|  |  |
| --- | --- |
|  | |
| Python Machine Learning Project  Data and Web Mining CA – Report | |
|  | |
| Ciaran Finnegan / Dermot Madsen  Student No : 10524150 / 10522567  10/02/2020 |  |
|  |  |

List of contents

1 Project Overview 3

1.1 High Level Description 3

1.2 The CRISP-DM Methodology / Reference Model 3

1.3 Development Environments 6

2 Business Understanding 7

2.1 Determine Business Objectives 7

2.2 Assess Situation 8

2.3 Determine Data Mining Goals 9

2.4 Produce Project Plan 9

3 Data Understanding 11

3.1 Collect Initial Data 11

3.2 Describe Data 12

3.3 Explore Data 13

3.4 Verify Data Quality 22

4 Data Preparation 24

4.1 Select Data 24

4.2 Clean Data 25

4.3 Construct Data 26

4.4 Integrate Data 31

4.5 Format Data 31

5 Modelling 32

5.1 Select Modelling Technique 32

5.2 Generate Test Design 32

5.3 Build Model 32

5.4 Assess Model 32

6 Evaluation 33

6.1 Evaluate Results 33

6.2 Review Process 33

6.3 Determine Next Steps 33

7 Deployment 34

7.1 Plan Deployment 34

7.2 Plan Monitoring and Maintenance 34

7.3 Produce Final Report 34

7.4 Review Project 34

8 Conclusion 35

8.1 Conclusion.. 35

9 Appendices and References 36

9.1 Appendix A Understanding Wine and Types 36

9.2 The White Wine Datset 38

9.3 References 39

# Project Overview

## High Level Description

This document covers our planned approach and execution of a data mining analysis on a dataset relating to the assessment and prediction of wine quality.

Following the CRISP-DM model, we laid out an objective for this Continuous Assessment exercise and followed a series of steps, often iteratively to arrive at a predictive model for wine quality based on known feature characteristics.

The following sections explain the business objectives, the assessment of data, and the selection, implementation and deployment of a model to provide a predictive guide to new wine quality.

## The CRISP-DM Methodology / Reference Model

In the mid to late 1990s, business markets were showing a sharp upturn in interest into the possibilities offered by data mining practices. The need for a standard process model, widely and freely available, became quickly apparent.

By 1999/2000, a process model named CRISP-DM (Cross-Industry Standard Process for Data Mining) had been produced by leading thinkers in the industry. It was based on practical, real-world experiences and sought input across a range of business domains.

As explained in the following sections of this document, we have taken the key principles of CRISP-DM to implement our CA project.

### Methodology

The CRISP-DM methodology is described as a hierarchical process mode with four levels that transition from the generic to the specific;

1. **Phases** – process blocks consisting of several generic tasks.
2. **Generic Tasks** – so called because they are intended to be robust and stable tasks that can apply in any data mining situation.
3. **Specialised Tasks** – a description as to how the generic tasks should be applied in specific situations. Very often these tasks can be performed in multiple orders and repeated a number of times.
4. **Process Instances** – this is a record of the actions, decisions, and results of an actual data mining engagement.

### Reference Model

The life cycle of a data mining project consists of six phases, as shown in this image below.



Figure 1

The sequence of the phases is not rigid. In our Wine Quality project moving back and forth between different phases was frequently required, as expected.

The outcome of each phase determines which phase, or particular task of a phase, has to be performed next. The arrows indicate the most important and frequent dependencies between phases.

As an example, in our Wine Quality CA we needed to…<provide some actual examples – when we have them…!>

The diagram above displays the following phases.

* Business Understanding
* Data Understanding
* Data Preparation
* Modelling
* Evaluation
* Deployment

The next sections of this document elaborate on these phases a little further and the remainder of the report describes the actual implementation of the CRISP-DM against our Wine Quality project.

### Business Understanding

Understand the project objectives and the requirements from a business perspective.

Convert this knowledge into an actual data mining problem definition, along with a project plan that will provide a framework to deliver the business objectives.

### Data Understanding

Start with initial data collection.

Proceed into activities that provide familiarity with the data, including data quality issues, data insight, and possible sub-sets within the data.

### Data Preparation

Activities to construct the final dataset that will be fed into the modelling too.

Data preparation tasks can be performed in multiple orders and over many interactions.

### Modelling

Select and apply various modelling techniques.

Calibrate parameters to provide optimal values within the model.

Revert to data preparation phase, if necessary.

### Evaluation

A high quality data analysis model has been built.

Assess that the model achieves the business objective.

### Deployment

The knowledge gained by the creation of the model will need to be organised and presented in a way that it can be used by the customer.

It is important for the customer to know up front what actions need to be carried out in order to actually make use of the created models.

## Development Environments

The majority of development took place within the RapidMiner toolkit and screenshots are provided to show the step-by-step working.



Figure 2

Supplementary data investigation was conducted using a Python project developed in Visual Studio 2019. The structure of the project was modularised around a framework similar to the CRISP-DM model.

