zerocode
learning

Learning Data Analytics Made Easy

USER GUIDE

DECISION TREE ANALYSIS

TABLE OF CONTENTS

INDEX

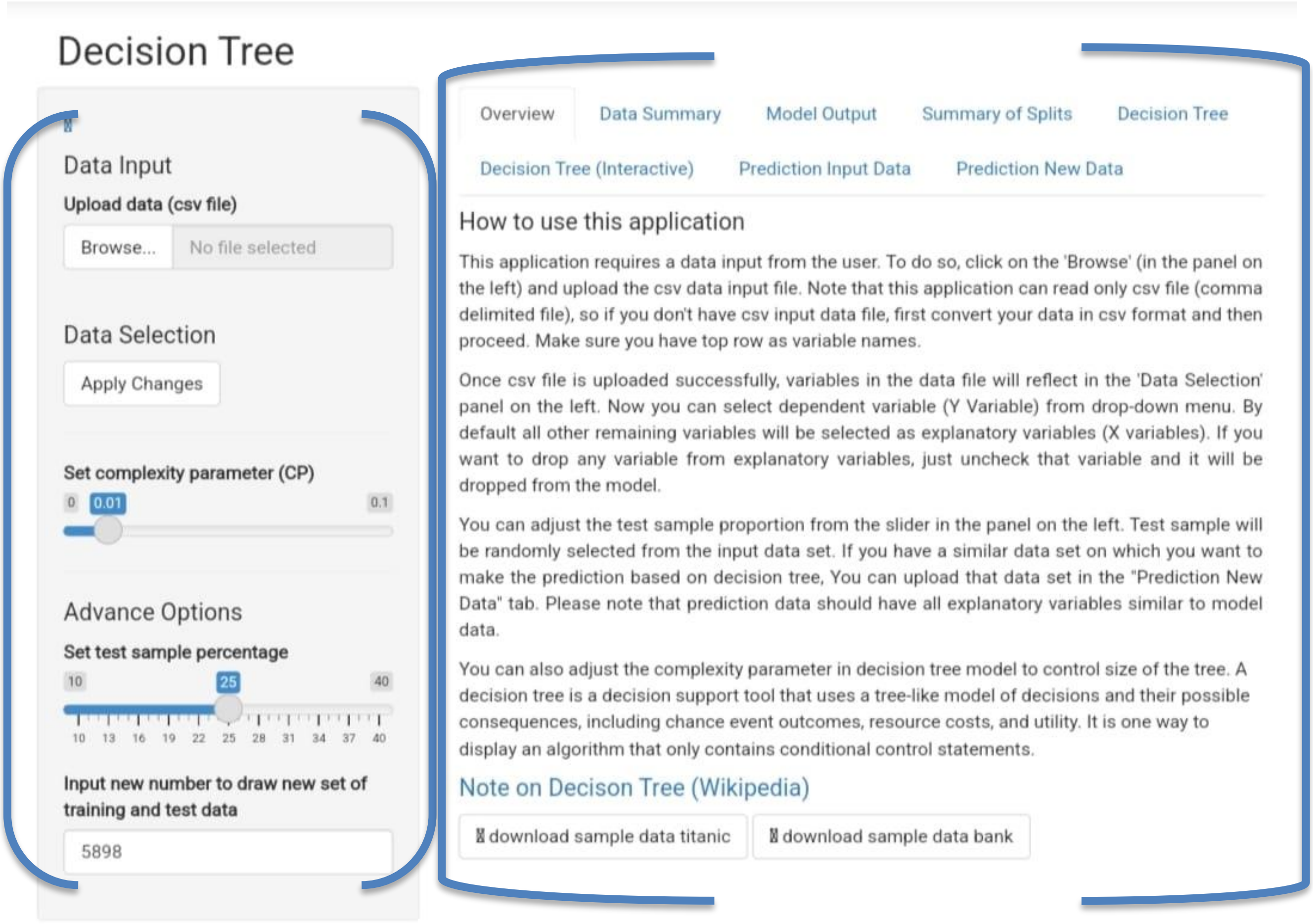
DECISION TREE ANALYSIS

01

Decision Tree, A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

LEFT PANEL (INPUT AREA)

