****

**Honours Degree in Computing**

**Data Analytics Assessment:**

**Analyse a dataset**

**Submitted by: Carl Brady, B00084475**

**Submission date:**

**17/12/2019**

**Declaration**

I declare that this material, which I now submit for assessment, is entirely my own work and has not been taken from the work of others, except where otherwise stated. I have identified and included the source of all facts, ideas, opinions, and viewpoints of others in the assignment references. Direct quotations from books, journal articles, internet sources, module text, or any other source whatsoever are acknowledged and the source cited are identified in the assignment references.

I understand that plagiarism, collusion, and copying are grave and serious offences and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying. I acknowledge that copying someone else’s assignment, or part of it, is wrong, and that submitting identical work to others constitutes a form of plagiarism. I have read and understood the colleges plagiarism policy 3AS08 (available [here](https://www.itb.ie/AboutITB/QualityAssurancePolicies/3AS08%20Institute%20Procedural%20Guidelines%20for%20Dealing%20with%20Instances%20of%20Plagiarism%20in%20Assignments%20and%20Examinations%20%2013%20June%202014.pdf)).

This material, or any part of it, has not been previously submitted for assessment for an academic purpose at this or any other academic institution.

I have not allowed anyone to copy my work with the intention of passing it off as their own work.

Name: Carl Brady Dated: 08/12/2019

# 

Table of Contents

[Business Understanding 4](#_Toc27434587)

[Business objective 4](#_Toc27434588)

[Data Mining objective 4](#_Toc27434589)

[Data Understanding 4](#_Toc27434590)

[Data Exploration 5](#_Toc27434591)

[Data Quality Verification 7](#_Toc27434592)

[Data Preparation 9](#_Toc27434593)

[Selecting Data 9](#_Toc27434594)

[Cleaning Data 9](#_Toc27434595)

[Constructing Data 10](#_Toc27434596)

[Modelling 10](#_Toc27434597)

[Modelling techniques 10](#_Toc27434598)

[Generating Test Design 10](#_Toc27434599)

[Building and Assessing the model 11](#_Toc27434600)

[Evaluation 11](#_Toc27434601)

[Conclusion 12](#_Toc27434602)

[References 12](#_Toc27434603)

[Appendix 13](#_Toc27434604)

[Rapid Miner Process 13](#_Toc27434605)

[Default Model Results 13](#_Toc27434606)

[Gini Index Results 14](#_Toc27434607)

[Information Gain Results 14](#_Toc27434608)

[Gain Ratio Results 15](#_Toc27434609)

[Naïve Bayes Results 15](#_Toc27434610)

[Decision Tree 16](#_Toc27434611)

# Business Understanding

The creation of this model is based upon the heart disease dataset. This dataset contains seventeen attributes and one class label. There are general attributes such as id, age, gender and specific attributes such as carotid thickness, chest pain type, serum cholesterol. Cardio-Vascular disease or Heart disease accounts for thirty-six percent of all deaths in Ireland [1], therefore it is important to create such a prediction model to attempt to prevent the disease.

## Business objective

* The business objective is to predict whether a person is susceptible to heart disease.

## Data Mining objective

* Build a classification model to predict whether a person may or may not be susceptible to heart disease based upon the data provided about the individual.

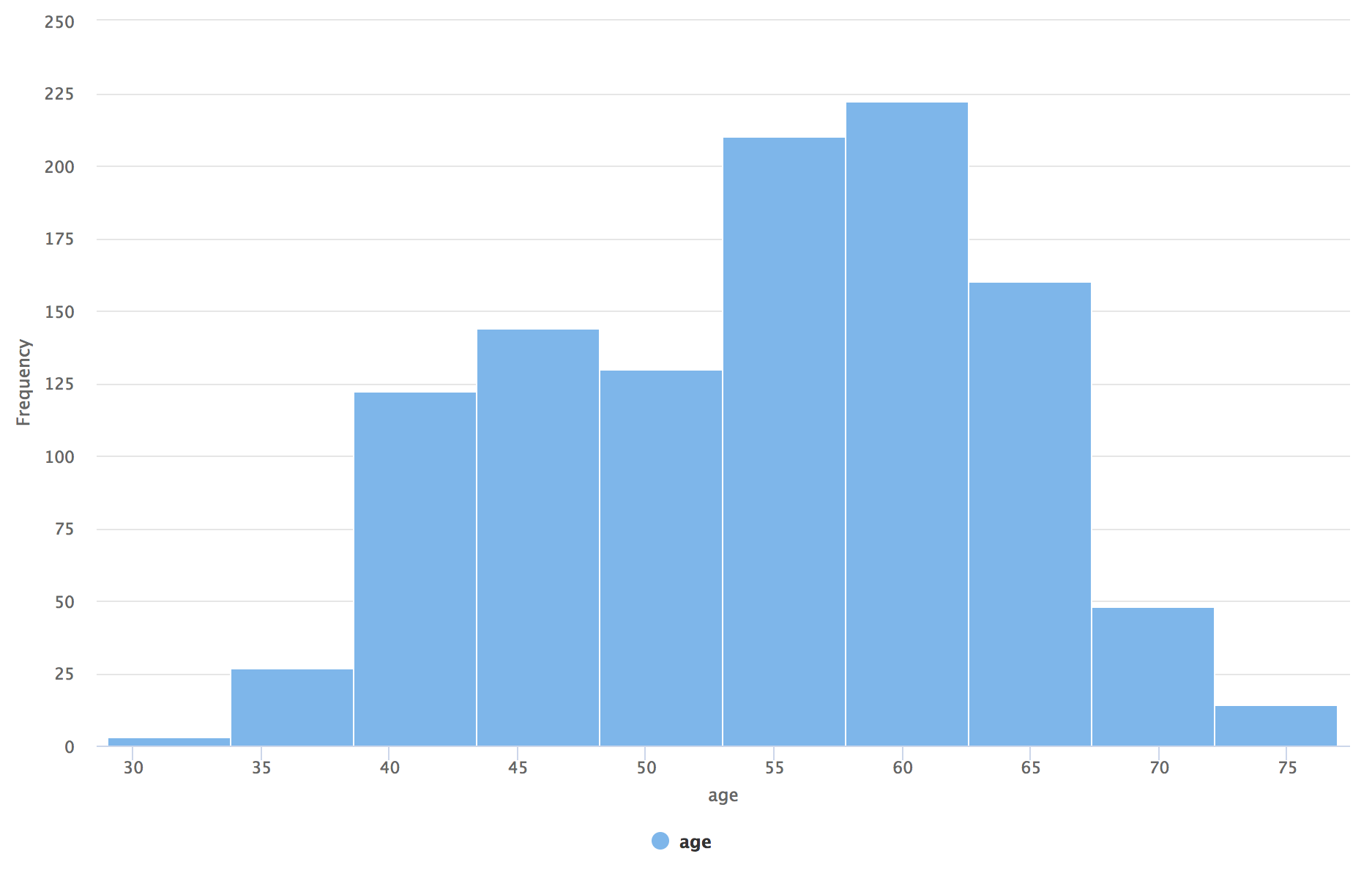
# Data Understanding

The data found in the dataset was placed into a table, each attribute was then described based on their data type, mean, min, max and standard deviation. If the attribute was categorical or ordinal, these values alongside a brief explanation were also inputted into the table.