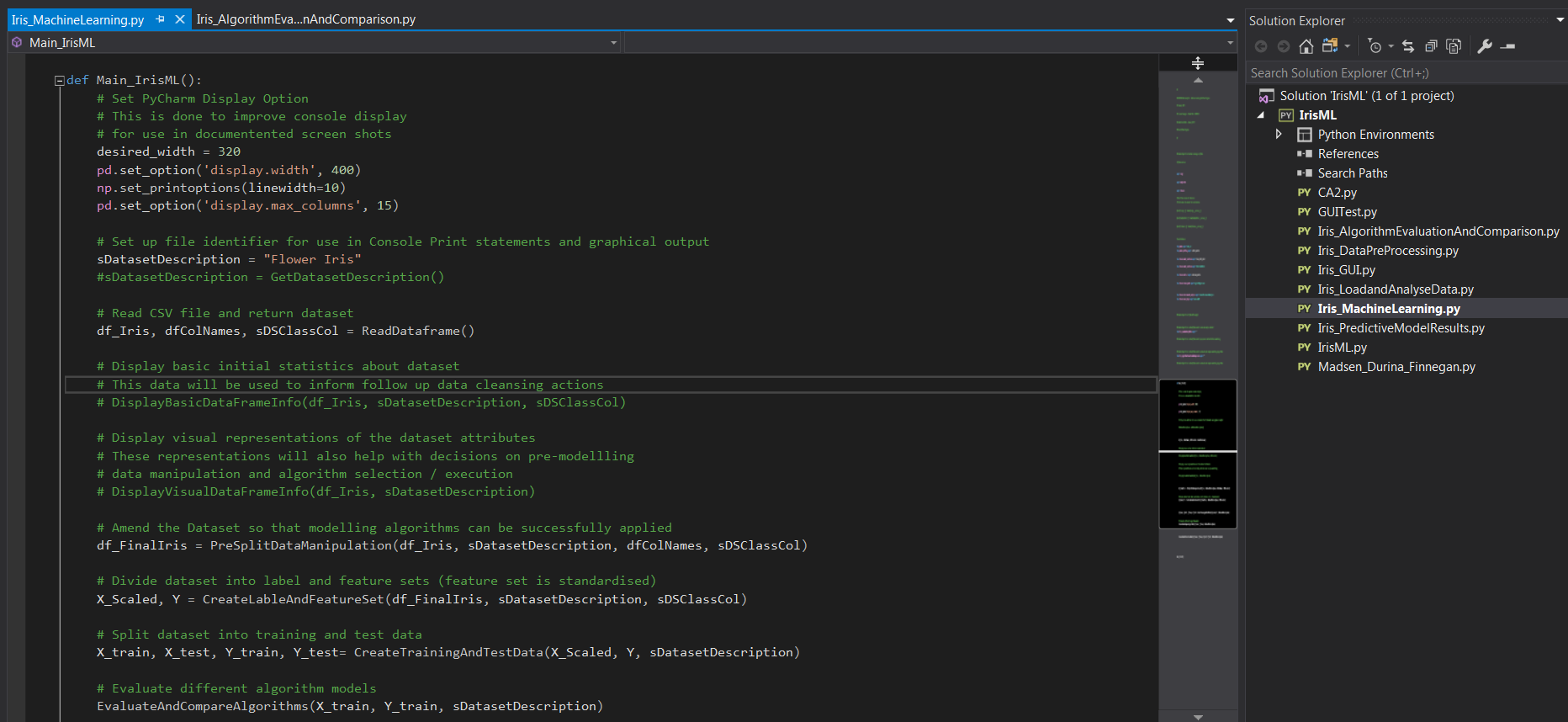


Figure 3

The Python project was used to provide quick to develop validations for the RapidMiner process and the ‘Performance’ outputs provided by that tool.

# Business Understanding

## Determine Business Objectives

The first objective of the data analyst is to thoroughly understand, from a business perspective, what the customer really wants to accomplish.

In our QA, the ‘customer’ and ‘business’ are obviously a theoretical concept. However, given that our chosen dataset relates to Wine Quality, we are assuming the role in the project of a chain of Off-Licence shops, who have a particularly speciality in selling Portuguese “Vinho Verde” red wine. Part of the business USP (unique selling point) is that staff is encouraged to be knowledgeable about the quality of this wine that they may recommend to customers. Although our imaginary Off-Licence chain promotes an awareness of wine amongst staff, very few employees would aspire to the level of sommelier in this brand of Portuguese wine and therefore it is necessary to provide guidance to staff when new stocks of wine arrive in-store.

Vineyards and Wine wholesalers will presumably provide recommendations on what wines are ‘good’ but a secondary objective of our business is to have a less subjective measure of quality for new wines. Our predictive model will therefore provide a more scientific basis for a quality rating, which can be applied across the entire outlet of shops, rather than relying on a human analysis, which could be open to interpretation.

A further secondary objective is that this model may provide guidance for similar in-store marketing of other ‘niche’ wine brands, should our business wish to replicate this approach to wine promotion.

How do we define success? A model is built based to predict the quality of new stocks of red wine based on the constituent chemical properties of the liquid. Our initial dataset will contain information to train and test our model (after various algorithm selections), and we will then conduct an additional test with new ‘unseen’ data to show that the model works well to provide an employee guide to wine quality.

## Assess Situation

This task involves more detailed fact-finding about all of the resources, constraints, assumptions, and other factors to be considered in determining the data analysis goals and project plan.

**Dataset Inventory**

We chose a dataset provided in the Kaggle website ([*https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009*](https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009)) that relates to the red wine variant of the Portuguese “Vinho Verde” red wine. (The primary source of the dataset is on the UCI Machine Learning Repository - [*https://archive.ics.uci.edu/ml/datasets/wine+quality*](https://archive.ics.uci.edu/ml/datasets/wine+quality)).

The dataset is publically available and provides a series of input variables, which are based on physiochemical tests, and an output variable that is a 0 – 10 score of quality.

**Assumptions**

The ‘quality’ measure, which is obviously the key characteristic we want to assess, has been assumed to come from feedback from wine industry specialists.

**Constraints**

This an assessment of quality based on the chemical constituents of the wine. There is no data relating to year or grape type, as might be expected with an assessment of wine, so we are assuming that a chemical analysis will provide the all the data points we need for a quantifiable assessment of quality.

There is also no indication of brand or price. This dataset is deliberately excluding these factors (or is unable to include them). Therefore that type of marketing data points will not influence the prediction of ‘quality’ as produced by our model.

## Determine Data Mining Goals

A data mining goal states a project objective in technical terms.

**Goal**

Our CA project aim is to build a predictive model that provides a 0 – 10 rating for a wine based a list of 11 chemical attributes in the liquid.

**Success Criteria**

We want our model to operate with a greater than 80% accuracy in its predictions of wine quality for new “Vinho Verde” red wines.

We may extend the modelling process so that a score of ‘7’ or greater is described as ‘Very Good’, ‘5 – 6’ receives a ‘Medium’ description, and anything else is ‘Poor’. Thus we refine our classification of the model outputs into simpler terms for the end user employees.

## Produce Project Plan

**Project Plan**

The framework of this document, even just reading from the Table of Contents, provides the general outline of activity.

In brief, our timelines are to complete the following activity by the following milestones (allowing for some iterations and back and forth before project completion);

Any project plan is a dynamic document, and this CA is no exception and we expected, and encountered, the need for many revisions.

* **Saturday January 25th**: Complete dataset selection and establish business objectives.
* **Saturday February 1st**: Complete Data Understanding, Data Preparation, and preliminary model assessment.
* **Saturday February 8th:** Complete Modelling and Evaluation, determine Production approach. Present to class.
* **Sunday February 9th:** Submit CA final report with recommendations.

**Assessment of Tools**

In order to gain an insight into the use of commonly used industry tool, the majority of the data mining approach was conducted in ***RapidMiner*** (as can be seen in the screenshots used throughout this document).

However, early stage data analysis and some preparation used ***Python*** scripting. This was partially because of familiarity with Python from earlier CA work on the course and also to provide some quick additional verification of the RapidMiner outputs.

Excel was used for part of the validation process on predictions made by the model on new data.

# Data Understanding

## Collect Initial Data

This involves the acquisition of data and loading into our chosen data mining tool kits.