OPERATIONAL ANALYSIS TAB (MAIN PANEL)



Decision Tree

Data Input

Upload data (csv file)

Browse... No file selected

Data Selection

Apply Changes

Set complexity parameter (CP)

0 0.01 0.1

Advance Options

Set test sample percentage

10 25 40

10 13 16 19 22 25 28 31 34 37 40

Input new number to draw new set of training and test data

5898

Overview | Data Summary | Model Output | Summary of Splits | Decision Tree | Decision Tree (Interactive) | Prediction Input Data | Prediction New Data

How to use this application

This application requires a data input from the user. To do so, click on the 'Browse' (in the panel on the left) and upload the csv data input file. Note that this application can read only csv file (comma delimited file), so if you don't have csv input data file, first convert your data in csv format and then proceed. Make sure you have top row as variable names.

Once csv file is uploaded successfully, variables in the data file will reflect in the 'Data Selection' panel on the left. Now you can select dependent variable (Y Variable) from drop-down menu. By default all other remaining variables will be selected as explanatory variables (X variables). If you want to drop any variable from explanatory variables, just uncheck that variable and it will be dropped from the model.

You can adjust the test sample proportion from the slider in the panel on the left. Test sample will be randomly selected from the input data set. If you have a similar data set on which you want to make the prediction based on decision tree, You can upload that data set in the "Prediction New Data" tab. Please note that prediction data should have all explanatory variables similar to model data.

You can also adjust the complexity parameter in decision tree model to control size of the tree. A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

Note on Decison Tree (Wikipedia)

download sample data titanic | download sample data bank

LEFT PANEL (INP)

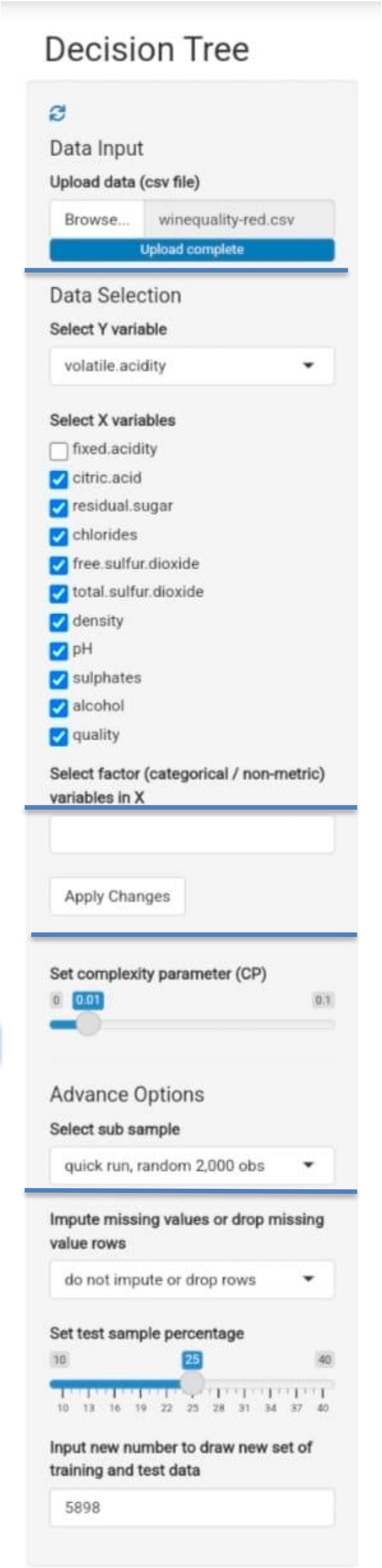
Upload your dataset here

Select your favorable variables required to base the analysis

Apply any changes if you want to do.