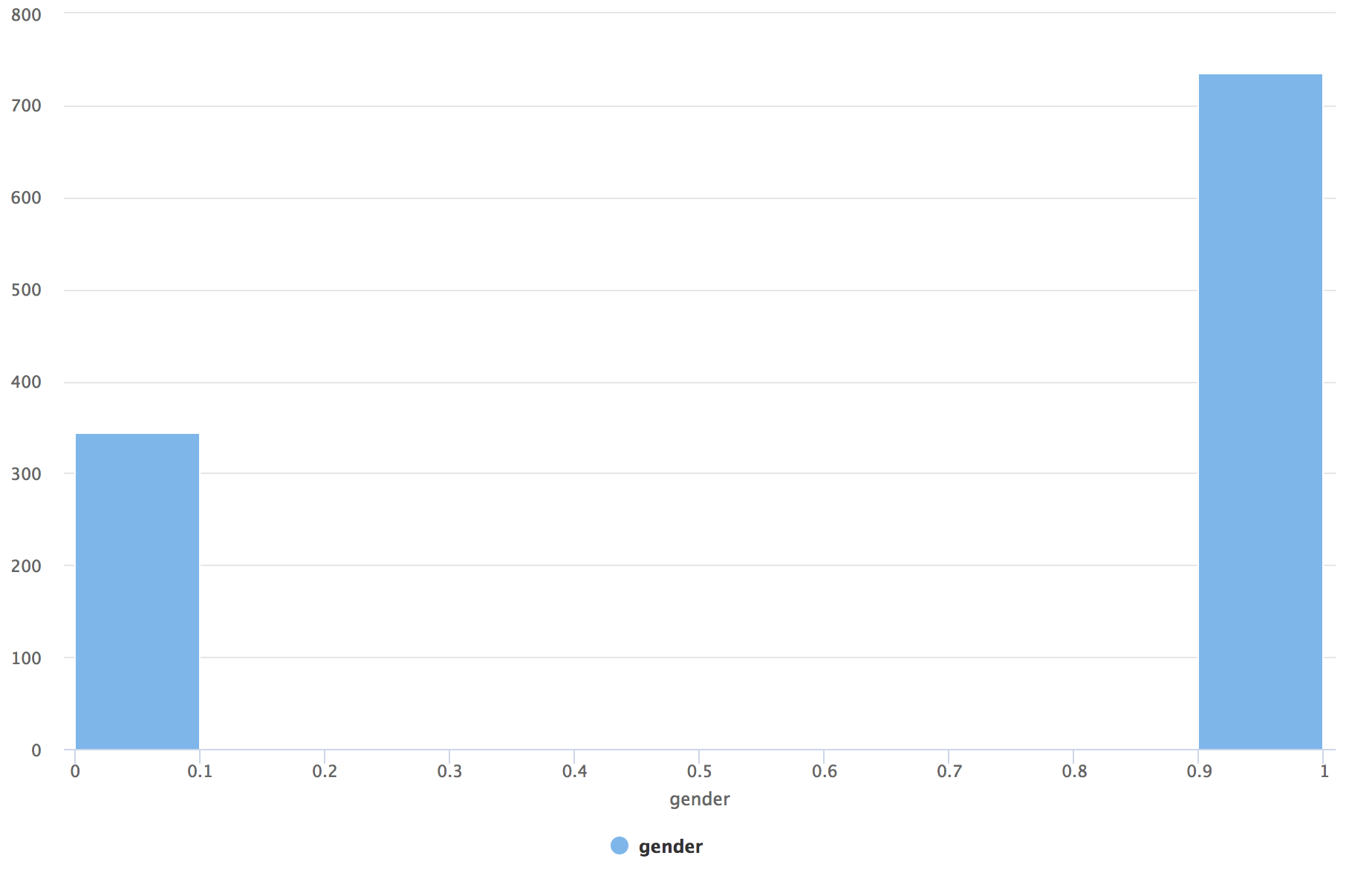
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Description | Data Type | Mean | Min | Max | St Dev | Categorical/Ordinal Values |
| Age | Ratio | 54.543 | 29 | 77 | 9.069 |  |
| Gender | Categorical | 0.681 | 0 | 1 | 0.466 | 0 & 1, represent Male & Female. |
| Chest Pain Type | Ordinal | 3.169 | 1 | 4 | 3.169 | 1,2,3,4 represent level of chest pain severity. |
| Resting Blood Pressure | Ratio | 259.180 | 94 | 139585 | 4241.547 |  |
| Serum Cholesterol | Ratio | 496.918 | 126 | 266895 | 8110.165 |  |
| Blood Sugar | Binary  Categorical | 0.148 | 0 | 1 | 0.355 | 0 & 1 represent Blood Sugar values. |
| Electro Cardiograph | Categorical | 1.031 | 0 | 2 | 0.995 | 0 & 2 represent Electro Cardiograph values. |
| Max Heart  Rate | Ratio | 149.540 | 71 | 202 | 22.550 |  |
| Angina | Binary  Categorical | 0.318 | 0 | 1 | 0.466 | 0 & 1 represent whether person suffers from angina. |
| Peak ST | Ratio | 2.022 | 0 | 1086.200 | 33.024 |  |
| Slope ST | Ordinal | 4.703 | 1 | 1691 | 72.603 |  |
| Vessels Colour | Categorical | 2.693 | 0 | 729 | 38.204 | 0,1,2,3 represent vessel colour ( 729 is an error value) |
| Thal | Ordinal | 4.738 | 3 | 7 | 1.950 | 3,6,7 represent Thal values. |
| Lipid | Ratio | 4.998 | 0.722 | 1080 | 32.742 |  |
| LDL | Ratio | 252.846 | 0 | 2264 | 344.193 |  |
| Carotid  Thickness | Ratio | 1.095 | 0.00 | 1080 | 32.846 |  |
| ID | Nominal | N/A | 1 | 1084 | N/A | ID listing |

## Data Exploration

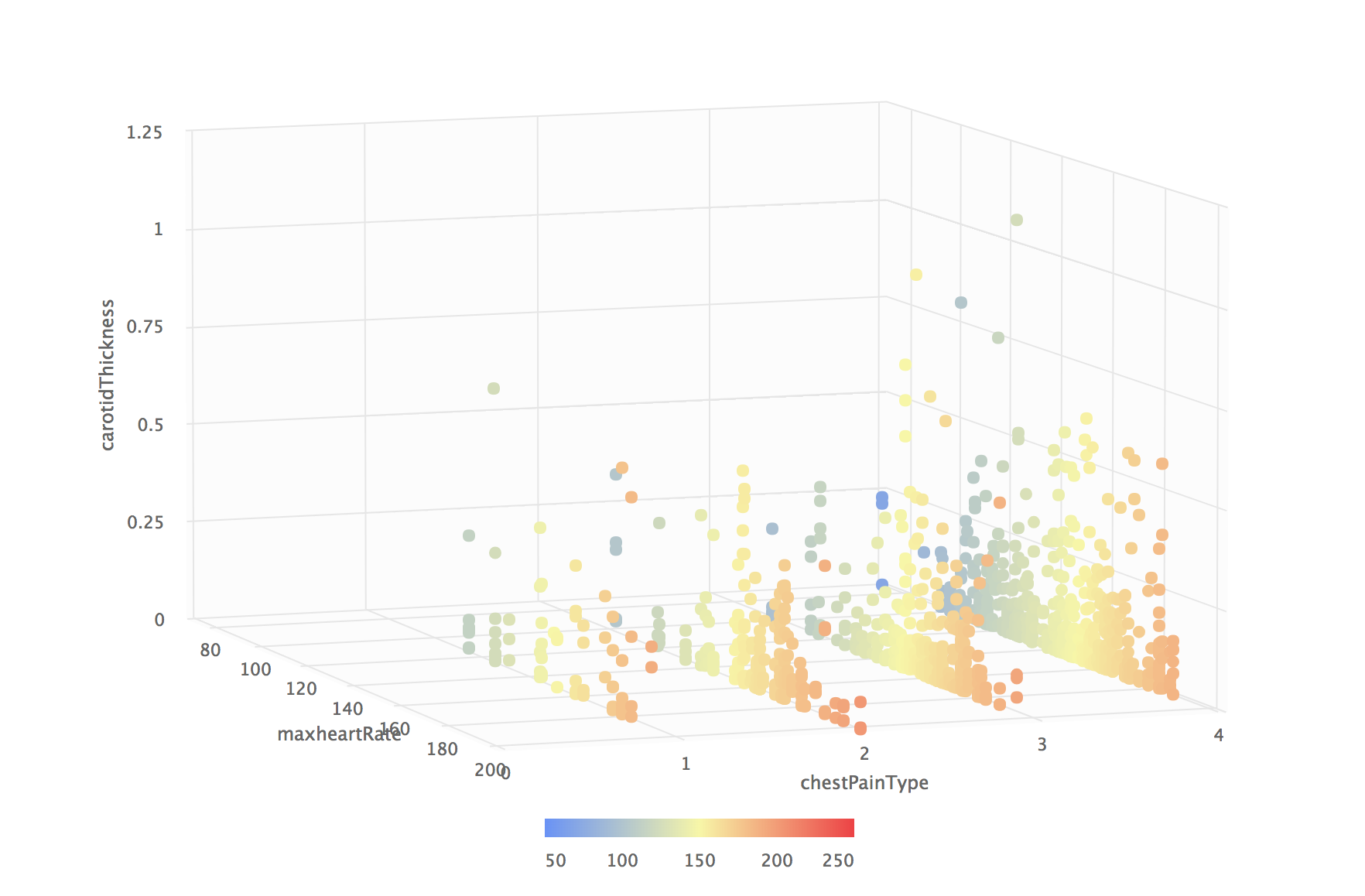
Upon exploration of the data, it was first noted that the age attribute was slightly biased towards people older than 52. Upon examining a histogram of the data from this attribute, it is possible to see this is correct and there is a slightly skewed distribution towards people above 52 years old.