**Initial Data Collection Report**

As described in Section 2.2 of this document the dataset for the CA is taken from the Kaggle website, specifically from the URL: <https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009>, which in turn references the original UCI Machine Learning source.



Figure n

In Appendix A of this document there is a brief guide to understanding wine types and composition, with particular relevance to the chemical data points in this dataset. (Source: *https://github.com/dipanjanS*).

Downloading the CSV file from Kaggle is a straightforward exercise and the CSV file itself is just 101 kB.

For preliminary data analysis the CVS file on Red Wine quality loads without issue into RapidMiner.



Figure n

## Describe Data

This involves an examination of the ‘gross’ (or ‘surface’) data and a report on the results.

**Data Description Report**

The file is on a CSV format, and contains 1600 row with 12 attribute columns in the following structure:

***Input variables*** (based on physicochemical tests):  
1 - fixed acidity  
2 - volatile acidity  
3 - citric acid  
4 - residual sugar  
5 - chlorides  
6 - free sulphur dioxide  
7 - total sulphur dioxide  
8 - density  
9 - pH  
10 - sulphates  
11 – alcohol

***Output*** variable (based on sensory data):  
12 - quality (score between 0 and 10)

A surface view in NotePad++ shows the following sample structure;



Figure n

In line with the description of the dataset in Kaggle, the dataset contains header information but the remainder of the dataset is purely numeric.

Given that this is a dataset aimed at relative new comers to the work of Machine Learning, it does not seem likely that there will be any invalid or missing data entries. We would expect this to be borne out in the analysis in the following sections of this document.

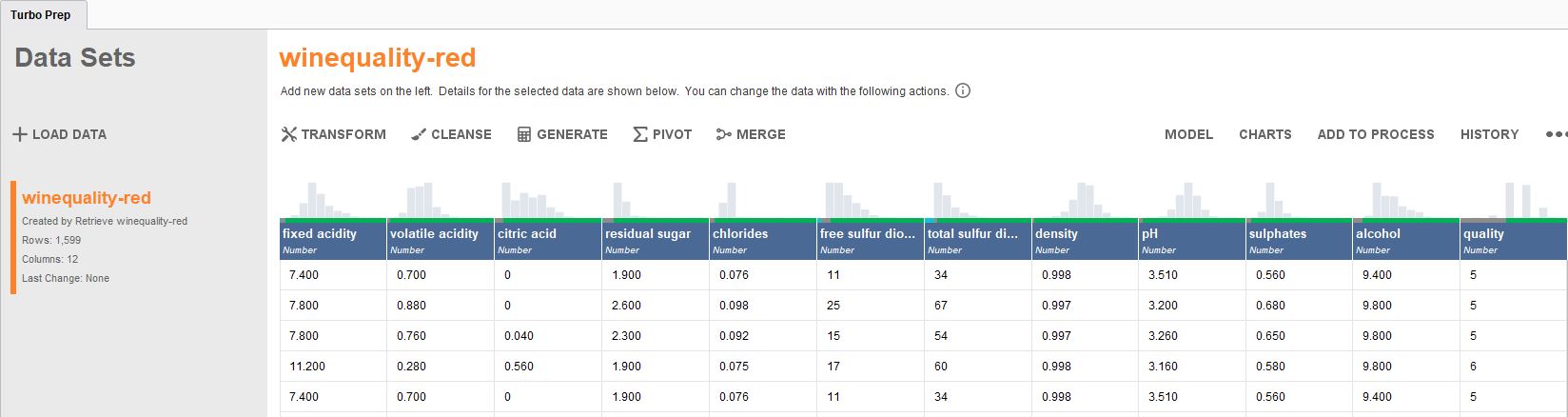
## Explore Data

This task addresses data mining questions using querying, visualization, and reporting techniques.

**Explore ‘Wine Quality’ Data**

*Assessing the Target attribute*.

The figure below shows the head rows in the dataset, in a much more readable format than with Notepad++.



As described earlier, there are 1600 rows in the dataset (one row with header information), and 12 columns.

The purpose of our data mining task to assess wine quality. Therefore the attribute column for ‘**quality**’ will be marked as our **label**. This is the value which we want to predict, based on the entries in the other attribute rows.

As will be explained in Section 4 of this CA report, we will carry out feature engineering to add an additional attribute to make the interpretation of the model easier for the end user by moving away from a numeric output to a qualitative description.

*Relationships between attributes. Correlations of attributes to the label value and between attributes themselves.*

In many datasets not every attribute will have a strong impact on the target label. The values of certain attributes may influence the model very little and can actually be ignored in the model building process.

In addition, certain attributes may be strongly correlation with each other, either negatively or positively. Commonly cited examples in data modelling are if measurements are stored using different standards – for example metric vs imperial. One of those attributes will be redundant because an increase in one attribute is reflected with a similar scale of increase in the other.

For our wine quality dataset we can execute a correlation analysis on the attributes to determine which attribute have the greatest impact on the resultant quality of the wine.

RapidMiner provides a process to carry out this analysis, either through a stand-alone process using the Correlation Matric Operator, or through the AutoModelling process.

After conducting such an analysis on our database, we can see the following correlation data between attributes and label;