Select the subsamples or the whole data for testing. Set complexity level for analysis

Deal with missing values either drop or immute it.
Set the percentage sample as per the requirement for analysis.



Decision Tree

Data Input

Upload data (csv file)

Browse... winequality-red.csv

Upload complete

Data Selection

Select Y variable

volatile.acidity

Select X variables

- ☐ fixed.acidity
- ☑ citric.acid
- ☑ residual.sugar
- ☑ chlorides
- ☑ free.sulfur.dioxide
- ☑ total.sulfur.dioxide
- ☑ density
- ☑ pH
- ☑ sulphates
- ☑ alcohol
- ☑ quality

Select factor (categorical / non-metric) variables in X

Apply Changes

Set complexity parameter (CP)

0 0.01 0.1

Advance Options

Select sub sample

quick run, random 2,000 obs

Impute missing values or drop missing value rows

do not impute or drop rows

Set test sample percentage

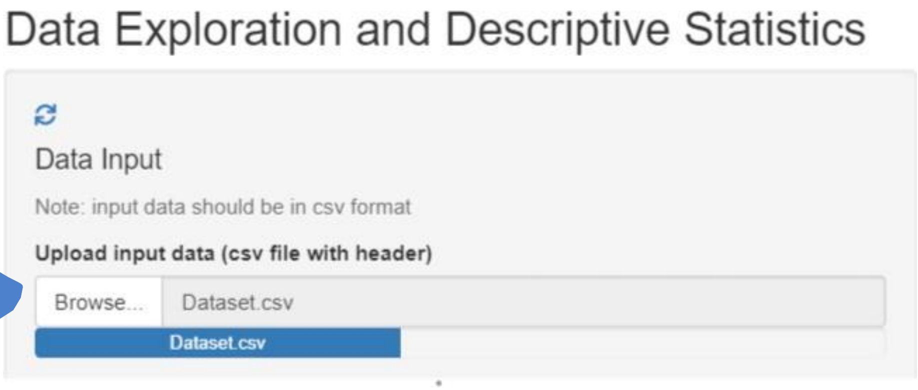
10 25 40

Input new number to draw new set of training and test data

5898

DATA INPUT
UPLOADING DATASET

- Click on browse
- Select the datafile that is in the form of csv format.(Ex program.csv)
- Browse the file and select the data to train your model for prediction.
- Top rows of the dataset should be of ‘variable names’.



Data Exploration and Descriptive Statistics

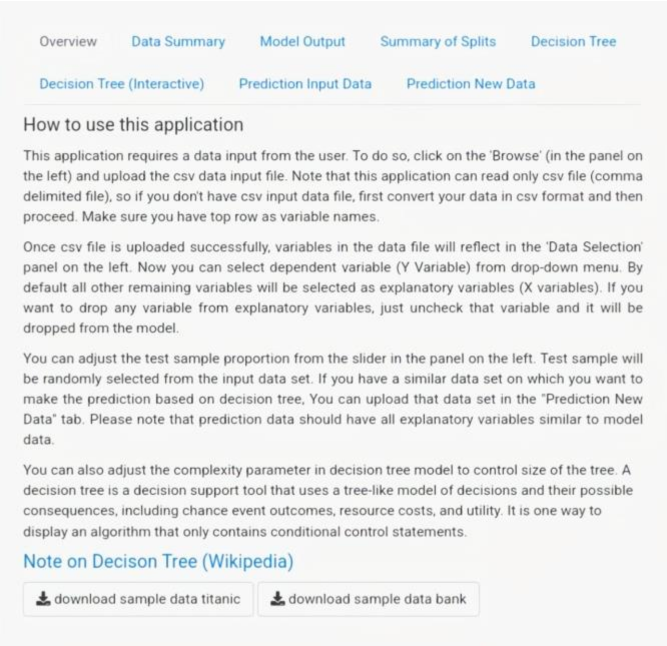
Data Input

Note: input data should be in csv format

Upload input data (csv file with header)

OVERVIEW TAB

This tab provides you with relevant study resources, tutorials, sample datasets and a short overview to start with, which helps you understand and comprehend your data correctly. This tab also provides you the basic idea about Decision tree , gives sample data and provides the description about Analysis.



