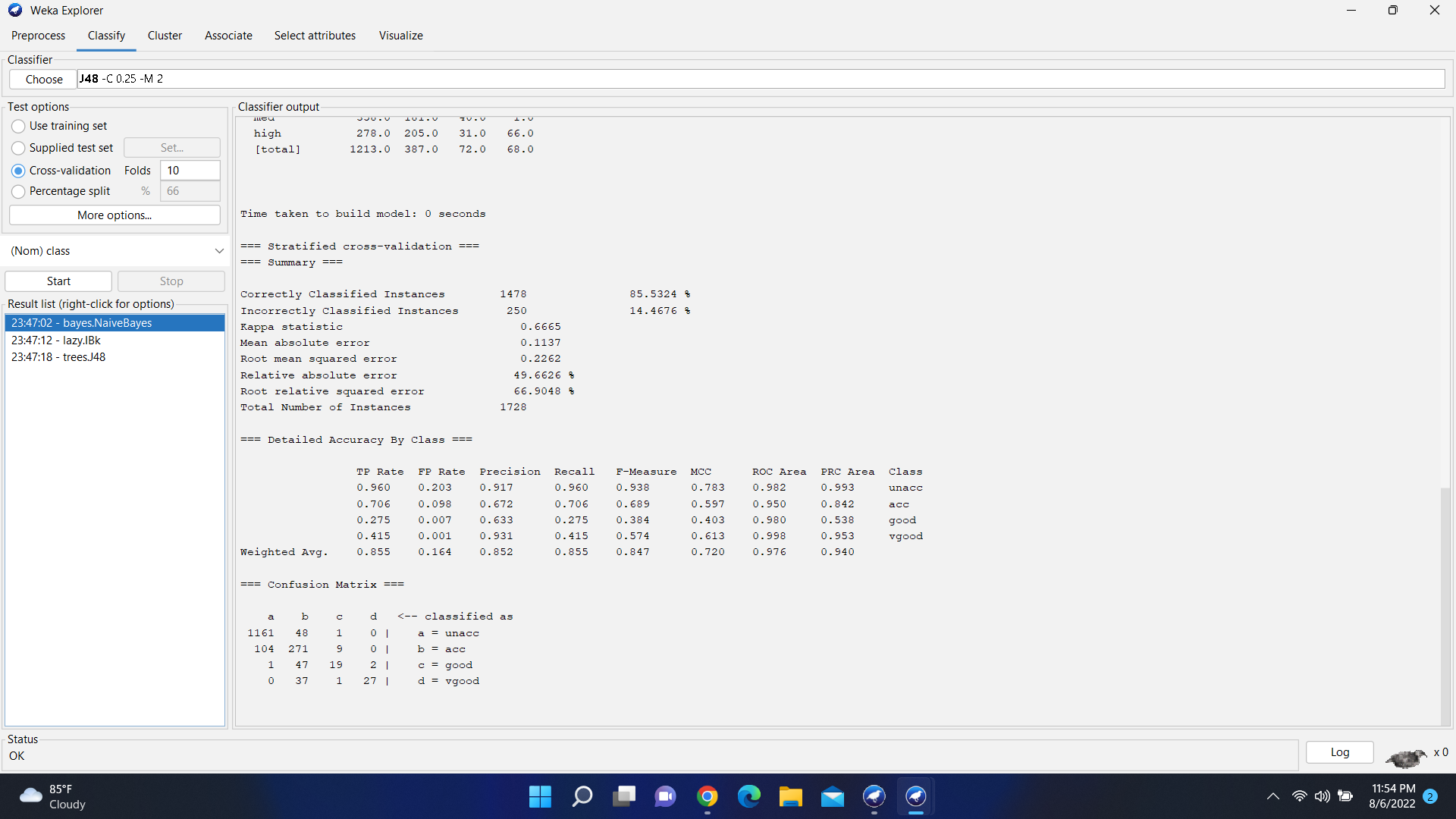


Fig: Calculation of Naïve bayes.

**Decision tree :**

Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving regression and classification problems too.