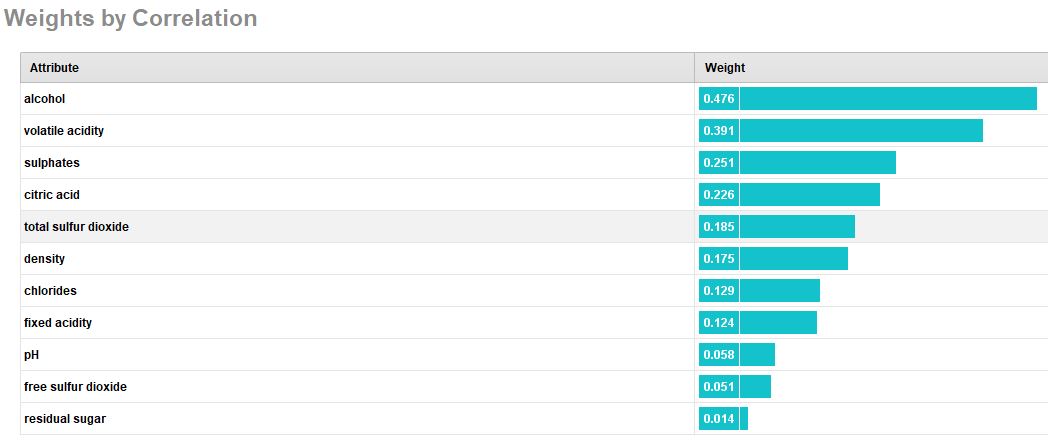


Figure n

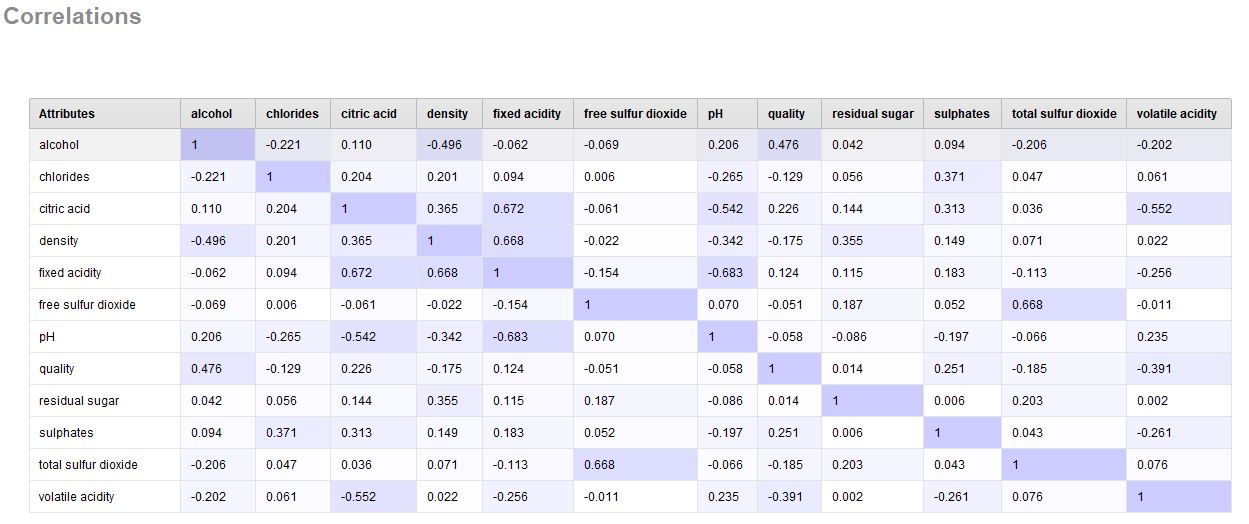
Alcohol content has the greatest impact on the resultant measure of ‘quality’.

Residual sugar appears to have the least impact on the determination of red wine quality in our sample dataset.

The diagram above shows the other relative metrics for each attribute in terms of their influence on ‘quality’.

Within the attributes themselves, correlations can be examined to determine if some attributes can be considered redundant or, at the very least, less important to include in the model calculations.

The diagram below shows the inter attribute correlations;



The darker colours indicate attributes in the feature set that have higher correlation values.

Not surprisingly the **pH** and **fixed acidity** attributes are showing are more marked negative correlation that many of the other features.

Likewise the **free sulphur dioxide** and **total sulphur dioxide** features are showing a reasonably positive correlation.

Although the Wine Quality dataset is not large in terms of rows and attributes, the principle holds that models can be improved in terms of creation and execution if the feature set is pruned to just those attributes that have the greatest impact in the model accuracy.

Taking the above considerations into account the list of attributes with which to train our potential models, which is described in detail in Section 5 of this document, can be pruned to improve performance.

*Simple Statistical Analysis of the Wine Quality Dataset*

RapidMiner provide a simple and graphical means to generate some basic statistical data on the attributes in the dataset.

The figures below provide a visual description of the histograms for each attribute in the Wine Quality dataset.

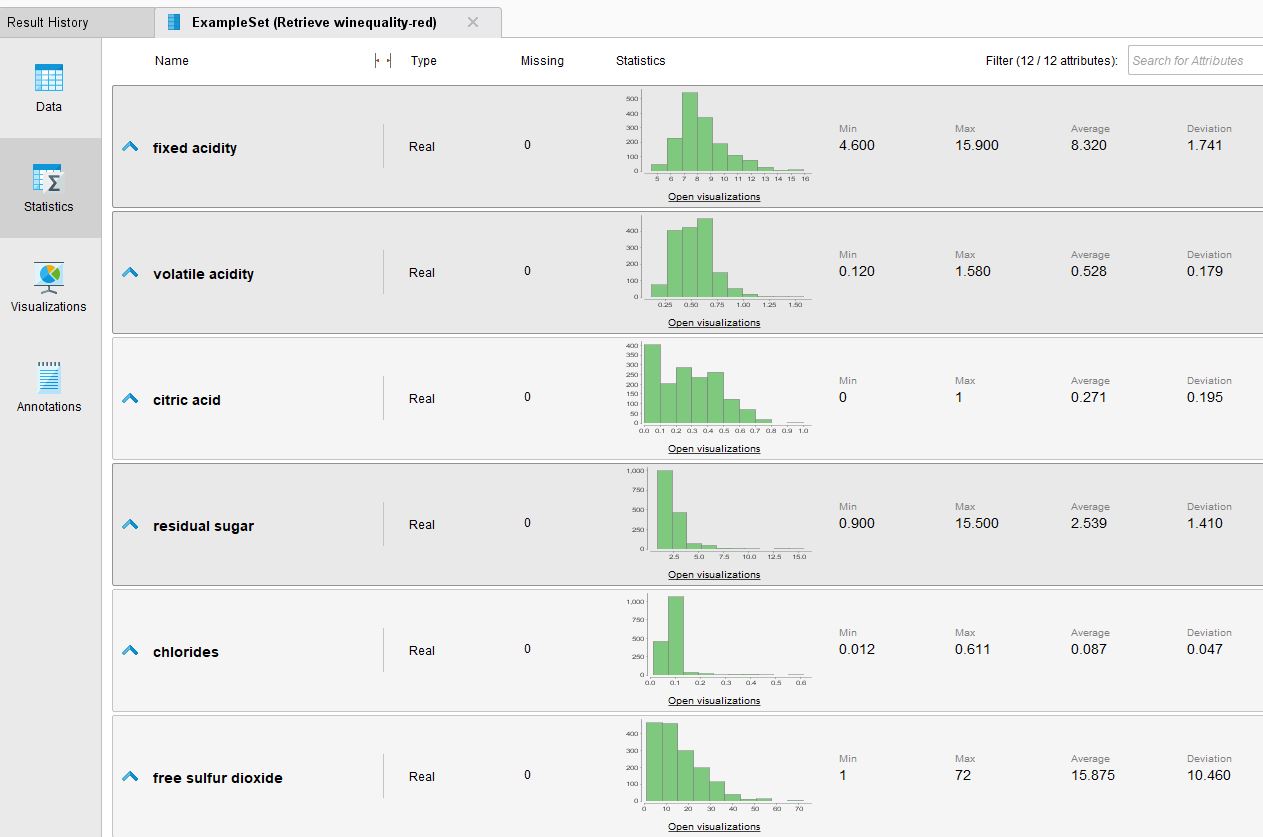


Figure n



Figure n

**Data Exploration Report**

*Initial findings / hypotheses and their impact on the remainder of the Wine Quality project.*