How to use this application

This application requires a data input from the user. To do so, click on the 'Browse' (in the panel on the left) and upload the csv data input file. Note that this application can read only csv file (comma delimited file), so if you don't have csv input data file, first convert your data in csv format and then proceed. Make sure you have top row as variable names.

Once csv file is uploaded successfully, variables in the data file will reflect in the 'Data Selection' panel on the left. Now you can select dependent variable (Y Variable) from drop-down menu. By default all other remaining variables will be selected as explanatory variables (X variables). If you want to drop any variable from explanatory variables, just uncheck that variable and it will be dropped from the model.

You can adjust the test sample proportion from the slider in the panel on the left. Test sample will be randomly selected from the input data set. If you have a similar data set on which you want to make the prediction based on decision tree, You can upload that data set in the "Prediction New Data" tab. Please note that prediction data should have all explanatory variables similar to model data.

You can also adjust the complexity parameter in decision tree model to control size of the tree. A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

Note on Decison Tree (Wikipedia)

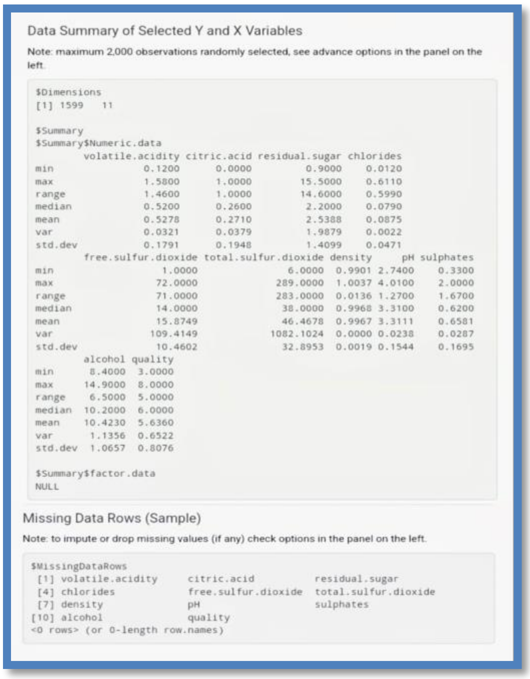
DATA SUMMARY TAB

04

It is very important to understand our data completely to infer meaningful insights and to get an overview of all the data points as a whole, but it is quite impossible to analyze thousand data points manually.

The **‘Data Summary’** option enables you to get a comprehensive evaluation through statistical measures that help us form the basis of our analysis.

It will display all the 'descriptive analytics' measures including mean, median, standard deviation, variance etc. for all the data variables present in the dataset. we can review the uploaded data and the contents of it, A brief summary of the data can be seen it includes range of data values, minimum and maximum value missing and null values etc.



Data Summary of Selected Y and X Variables

Note: maximum 2,000 observations randomly selected, see advance options in the panel on the left.

```
$Dimensions
[1] 1599   11

$Summary
$Summary$Numeric.data
        volatile.acidity citric.acid residual.sugar chlorides
min               0.1200      0.0000         0.9000    0.0120
max               1.5800      1.0000        15.5000    0.6110
range             1.4600      1.0000        14.6000    0.5990
median            0.5200      0.2600         2.2000    0.0790
mean              0.5278      0.2710         2.5388    0.0875
var               0.0321      0.0379         1.9879    0.0022
std.dev           0.1791      0.1948         1.4099    0.0471
        free.sulfur.dioxide total.sulfur.dioxide density     pH sulphates
min                  1.0000               6.0000  0.9901 2.7400    0.3300
max                 72.0000             289.0000  1.0037 4.0100    2.0000
range               71.0000             283.0000  0.0136 1.2700    1.6700
median              14.0000              38.0000  0.9968 3.3100    0.6200
mean                15.8749              46.4678  0.9967 3.3111    0.6581
var                109.4149            1082.1024  0.0000 0.0238    0.0287
std.dev             10.4602              32.8953  0.0019 0.1544    0.1695
        alcohol quality
min      8.4000  3.0000
max     14.9000  8.0000
range    6.5000  5.0000
median  10.2000  6.0000
mean    10.4230  5.6360
var      1.1356  0.6522
std.dev  1.0657  0.8076

$Summary$factor.data
NULL
```