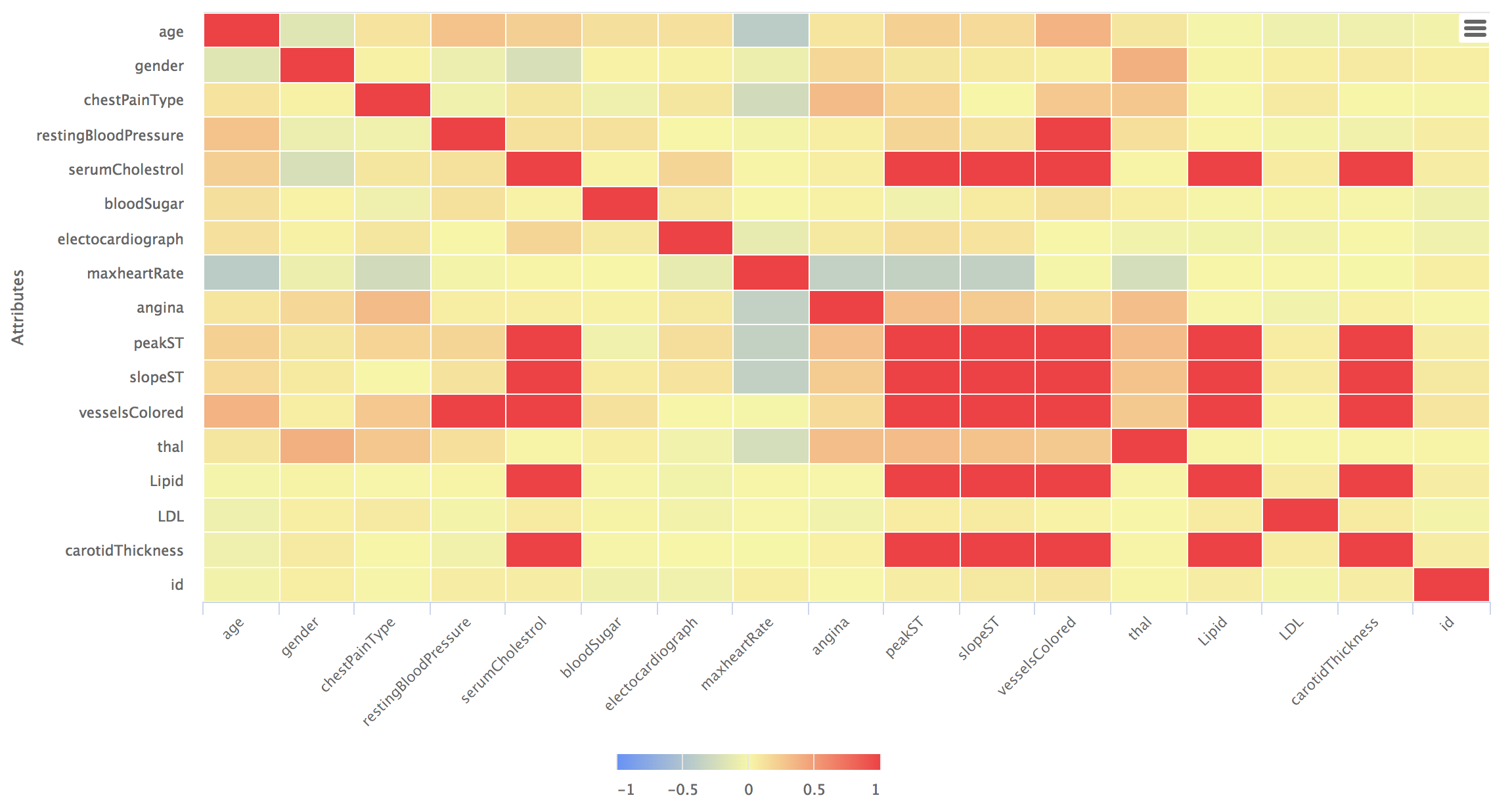
Also noted is that there is twice the amount of gender one as there is compared to gender zero. Gender zero accounted for 345 of the rows that appear in the dataset in comparison to 725 for gender one.



By grouping attributes together to establish if the attributes may be predictive or correlated, it was found that there is there predictive relationship between the chest pain severity and carotid thickness and max heart rate, the higher the level of chest pain, the higher the carotid thickness.

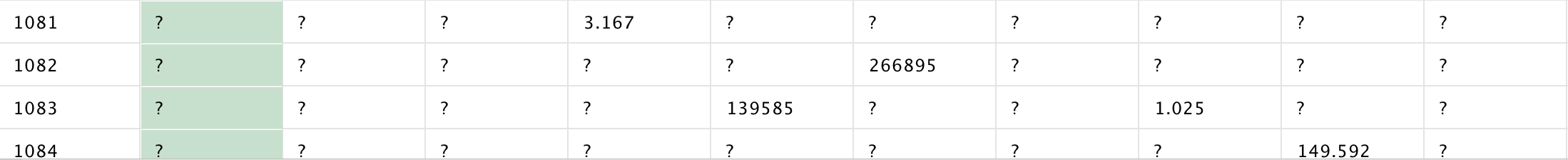


A correlation matrix was used to determine if any attributes were positively or negatively correlated.

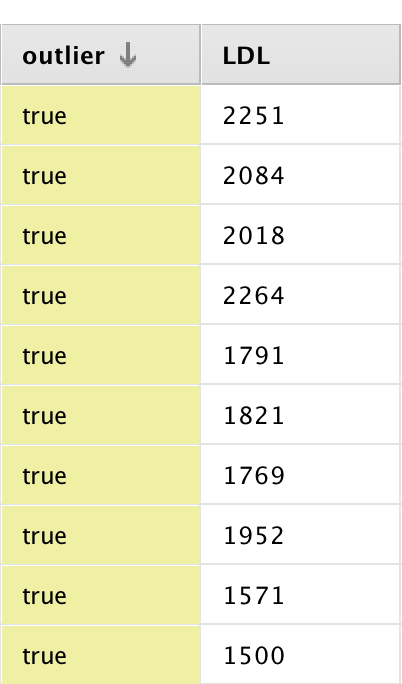


## Data Quality Verification

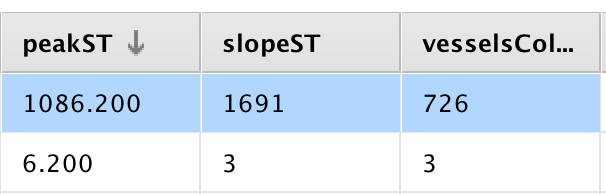
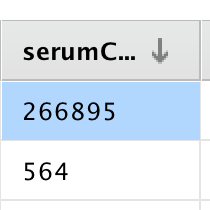
Upon data exploration, it was found that the last four rows contained numerous missing values. The missing values found in all columns except for CRP and LDL were present in these four rows. It was found that CRP had over 40% missing values and LDL had 194 missing values (roughly 17%).