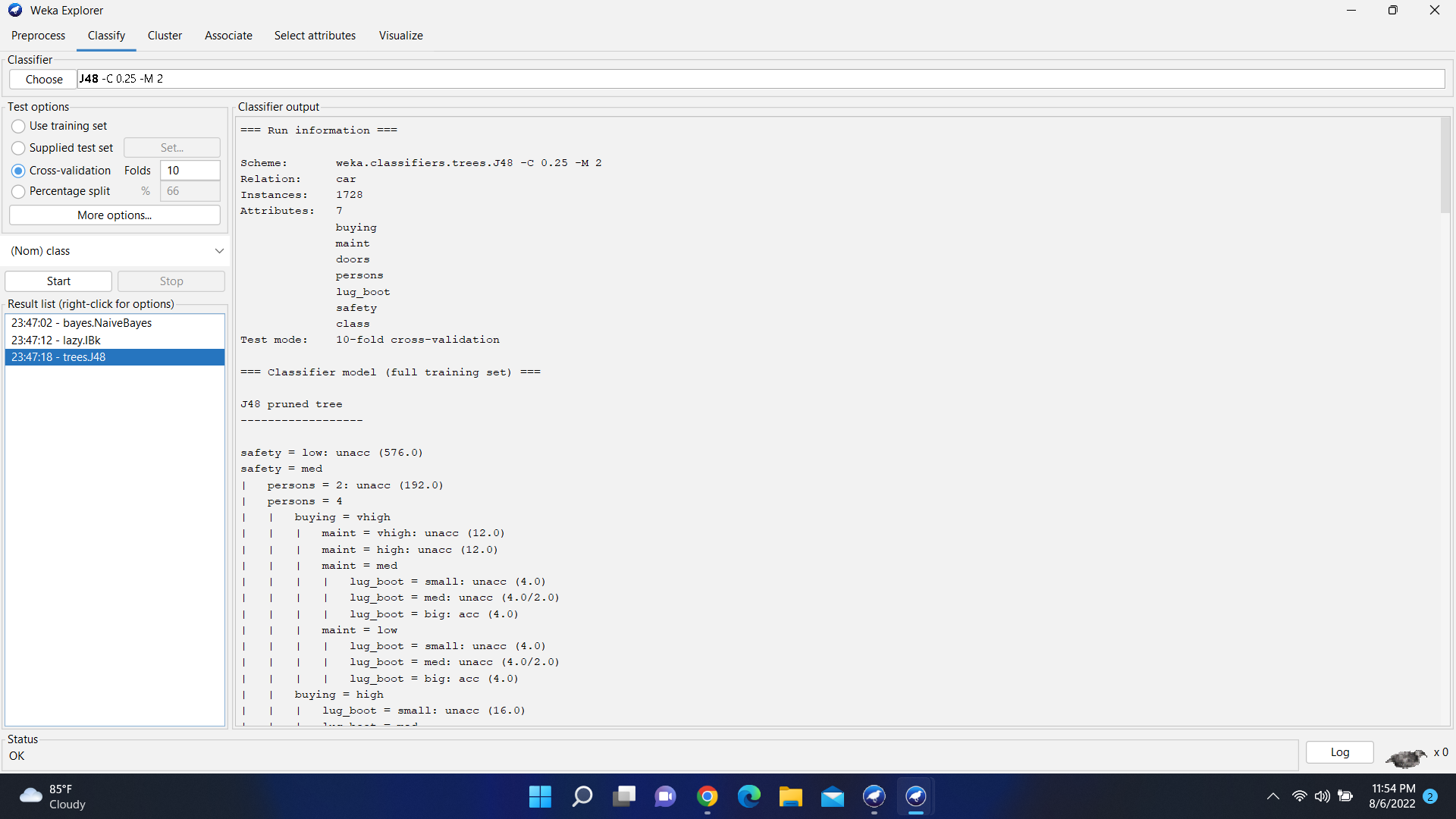
The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data(training data).