Missing Data Rows (Sample)

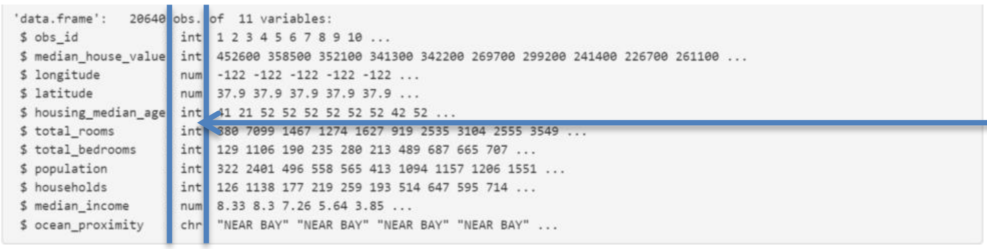
Note: to impute or drop missing values (if any) check options in the panel on the left.

```
$MissingDataRows
 [1] volatile.acidity     citric.acid          residual.sugar
 [4] chlorides            free.sulfur.dioxide  total.sulfur.dioxide
 [7] density              pH                   sulphates
[10] alcohol              quality
<0 rows> (or 0-length row.names)
```

This includes the minimum value maximum value , range between data values ,mean ,median ,mode with standard deviation that is the terms of statistics

Info about missing values

It also segregates dataset variables into respective data types, such as integer, whole numbers, character etc.



```
'data.frame':   20640 obs. of  11 variables:
 $ obs_id            : int  1 2 3 4 5 6 7 8 9 10 ...
 $ median_house_value: int  452600 358500 352100 341300 342200 269700 299200 241400 226700 261100 ...
 $ longitude         : num  -122 -122 -122 -122 -122 ...
 $ latitude          : num  37.9 37.9 37.9 37.9 37.9 ...
 $ housing_median_age: int  41 21 52 52 52 52 52 52 42 52 ...
 $ total_rooms       : int  880 7099 1467 1274 1627 919 2535 3104 2555 3549 ...
 $ total_bedrooms    : int  129 1106 190 235 280 213 489 687 665 707 ...
 $ population        : int  322 2401 496 558 565 413 1094 1157 1206 1551 ...
 $ households        : int  126 1138 177 219 259 193 514 647 595 714 ...
 $ median_income     : num  8.33 8.3 7.26 5.64 3.85 ...
 $ ocean_proximity   : chr  "NEAR BAY" "NEAR BAY" "NEAR BAY" "NEAR BAY" ...
```