Outlier detection was used to detect outliers, ten outliers were selected to be outputted to examine potential outliers in the dataset. Upon completion of this process, it was discovered that the ten outliers found from the operator all were due to the LDL values being very high in comparison to the average value (252.84), this finding may not be due to incorrect data but rather due to extremities in the data.



Manual outlier detection was explored as the results of the rapid miner operator did not seem to detect some outliers elsewhere. Interestingly, outlier detection did not seem to detect the outliers that occur in ‘resting blood pressure’, ‘serumCholesterol’, ‘peakST’, ‘slopeST’ and ‘vesselsColour’ immediately, which all have obvious outliers. It is noteworthy that all the mentioned outliers appear in rows 1081 – 1084, which have already been noted to having many missing values, this indicates the outliers are most likely input errors.



# Data Preparation

To create a baseline accuracy, the cross-validation operator was used. Within the cross-validation operator, the model was applied and a performance operator was used to create the baseline accuracy figure. The result of this initial test can be seen below.

## Selecting Data

To remove useless attributes, the remove useless attributes operator was used. The numerical min deviation was set to 0.3. The nominal useless above was set to 0.9, meaning that if the most frequent value occurs for more than 90% of the given attribute, then we can assume that this attribute is not predictive to the business objective. We also set the nominal useless below to 20% as if the most frequent value is less than 20% of an attribute it is also not very predicative. The results from this run removed gender, blood sugar and angina. The ‘id’ attribute was then manually removed as this attribute has no benefit to our predictive model.

## Cleaning Data

Rows 1081 – 1084 were removed due to having so many missing values, this was accomplished by using filter example operator. The CRP column was removed due to having over 40% missing values, as this value is so high it makes sense to remove the column rather than replace the missing values. The only missing values that were present in the dataset at this stage was the LDL.

The missing values in LDL were replaced with the average value using the ‘replace missing values’ function. Upon inspection of the dataset, it could be seen there were no more missing values at this stage and the average and standard deviation were no longer effected by incorrect outliers found in rows 1081 – 1084. LDL values were also replaced using imputation, this was accomplished by using the ‘impute missing values’ operator. Inside this operator, the KNN algorithm was used to determine the value to replace each missing value with. Both implementations will be used in the modelling stage.

## Constructing Data

As some attributes had a large range of values, the ‘normalise’ operator was used to condense the attributes into a smaller range. The attributes that were normalised were age, LDL, resting blood pressure and serum cholesterol.

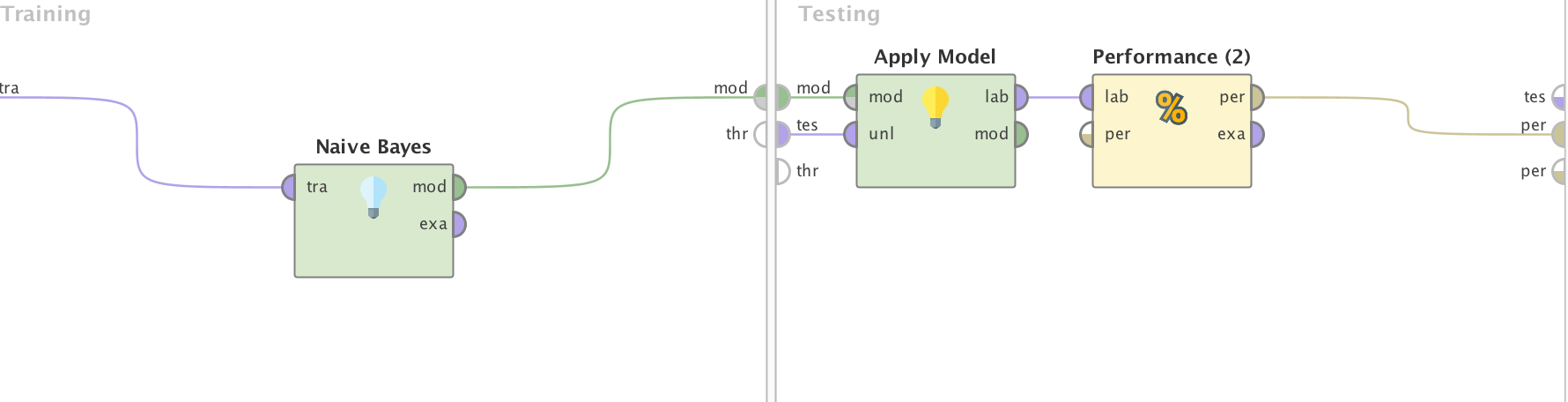
# Modelling

Modelling techniques

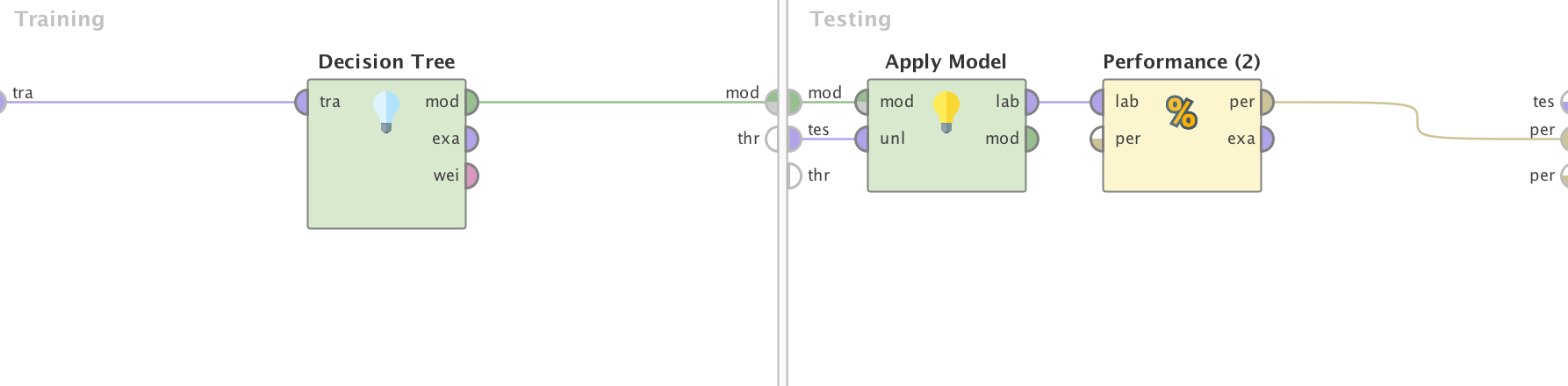
A decision tree was used as the data mining technique as it is easy to understand. It also gives great visualisation into how the algorithm works and what decisions were deemed important in predicating the business objective. Naïve Bayes was also implemented due to its use for probability base scenarios, which in the case of the business objective it was to predict of a person being susceptible to heart disease.

## Generating Test Design

Both algorithms were placed inside the training component of the cross-validation operator. In the testing component, ‘apply model’ and a ‘performance’ operator was used to output a more detailed performance overview.



Naïve Bayes Implementation



Decision Tree Implementation

## Building and Assessing the model

Decision Tree:

The decision tree algorithm was tested with multiple parameters. The gain ratio criterion was first selected with a maximum depth of 10, the results from this test showed an accuracy of 83.15%, which was just slightly above the original base line accuracy, additional depth sizes of 10 and 20 were also implemented but did not event in a greater difference to the performance output. The information gain and Gini index criterion showed significant improvement. Both implementations showed an improvement on the gain ratio, information gain showed the greatest improvement on accuracy, precision and recall.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Accuracy | Precision | Recall |
| Default Model | 61.58% +/- 4.15% | 58.12% +/- 5.24% | 49.53% +/- 9.23% |
| Decision Tree  (ratio gain) | 81.85% +/- 3.88% | 87.72% +/- 7.41% | 70.01% +/- 9.85% |
| Decision Tree  (info gain) | 89.91% +/- 4.42% | 89.82% +/- 5.72% | 87.59% +/- 7.20% |
| Decision Tree  (Gini index) | 87.82% +/- 3.82% | 89.34% +/- 5.50% | 87.80% +/- 6.39% |
| Naïve Bayes | 80.28% +/- 3.44% | 81.39% +/- 4.58% | 72.91% +/- 8.05% |

The decision tree algorithm used the chest pain type attribute as the root node, max heart rate and carotid thickness were also seen to be used as determining factors following the root node. The full decision tree is too large to place in the document but the first nodes can be seen below.