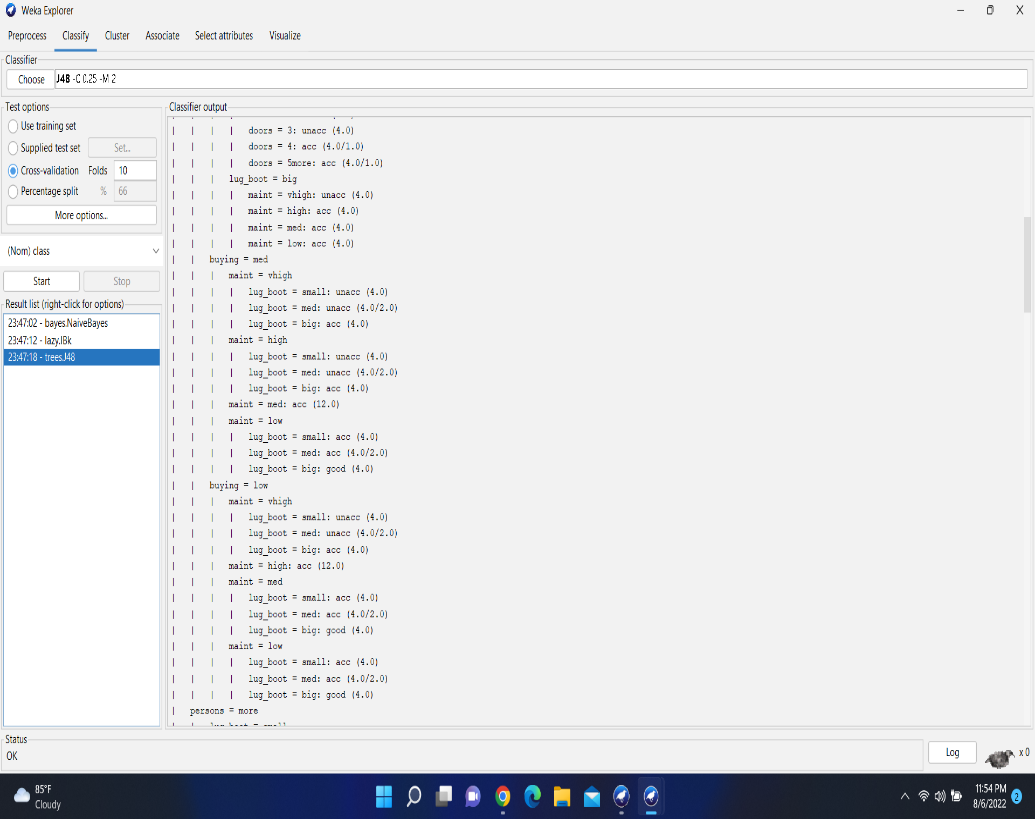
**Advantages:**

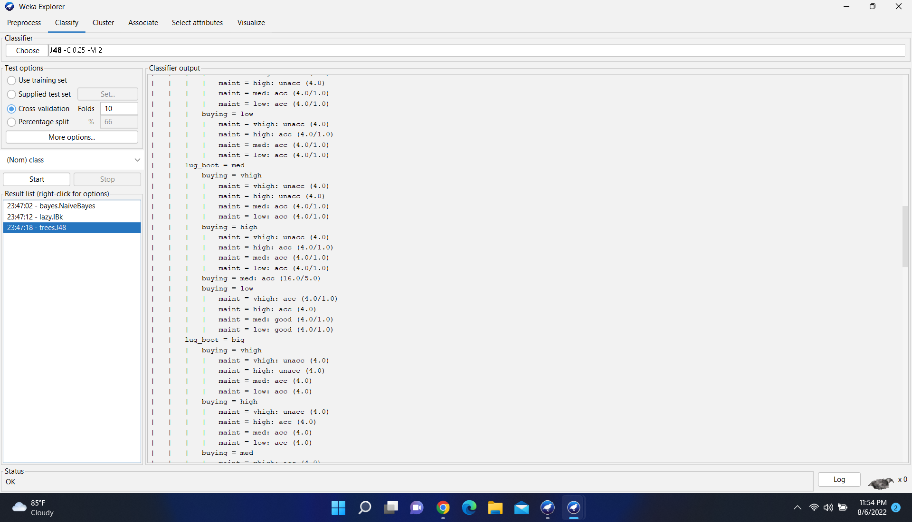
* Compared to other algorithms decision trees requires less effort for data preparation during pre-processing.
* A decision tree does not require normalization of data.
* A decision tree does not require scaling of data as well.
* A Decision tree model is very intuitive and easy to explain to technical teams as well as stakeholders.