Data types

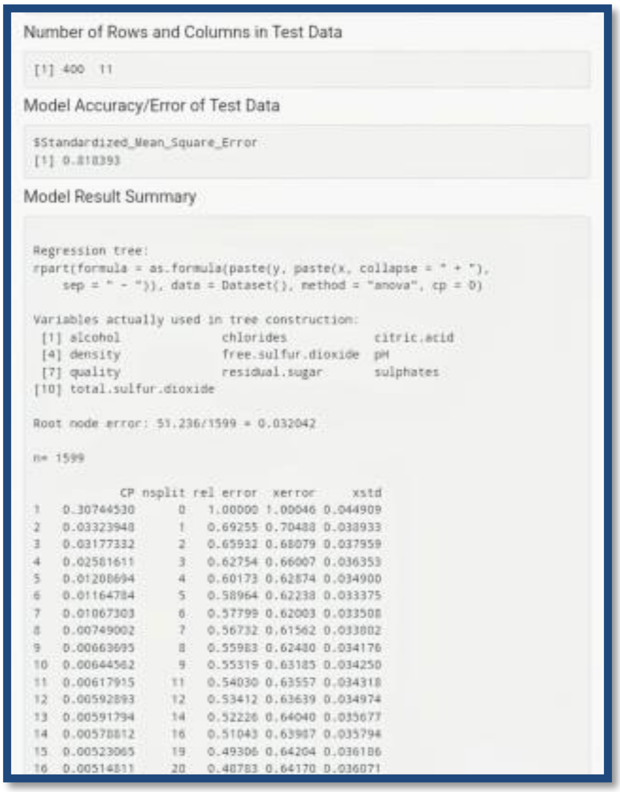
Use the left panel to transform selected variables as per the requirement of analysis ,correspondingly the data summary will also change.

05

MODEL OUTPUT TAB

Model output gives overall summary about result. This includes number of rows and columns, accuracy, result for overall attributes.

We can see that this tab provides the clarity regarding variables and their influence in the analysis.



Number of Rows and Columns in Test Data

```
[1] 400  11
```

Model Accuracy/Error of Test Data

```
$Standardized_Mean_Square_Error
[1] 0.818393
```

Model Result Summary

```
Regression tree:
rpart(formula = as.formula(paste(y, paste(x, collapse = " + "), 
    sep = " ~ ")), data = Dataset(), method = "anova", cp = 0)

Variables actually used in tree construction:
 [1] alcohol              chlorides            citric.acid         
 [4] density              free.sulfur.dioxide  pH                  
 [7] quality              residual.sugar       sulphates           
[10] total.sulfur.dioxide

Root node error: 51.236/1599 = 0.032042

n= 1599

            CP nsplit rel error  xerror     xstd
1   0.30744530      0   1.00000 1.00046 0.044909
2   0.03323948      1   0.69255 0.70488 0.038933
3   0.03177332      2   0.65932 0.68079 0.037959
4   0.02581611      3   0.62754 0.66007 0.036353
5   0.01208694      4   0.60173 0.62874 0.034900
6   0.01164784      5   0.58964 0.62238 0.033375
7   0.01067303      6   0.57799 0.62003 0.033508
8   0.00749002      7   0.56732 0.61562 0.033802
9   0.00663695      8   0.55983 0.62480 0.034176
10  0.00644562      9   0.55319 0.63185 0.034250
11  0.00617915     11   0.54030 0.63557 0.034318
12  0.00592893     12   0.53412 0.63639 0.034974
13  0.00591794     14   0.52226 0.64040 0.035677
14  0.00578812     16   0.51043 0.63987 0.035794
15  0.00523065     19   0.49306 0.64204 0.036186
16  0.00514811     20   0.48783 0.64170 0.036071
```