*Standardise and normalise the feature attributes*

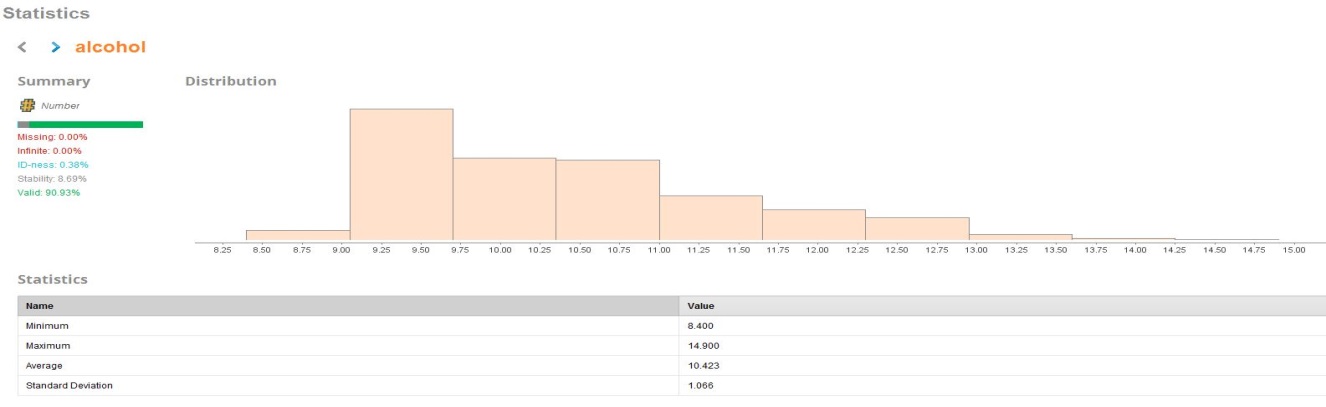
All the values in the dataset are numeric, which will simply some the data preparation tasks as the data mining process needs numeric data to build a mathematical model.

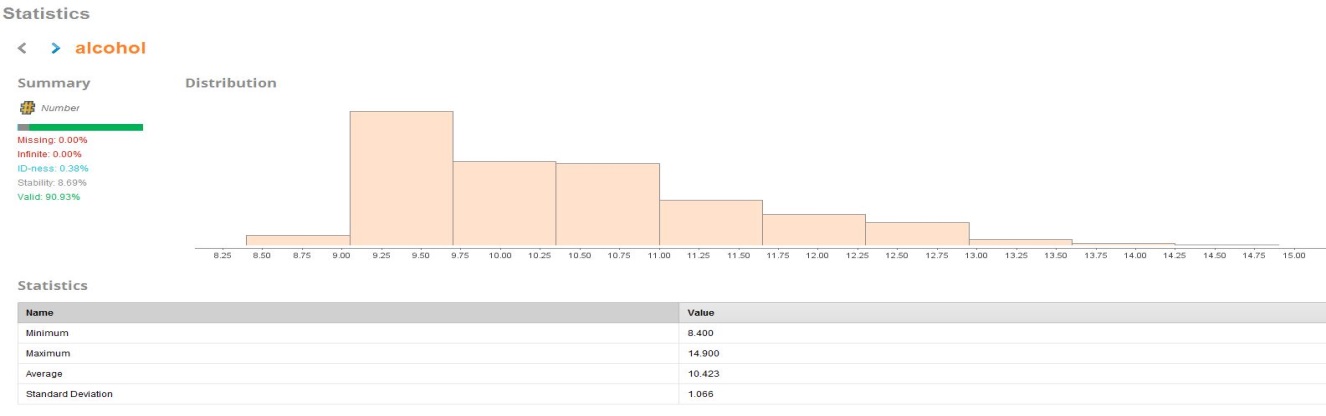
However, although the range of numeric data is not particularly large there are still some features that use a noticeably difference scale, for example the free sulphur dioxide range of values.

In order to prevent such attributes from skewing the resultant data model, all the elements in the feature set will be normalized. This process is described in more detail in Section 4 of this report.

*Quality of Data – No Missing Rows*

The following three screen shots are from the AutoModel output of Rapid Miner. They are a sample of the ‘General’ output from the AutoModel process.





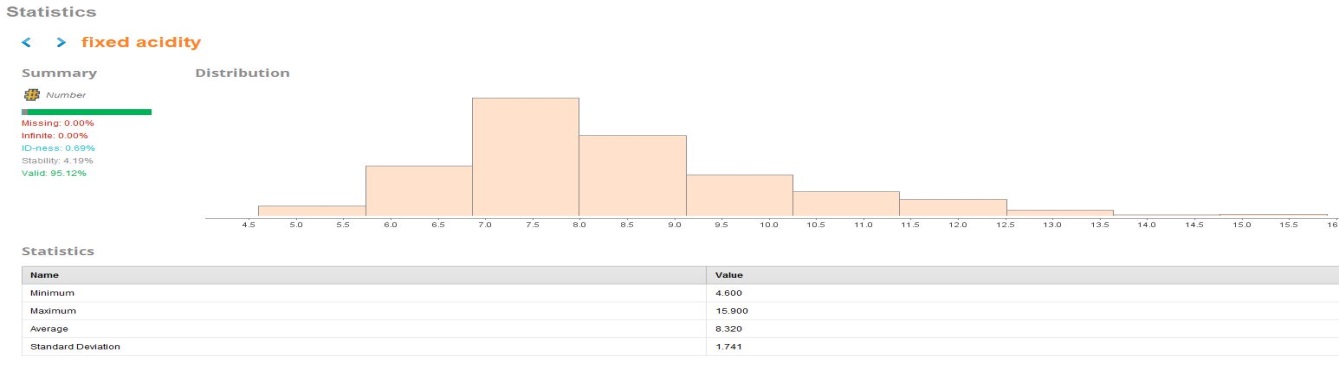


Figure n

The RapidMiner analysis shows that there are no missing data elements in any of the attributes in the dataset.

This will reduce the complexity in the data preparation stage of the project as it will not be necessary to impute missing data, nor will it be necessary to remove incomplete rows from the dataset.

We can see this again in the list of attributes in the ‘Result’s tab for the dataset in RapidMiner.