# Evaluation

Creating a model to determine whether a person is or is not susceptible to heart disease is a challenging process. The multiple attempts of implementation resulted in a maximum prediction accuracy of 89.91%. Although this accuracy is high, it is not reliable enough to determine in fact whether a person will or will not develop or already have heart disease but it is reliable enough to determine whether a person should be sent for testing. This model can determine whether a person is or is not susceptible to heart disease with as little as two pieces of information in some circumstances and a maximum of ten. Although there are many factors, chest pain type and max heart rate can be seen as important contributors towards the making of a decision.

When evaluating the model, it was found that decision tree was the best algorithm to use for classification. There are many algorithms that were not yet tested on this dataset such as neural networks and regression, these algorithms are planned to be implemented later. The author feels more time should have been spent on data preparation and pre-processing and further exploration of the dataset, unfortunately due to a decision to change mid-development from python to rapid miner, there were timing constrictions for development.

## Conclusion

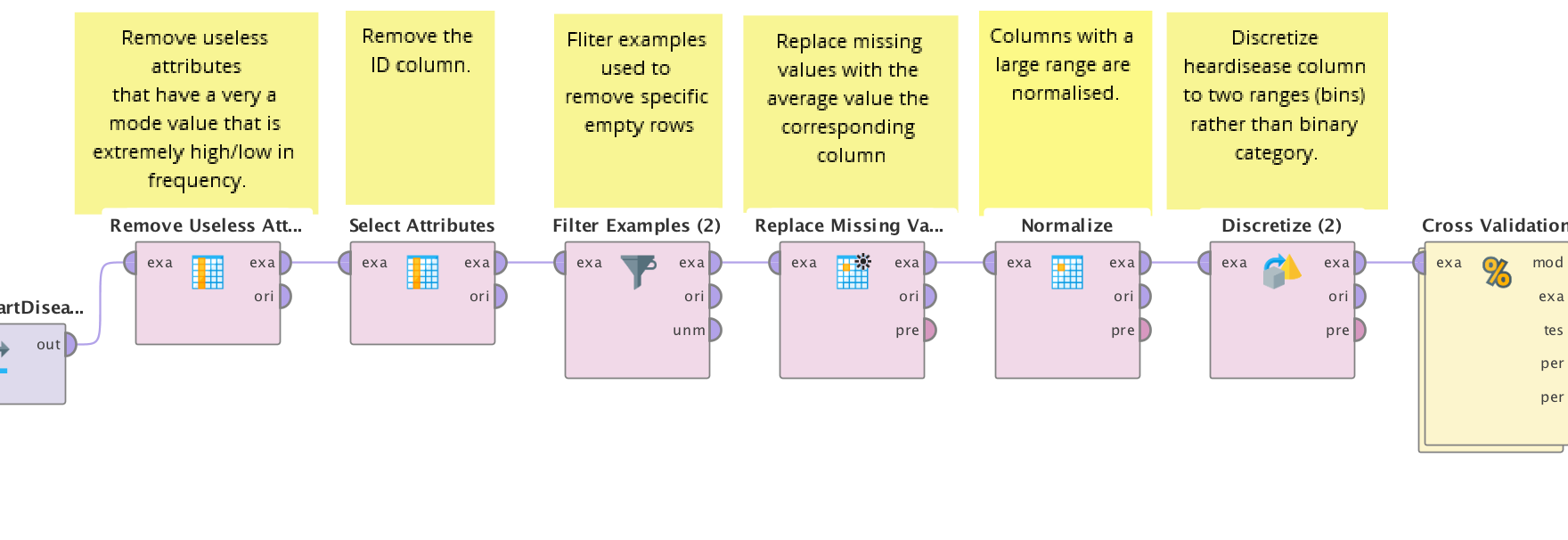
To conclude, the creation of a classification model to complete the business objective is a hard task, it may have been easier to create a model if the author had a greater understanding of the attributes themselves as it may would have been easier to identify correct / incorrect outliers but this may potentially have added bias to the model as the predictive attributes may have been assumed incorrectly and important attributes removed. Therefore, the best business solution in this scenario is to ensure that the dataset is one hundred percent correct before being given to the data analytics team.

# References

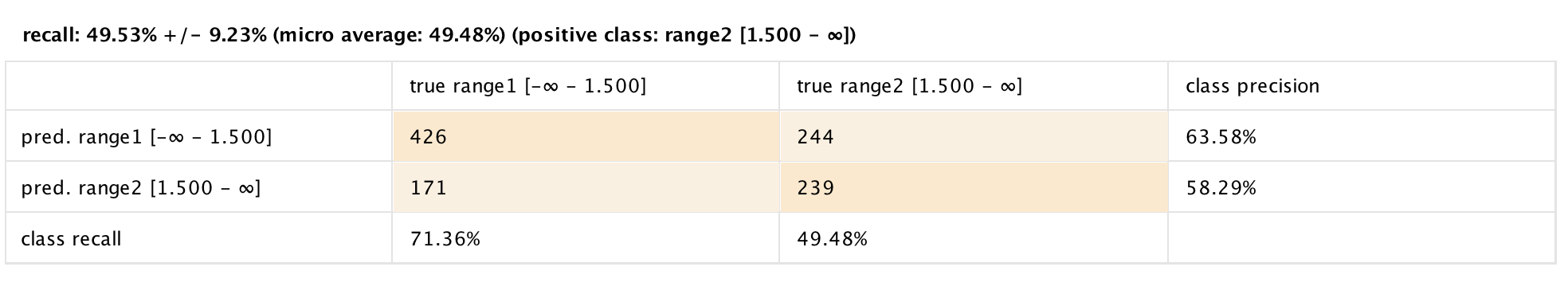
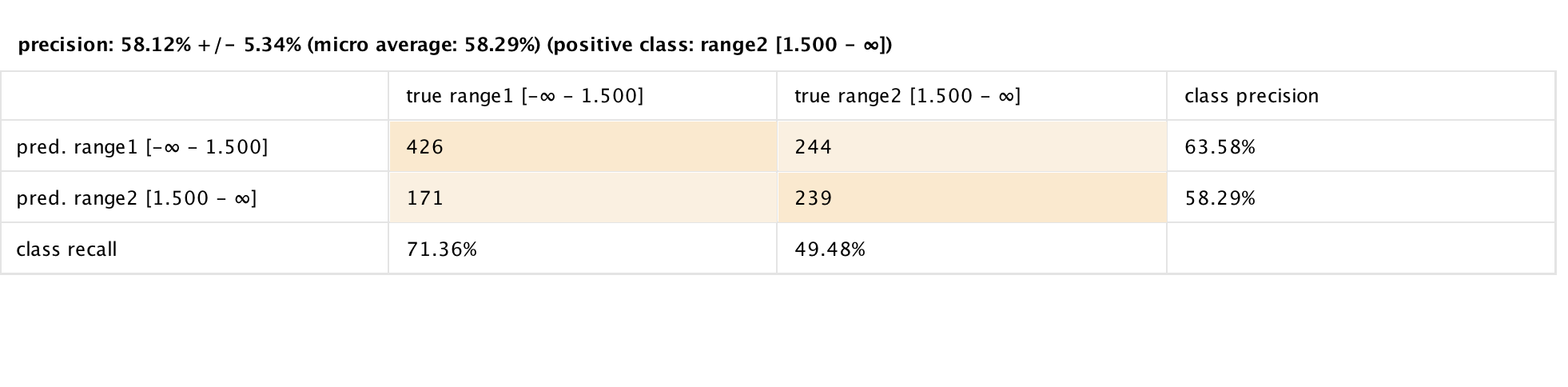
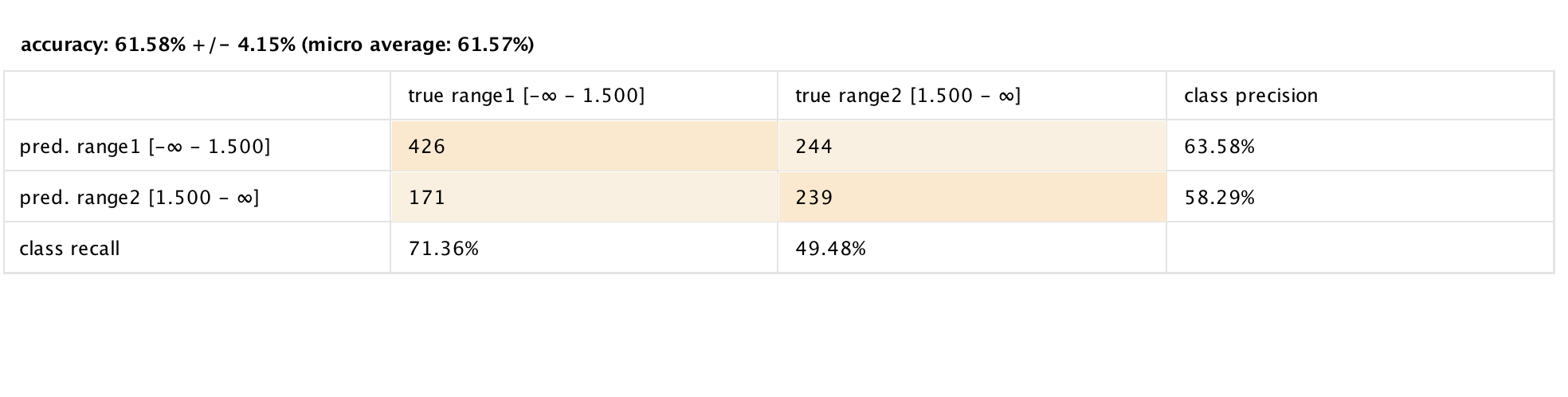
[1] HSE. 2019. Coronary heart disease. [ONLINE] Available at: <https://www.hse.ie/eng/health/az/c/coronary-heart-disease/>. [Accessed 11 December 2019].