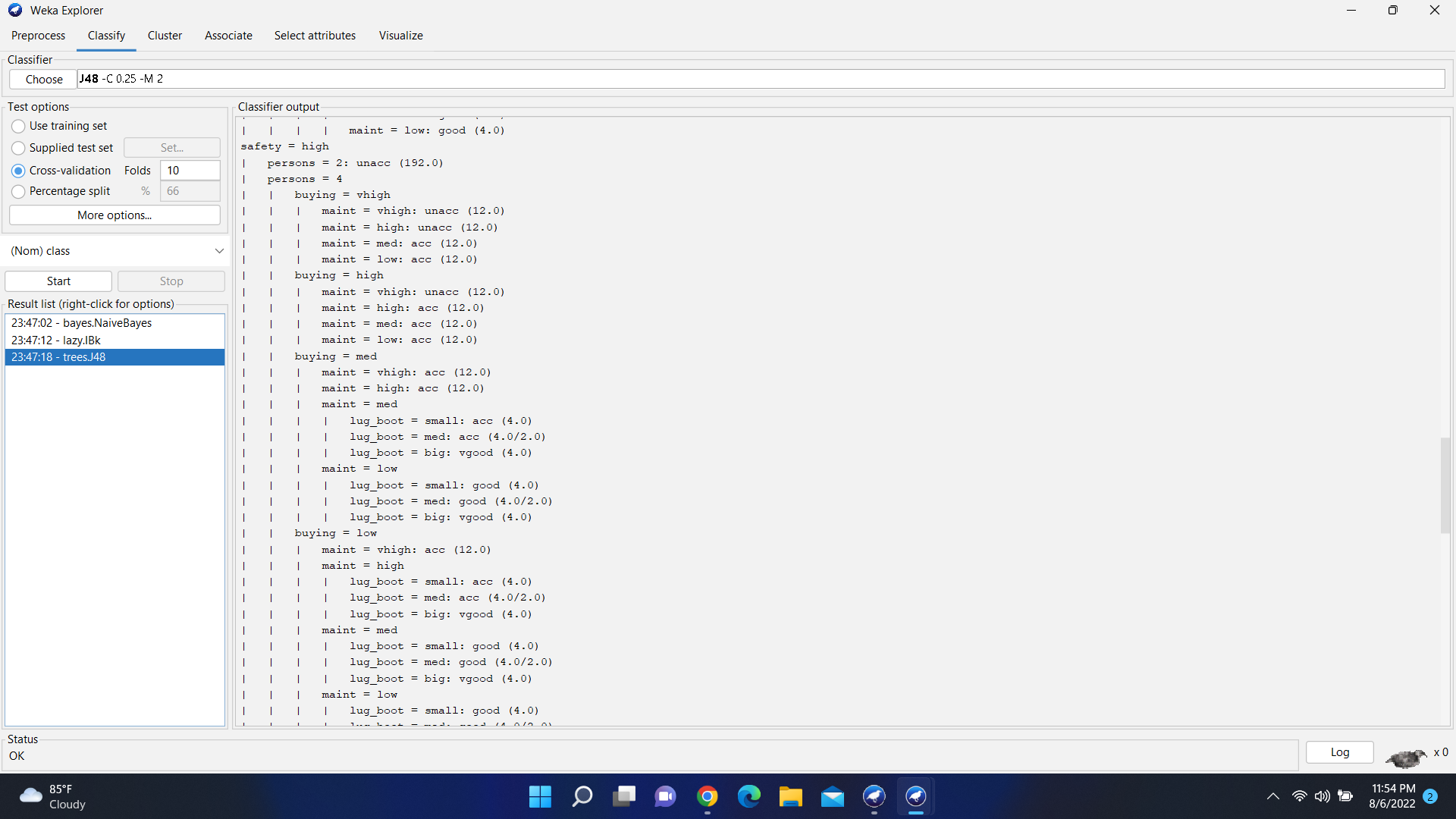
**Drawbacks:**

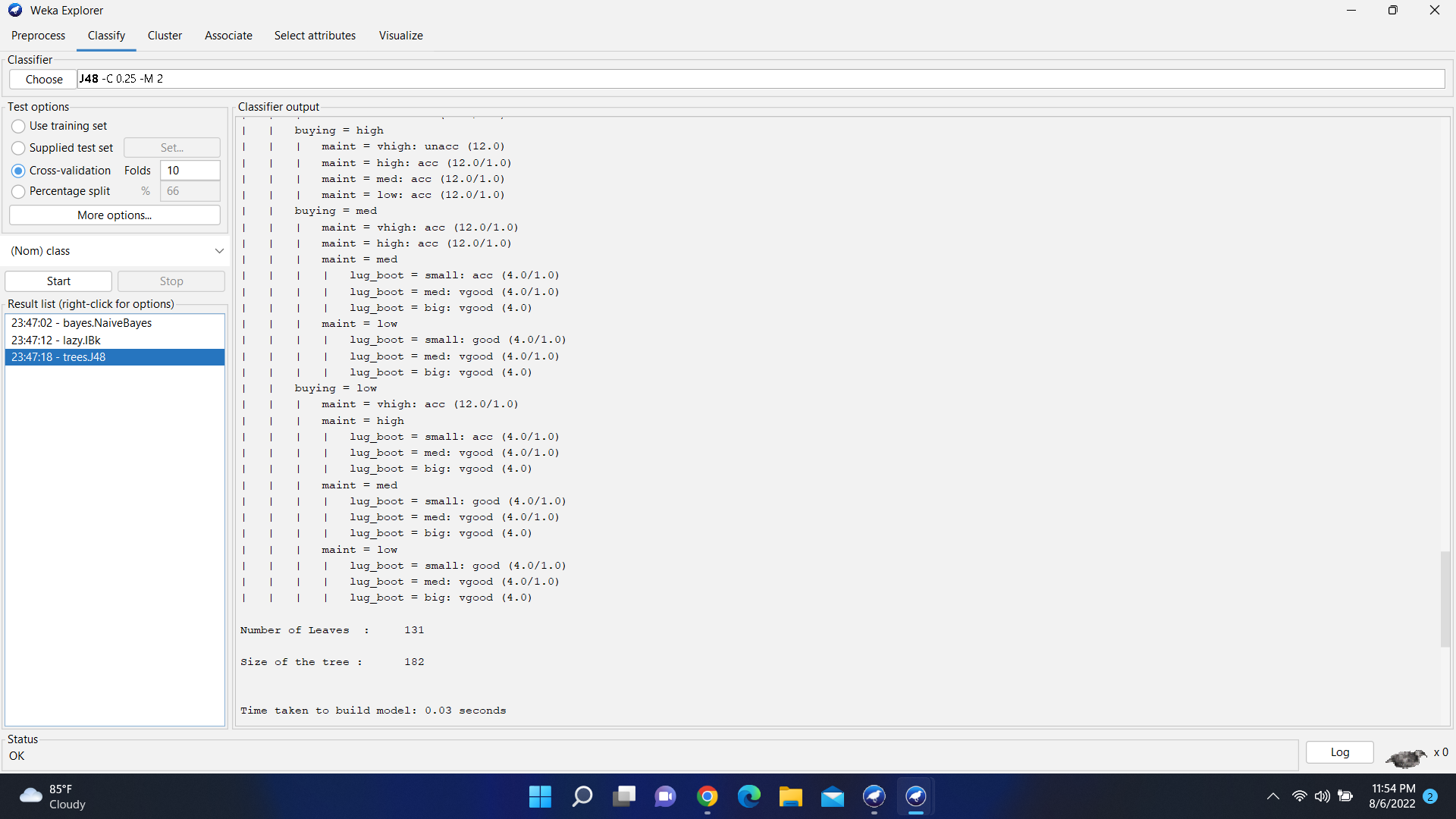
* A small change in the data can cause a large change in the structure of the decision tree causing instability.
* For a Decision tree sometimes calculation can go far more complex compared to other algorithms.
* Decision tree often involves higher time to train the model.
* Decision tree training is relatively expensive as the complexity and time has taken are more.



Fig: Calculation of decision tree.

Fig: Calculation of decision tree

.