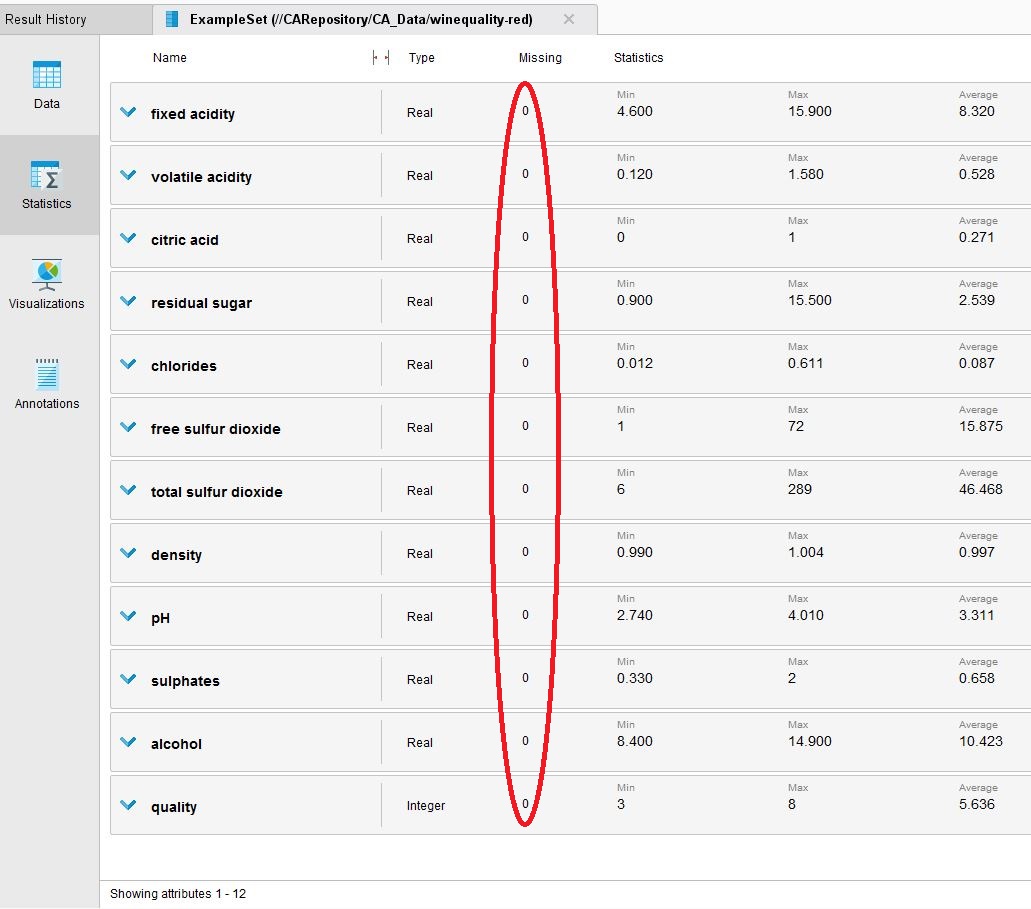


Figure n

*Quality of Data – Duplicate Values*

A quick Python based routine was run independently to confirm that no rows in the Wine Quality dataset are duplicates.

<Python screenshot>

*Quality of Data – Zero Values*

However, although our analysis shows no missing rows or obvious errors in the format of data in the columns, we ran an additional check for ‘zero’ values.

A zero entry, particularly when there is a significant proportion of information in a dataset that is numeric format, can also be an indication of missing data.

A quick EXCEL analysis of the Wine Quality csv file will show that there are **132** rows for the ‘citric acid’ attribute that have a zero value.

A quick validation in our supplementary Python program will show a similar analysis.

<Python screenshot>

Is this a legitimate data entry, or are these data rows incomplete?

To investigate further we referenced supporting material on Kaggle, and elsewhere, that explains the fermentation process for wine in more detail. Citric acid is typically only found in small quantities in wine, and it is used to add what is described as ‘freshness’ to the wine by enhancing flavour.

It is not unusual that the citric acid component in a given type of wine to be completely consumed during the wine fermentation process, and it is not always added back into the process.

Therefore we have concluded that a ‘zero’ entry for citric acid is a valid data point and will not need to be addressed in the data preparation phase of this project.

*Balancing the Wine Quality Dataset*

There is a concern on the spread of wines in the dataset in terms of quality.

RapidMiner allows for a more details bar chart view on the ‘quality’ data in the Wine Quality dataset.