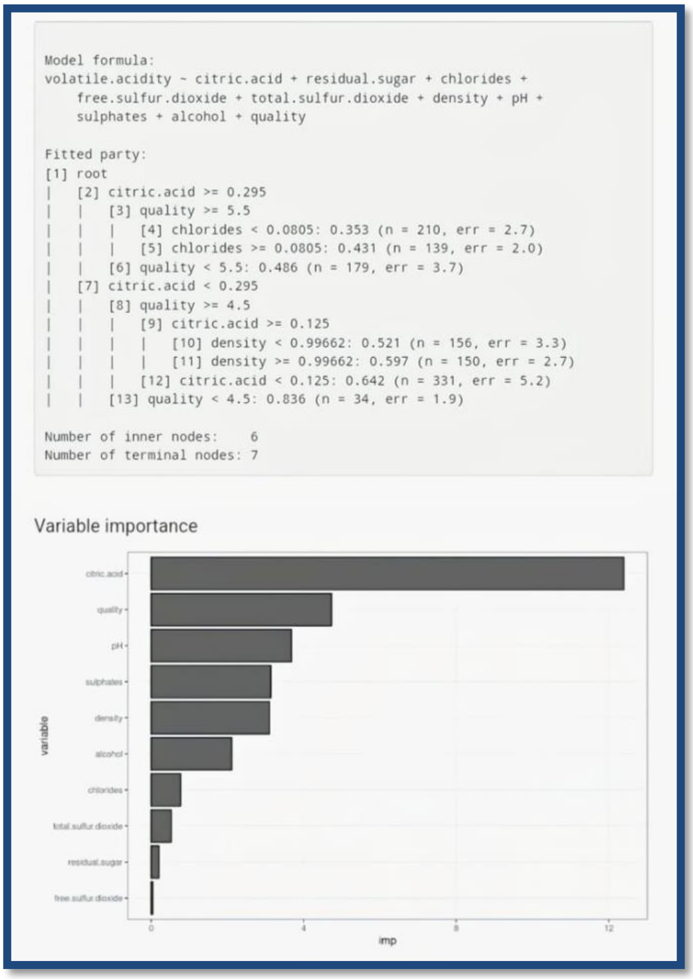
Use the left panel to modify/deal with the outliers identified here.

06

SUMMARY OF SPLITS TAB

A decision tree makes decisions by splitting nodes into sub-nodes. This process is performed multiple times during the training process until only homogenous nodes are left. And it is the only reason why a decision tree can perform so well. A decision tree makes decisions by splitting nodes into sub-nodes. This process is performed multiple times during the training process until only homogenous nodes are left.

Variable Importance: Variable importance is determined by calculating the relative influence of each variable: whether that variable was selected to split on during the tree building process, and how much the squared error (over all trees) improved (decreased) as a result.



```
Model formula:
volatile.acidity ~ citric.acid + residual.sugar + chlorides +
    free.sulfur.dioxide + total.sulfur.dioxide + density + pH +
    sulphates + alcohol + quality

Fitted party:
[1] root
|   [2] citric.acid >= 0.295
|   |   [3] quality >= 5.5
|   |   |   [4] chlorides < 0.0805: 0.353 (n = 210, err = 2.7)
|   |   |   [5] chlorides >= 0.0805: 0.431 (n = 139, err = 2.0)
|   |   [6] quality < 5.5: 0.486 (n = 179, err = 3.7)
|   [7] citric.acid < 0.295
|   |   [8] quality >= 4.5
|   |   |   [9] citric.acid >= 0.125
|   |   |   |   [10] density < 0.99662: 0.521 (n = 156, err = 3.3)
|   |   |   |   [11] density >= 0.99662: 0.597 (n = 150, err = 2.7)
|   |   |   [12] citric.acid < 0.125: 0.642 (n = 331, err = 5.2)
|   |   [13] quality < 4.5: 0.836 (n = 34, err = 1.9)

Number of inner nodes:    6
Number of terminal nodes: 7
```

Variable importance

As we can see variable importance and summary of splits are very much necessary in minimising the variables .
We can see the model formula and the complete root to leaf node analysis

We can see the influence of the each variable present in the data. We have importance and variables considered to plot as per the data the graph is plotted