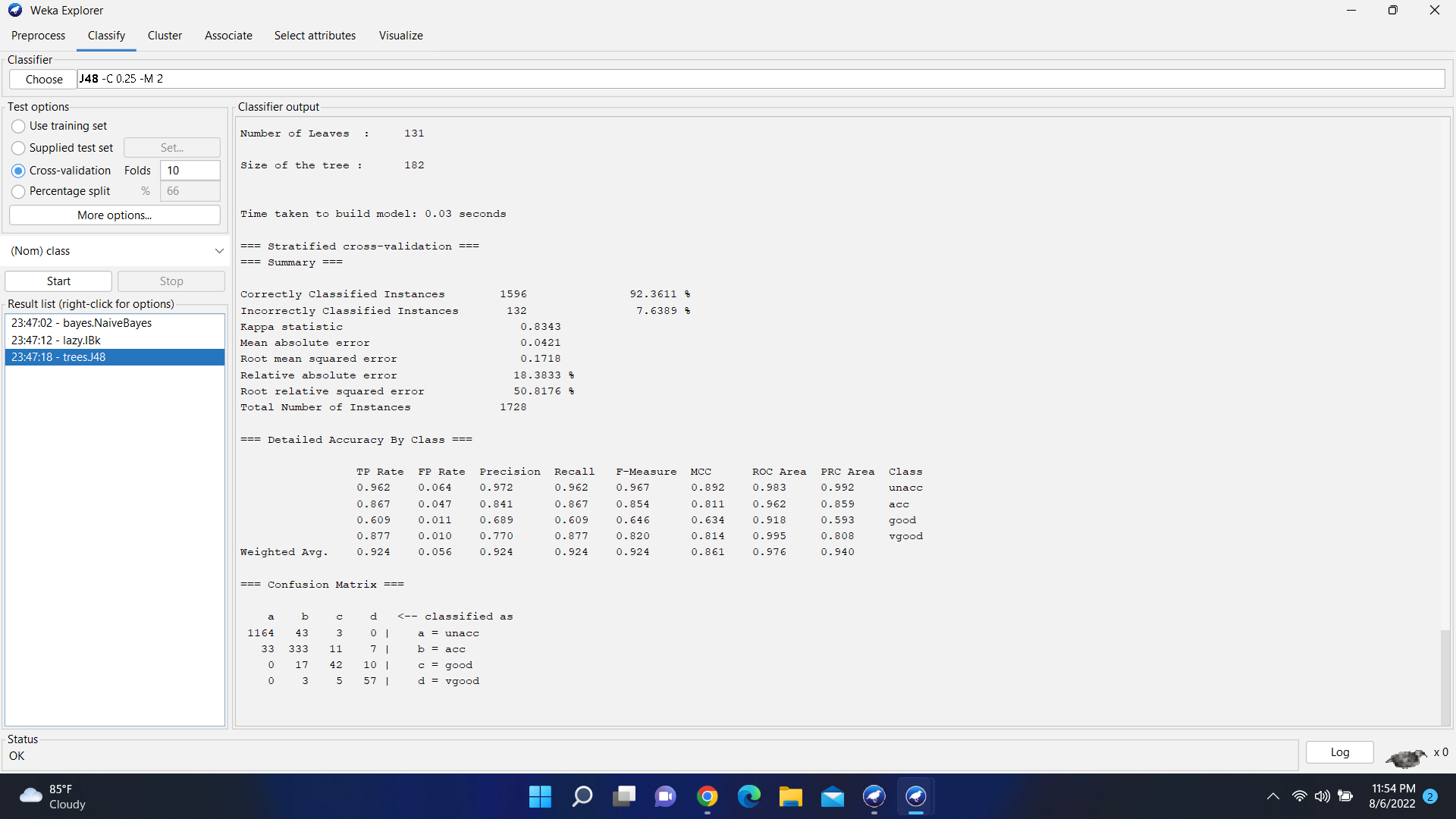


Fig: Calculation of decision tree.

**Table of Accuracy:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | KNN | Naïve bayes | Decision Tree |
| Accuracy | 90.3306% | 84.6281% | 89.5377% |
| Testing accuracy | 9.6694% | 15.3719% | 10.4623% |
| Relative absolute  error | 49.4805% | 50.5123% | 24.7675% |
| Root relative squared error | 60.1505% | 67.5251% | 60.3555% |
| Confusion matrix | [830 13 0 0 ]  [54 209 1 0 ]  [2 36 14 0 ]  [0 9 2 40] | [799 43 1 0 ]  [73 180 11 0 ]  [0 38 12 2 ]  [0 18 0 33] | [246 16 1 0]  [4 68 4 4]  [0 6 5 4]  [0 4 0 9] |

**Discussion & Conclusion:** In this project, different algorithms are applied on the car evaluation dataset.Here we can see that KNN has the highest accuracy than other algorithms and it is the most suitable module for our dataset. We able to get testing accuracy for all the algorithms. But KNN, testing accuracy is better than others. All the attributes plays a vital role for customers whether the car is accepted or unaccepted class. Safety and a person’s capacity are the main factors in rejecting car classes as unacceptable. No of doors plays no importance in deciding the classes of the car.