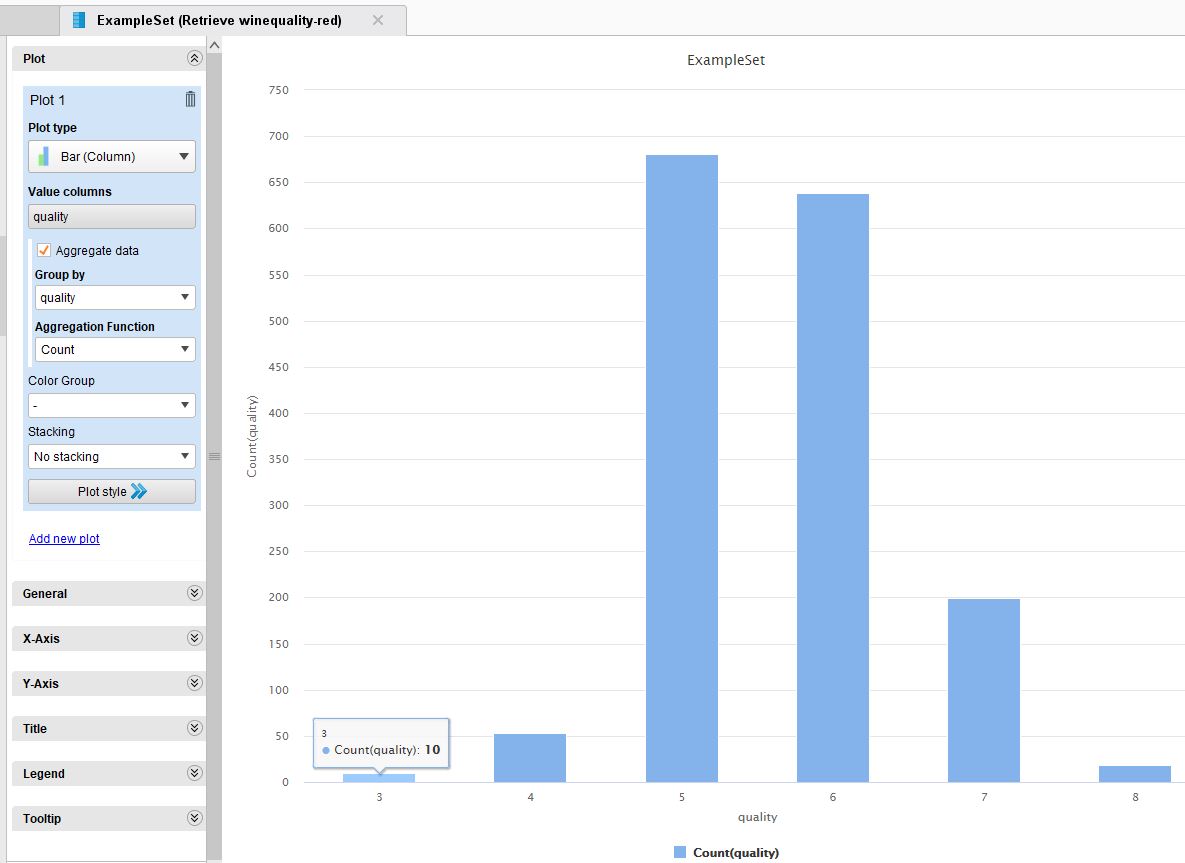


Figure n

There is an immediate concern that the dataset has a high count volume of wines with a ‘quality’ rating of **5** or **6**. However there are relatively few of either the very low or very high quality red wine types.

The Wine Quality dataset is unbalanced and does not have an even spread of wine qualities. This will impact on the accuracy of any model we attempt to build.

Section 4 of this report looks at steps we need to take to balance the Wine Quality dataset before we attempt to train and evaluate any models.

The bar chart above provides a good graphical representation of the spread. For a simpler ‘at-a-glance’ view the Python project output at this phase of data exploration shows the following distribution of ‘quality’ classes.

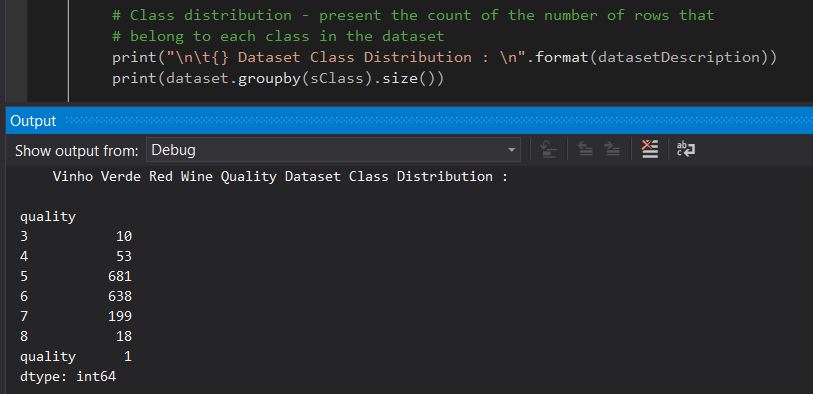


Figure n

## Verify Data Quality

Is the data complete? Does it contain errors and/or missing data? If so, how common are these issues?

**Data Quality Report**

Data quality has been shown to be very good based on the source description and our own data analysis.

There are no missing entries and no obvious errors in the dataset. There are no documented errors in the comments section on the Kaggle page from which the dataset was downloaded and this provides additional assurances.

*Checking for Outliers*

It is impractical to visually assess outliers with the numbers of features in the Wine Quality dataset.

However, RapidMiner provides Operators to detect outliers in datasets.

The following simple RapidMiner process was created to use Local Outlier Factors to generate outlier scores for each row in the Wine Quality dataset.

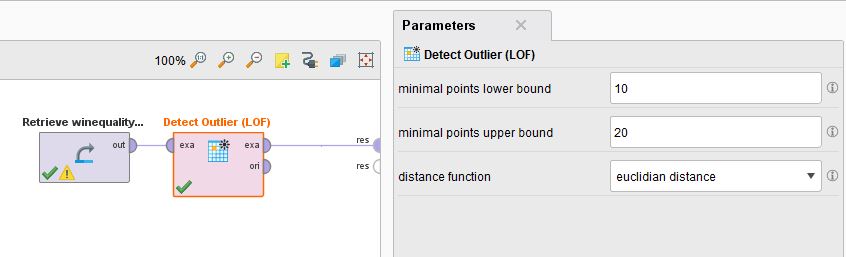


Figure n

The data shows two distinct outliers in the result set.

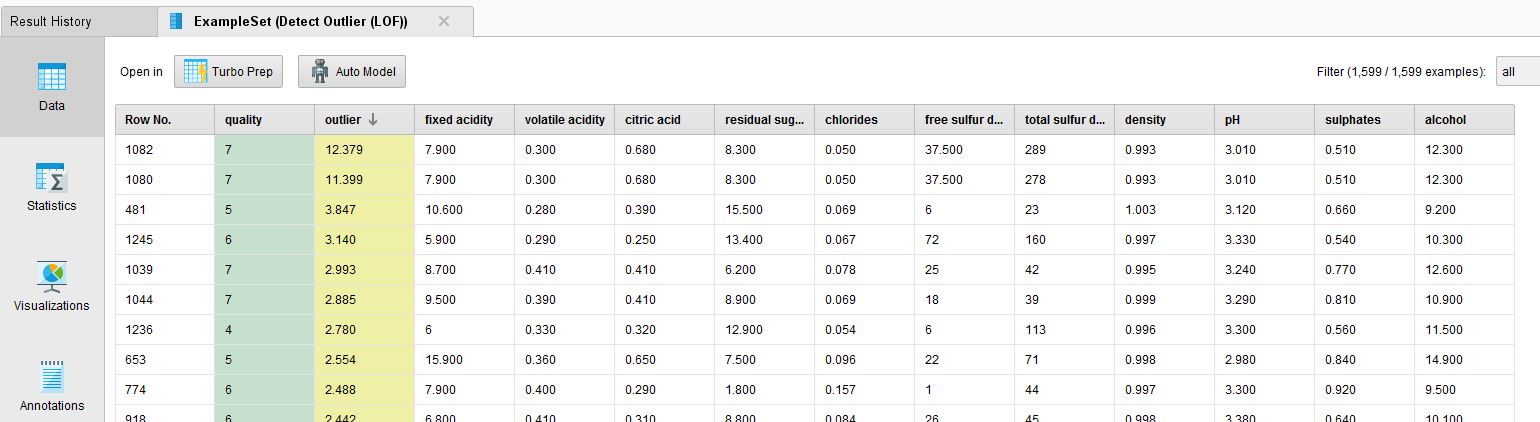


Figure n

A more graphical representation in RapidMiner displays the outliers in the following Scatter Plot;