# Appendix

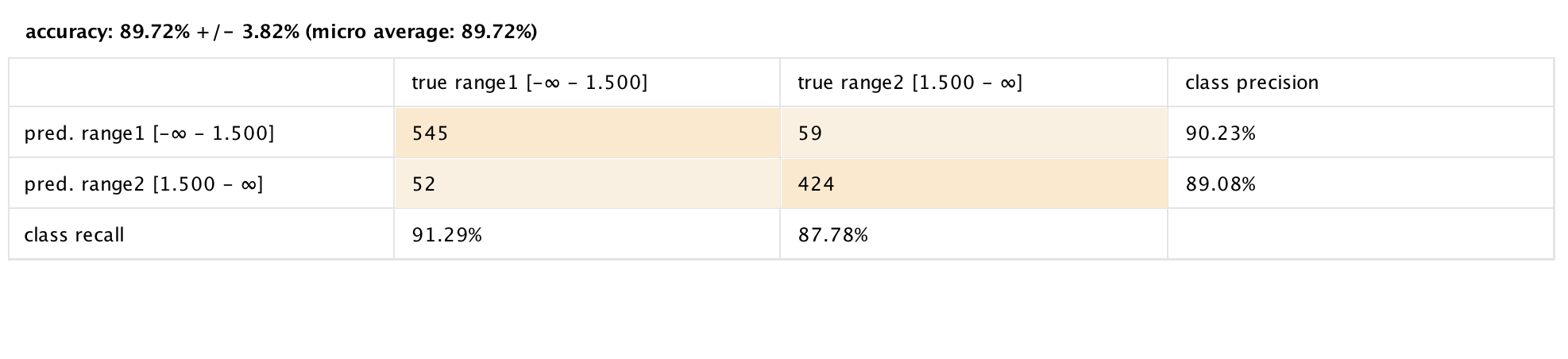
Rapid Miner Process

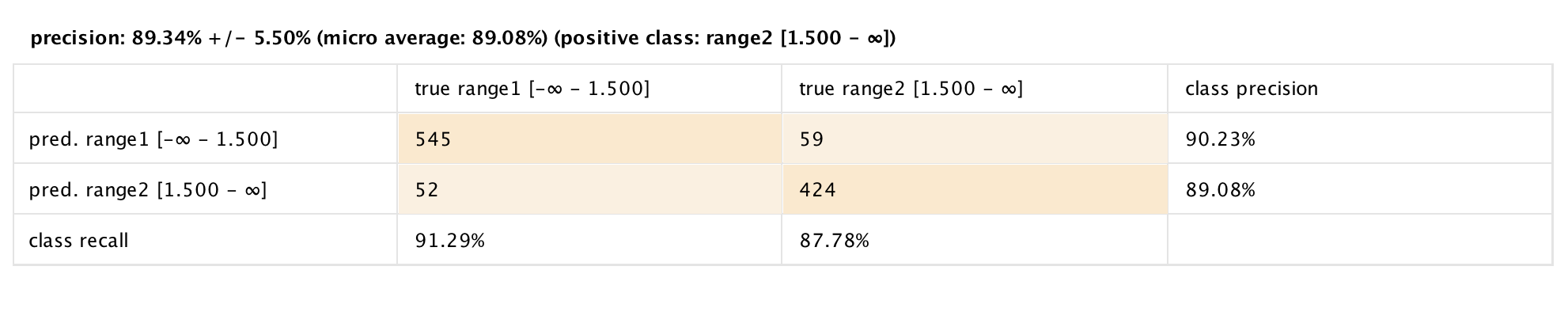


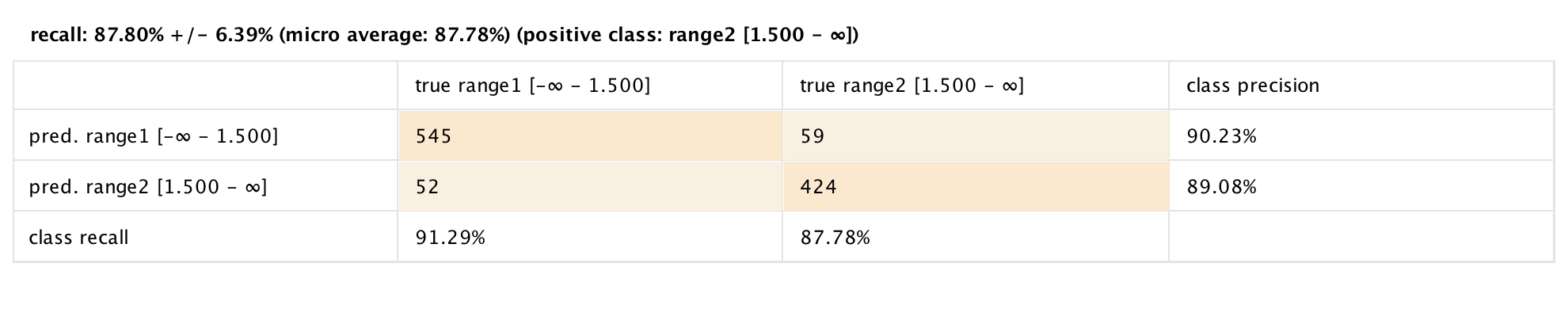
Default Model Results



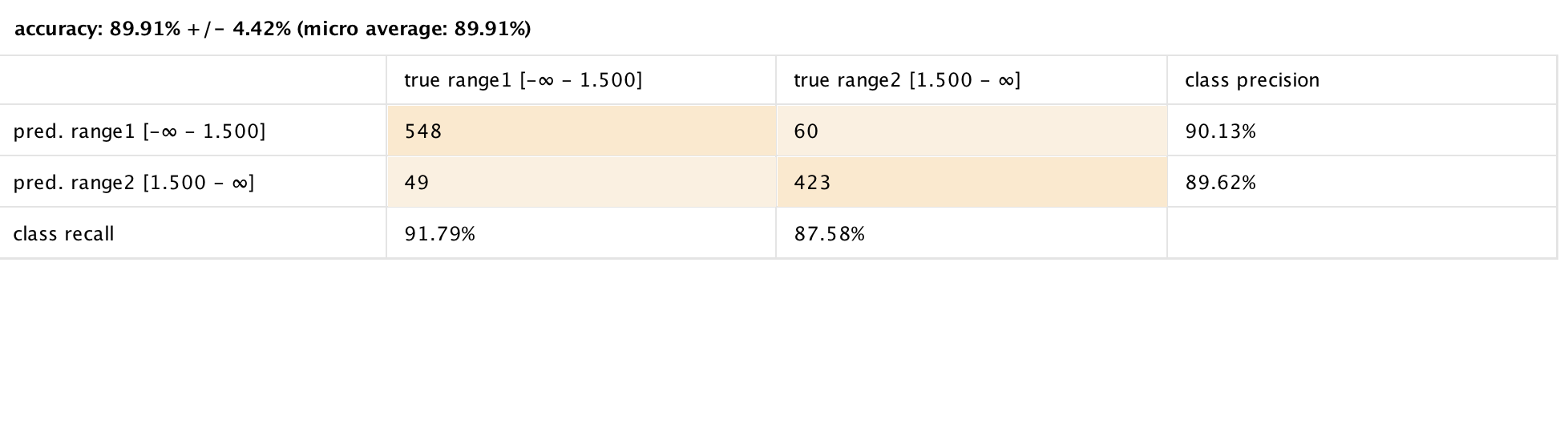