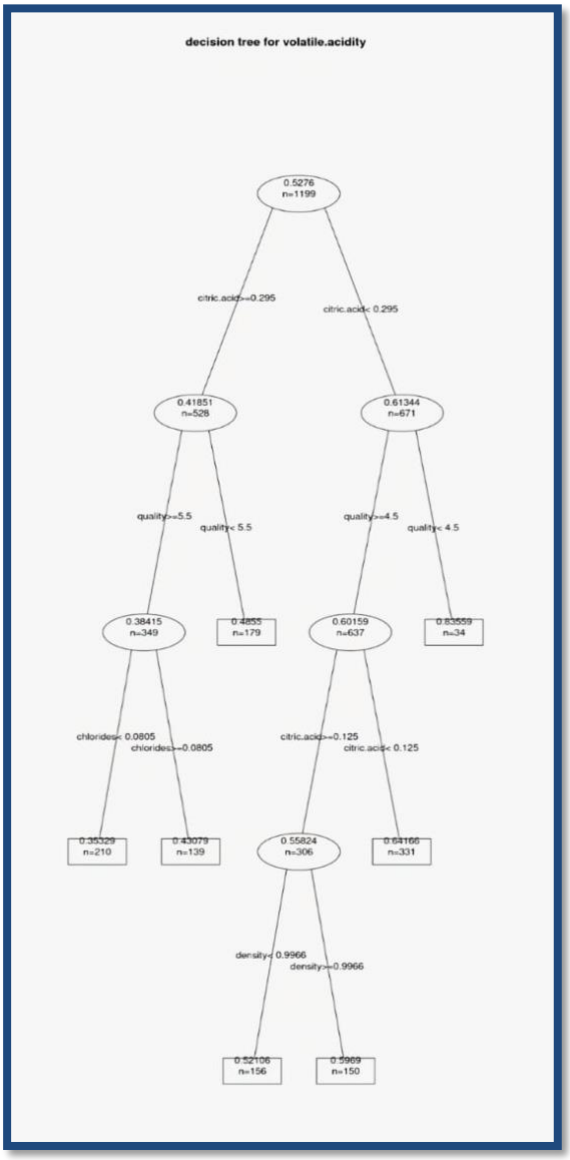
Variable importance analysis provides the tools to assess the importance of input variables when dealing with complex interactions, making the machine learning model more interpretable and computationally more efficient. In classification problems with imbalanced datasets, this task is even more challenging.

07

DECISION TREE

A decision tree is a flowchart that starts with one main idea and then branches out based on the consequences of your decisions. It's called a “decision tree” because the model typically looks like a tree with branches.

Decision trees help you to evaluate your options. Decision Trees are excellent tools for helping you to choose between several courses of action.
They provide a highly effective structure within which you can lay out options and investigate the possible outcomes of choosing those options.

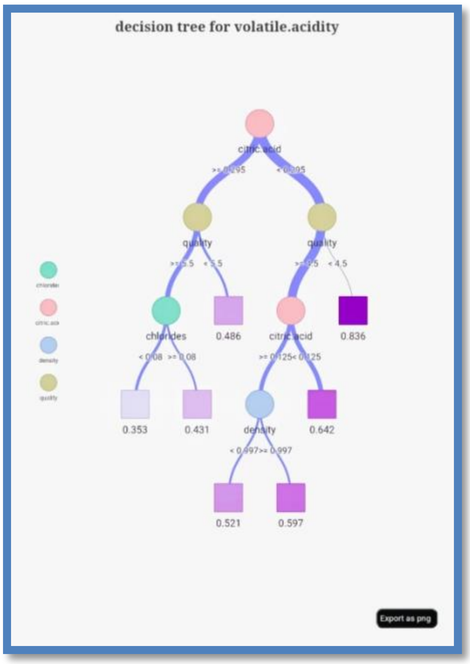


This is the actual tree that tests the quality of the red wine based on the input data
We can see the root node along Parent and child nodes, the last level is referred as leaf node and levels represents the height of the tree

Use the left panel to impute or drop the missing values identified here

08

DECISION TREE (INTERACTIVE)



Interactively exploring a filtered decision tree helps to keep a clear view of the decision process.
A decision tree guides a user from an initial question into one of the multiple possible end states.
We can zoom and click on the nodes to know the details regarding various aspects included in the analysis.
This gives better description and detail oriented regarding paths and the node analysis also variable influence over the root to leaf i.e top to bottom order of decision tree.