Gini Index Results

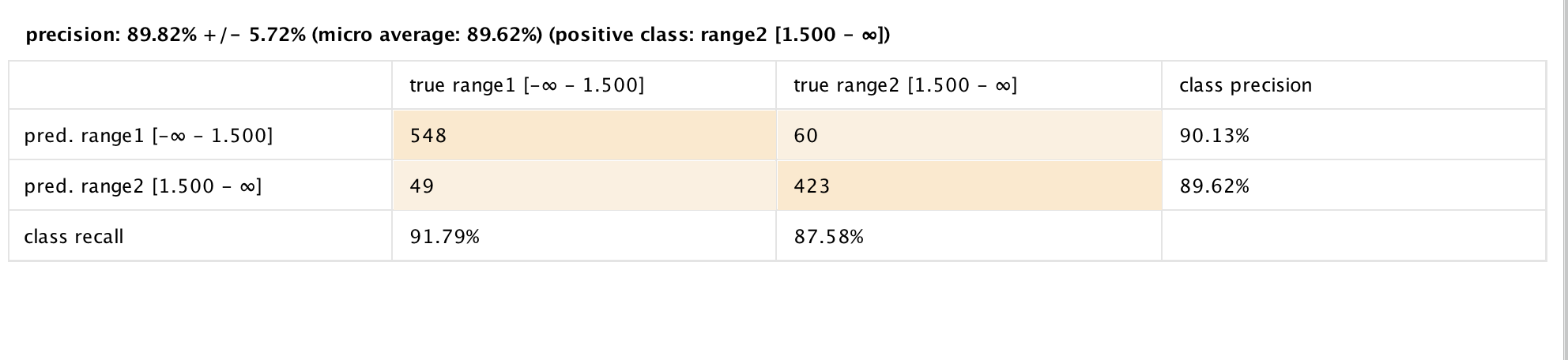


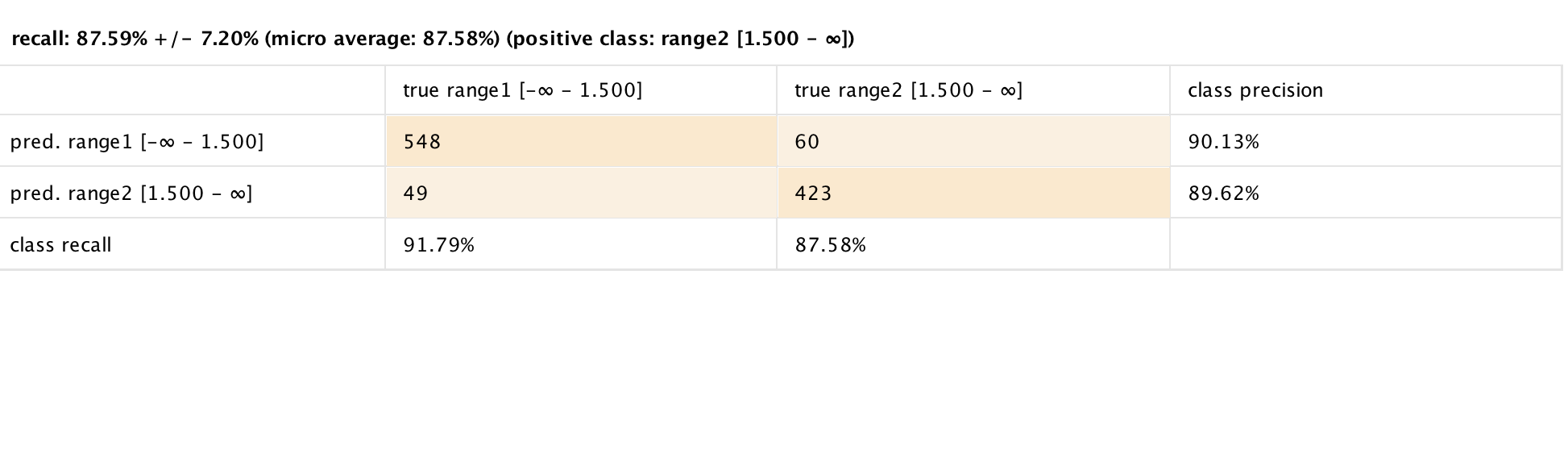




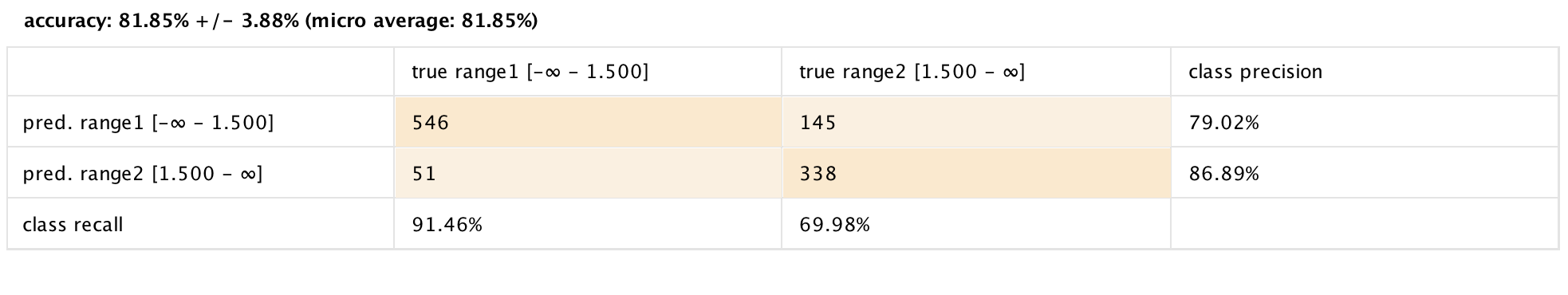
Information Gain Results

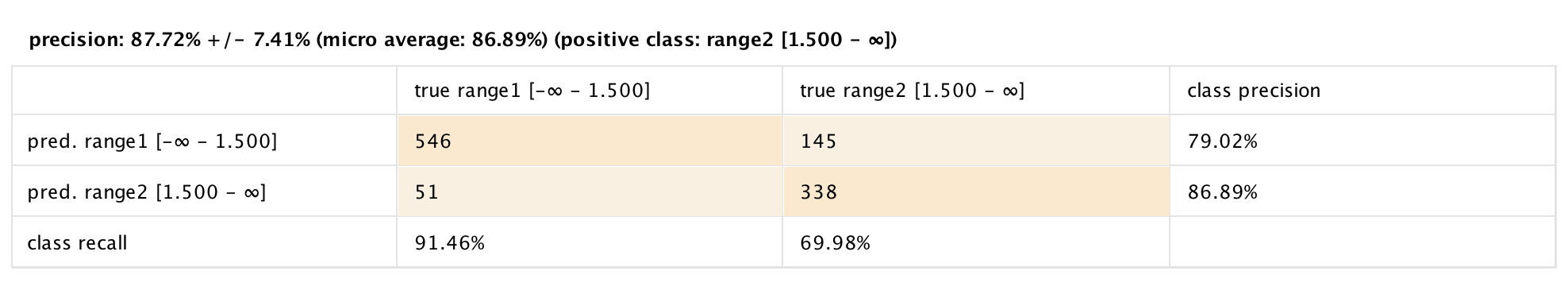


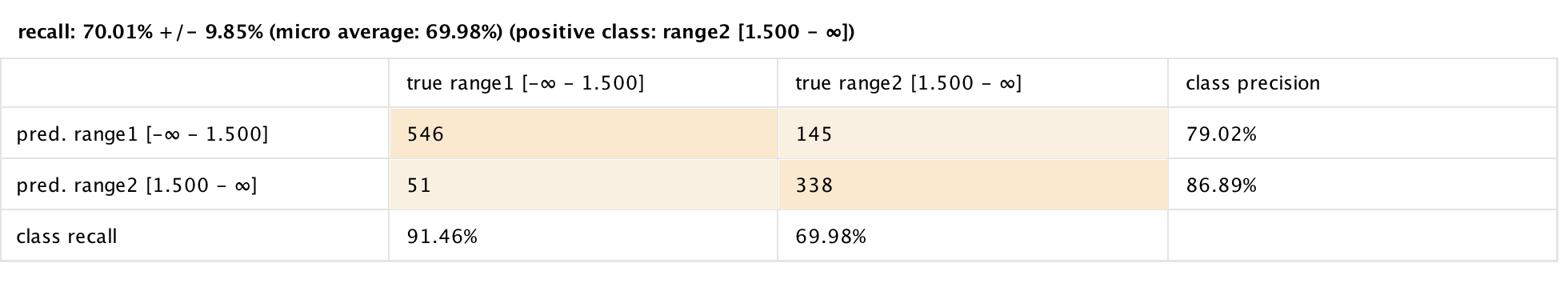




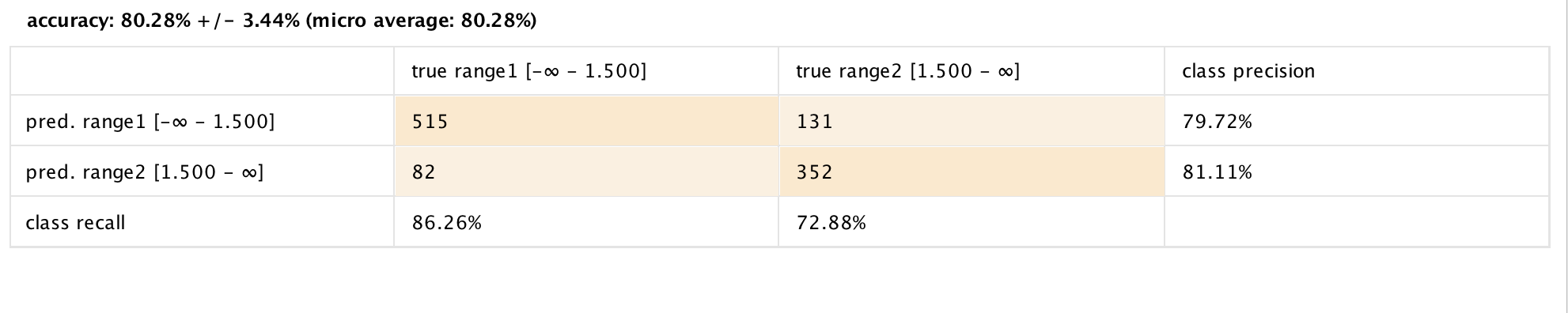
Gain Ratio Results

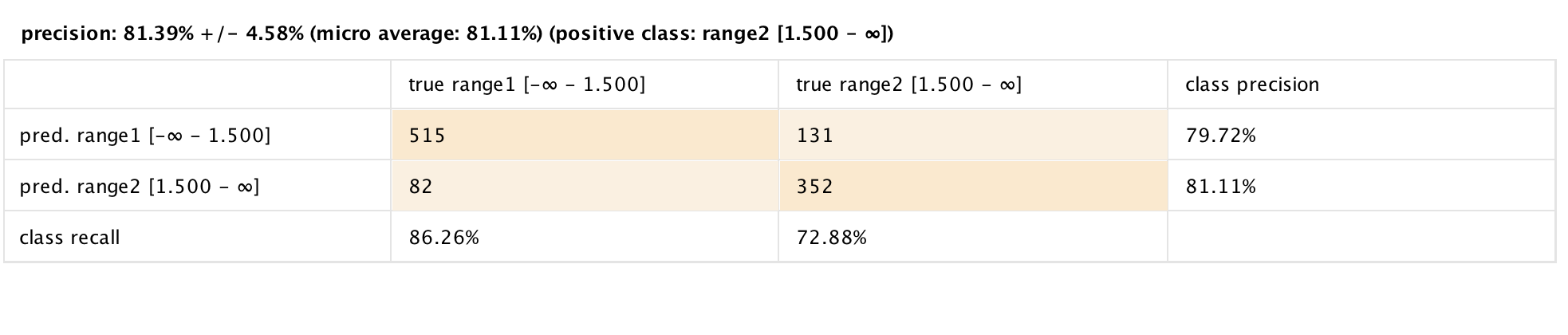


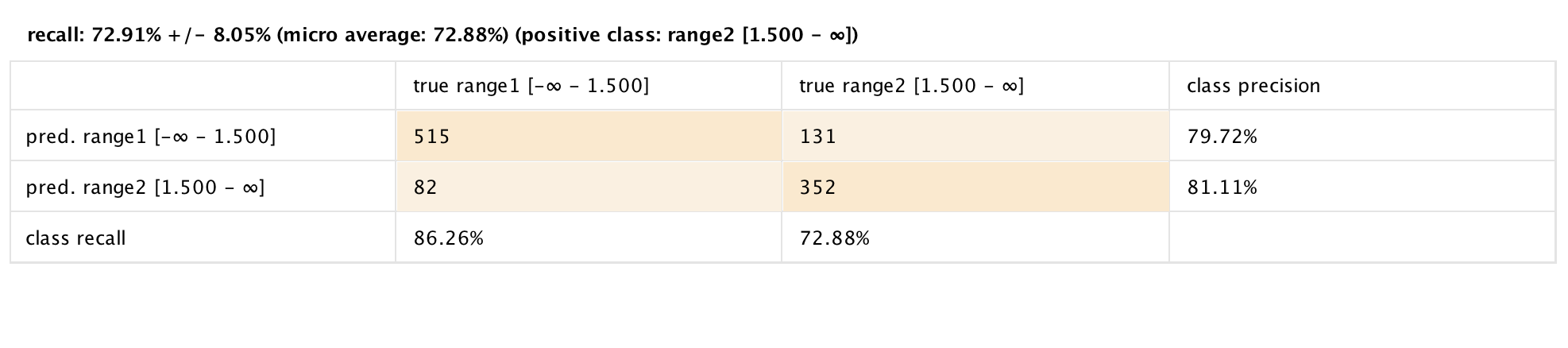




Naïve Bayes Results







Decision Tree