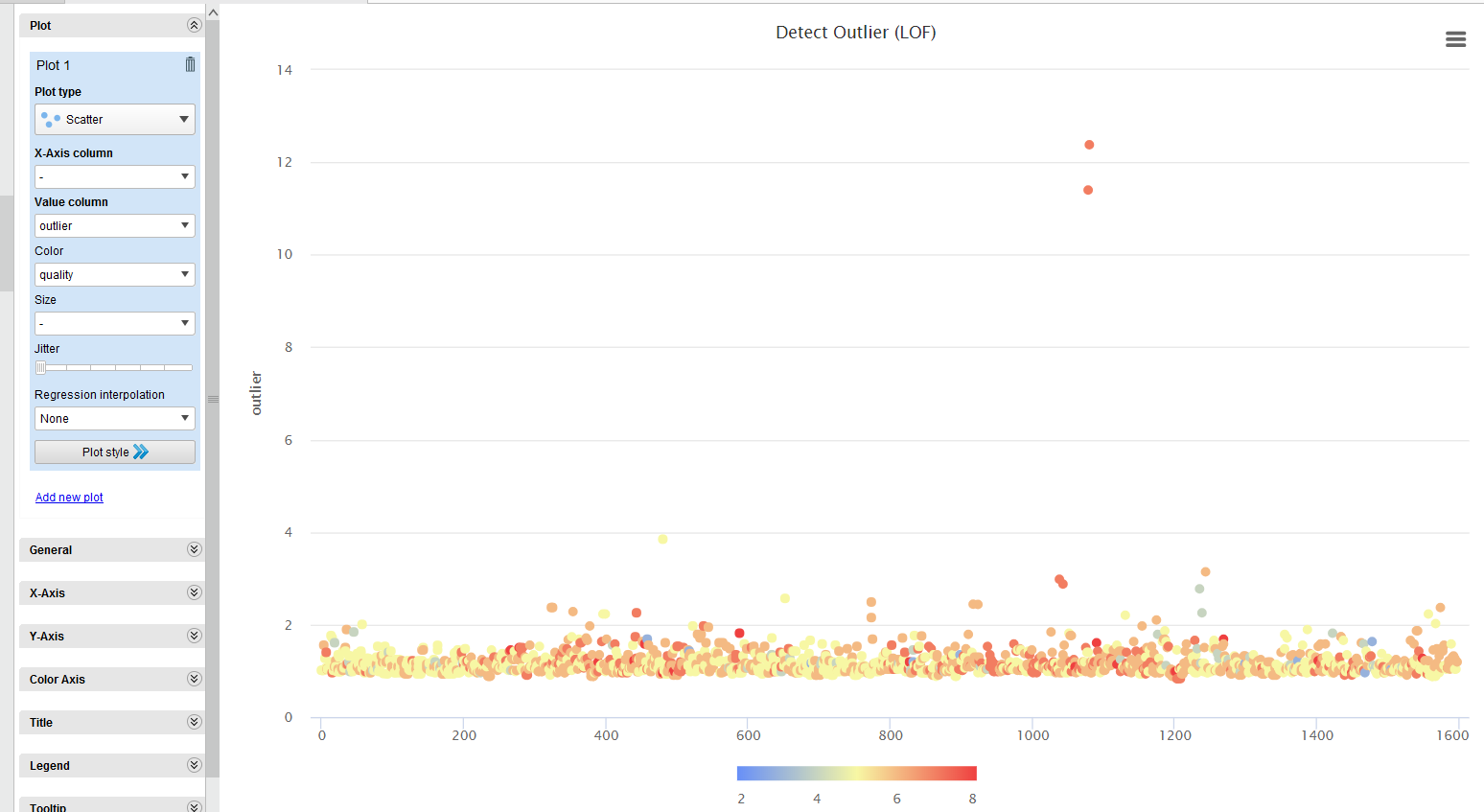


Figure n

A key point to consider is that the ‘outliers’ are marked as ‘high quality’ wines, which can be seen by the colour coding of the entries in the scatter plot above.

As already described, there are very few high quality wines in the dataset hence it is logical that this data would appear as outliers.

Deleting the outliers in this dataset would actually remove key information used in the building of the predictive model. As discussed in Section 4 of this document, it will be necessary to augment the dataset by artificially adding ‘high quality’ data rows to accompany these outliers.

# Data Preparation

The output of this phase of the project is the creation of an adapted dataset, which will be used for modelling and major analysis.

Our Wine Quality dataset will be transformed into a format that allows effective modelling and evaluation.

## Select Data

Our ‘business’ is using this Kaggle Wine Quality dataset to answer the requirement to predict the quality of new red wines delivered to our outlet, based on its chemical composition.

*Data Volumes and Technical Constraints*

The Wine Quality dataset is relatively small, with 1600 rows, so there is no requirement to reduce the number of rows upon which we will build our model in RapidMiner. A dataset this size is not expected to be excessively computationally expensive, even with the more elaborate algorithms we are expected to evaluate.

Our personal laptops should have no practical processing limitation with a dataset of this size and it is therefore not necessary to ‘slice’ or partition the data in any way.

*Quality*

There are no quality issues that warrant the removal of any data rows.

*Selection of Attributes*

There are twelve columns in the original Wine Quality dataset from Kaggle. As such, feature selection might not appear to be a crucial element in data preparation.

However, many datasets contain attributes that have very limited impact on the final values predicted by the type model for which we are searching. Datasets will also often contain attributes that provide close to identical information or are heavily related to each other, and hence introduce a certain redundancy.

It is good practice to eliminate those features that provide relatively limited value, and this is a principle we applied to our Wine Quality dataset.

AutoModel analysis in RapidMiner provided the guideline to a reduced feature set. Section 3.3 of this document provided screen shots of the correlation values attached to each attribute, both in terms of their influence on the ‘target’ value (quality) and their relationship to each other.

There was a considerable amount of iteration from model selection and evaluation back to feature selection to determine the impact of various attribute lists.

As an example, we looked at the recommended feature set for a ‘RandomForest’ algorithm used on the Wine Quality dataset and the following list was suggested;

<screenshot>

A reduced feature set can remove computational complexity when an algorithm is being applied to build a model. A more relevant set of features can also help with the final accuracy of the model and possibly understanding some of the underlying workings.

For our Wine Quality dataset the following Feature Selection operator in RapidMiner was put in place within the overall Data Preparation Phase processes.

<screenshot>

## Clean Data

In general, this task in the Data Preparation Phase is intended to raise the data quality to the level required by the selected analysis technique.

*Clean Data Set*

As discussed, data quality and ‘cleanliness’ was not an issue with this Kaggle dataset on Wine Quality.

The numerical data in each attribute was in a consistent format and data type.

There were no categorical attributes in the original dataset that could have introduced error or ambiguity into our modelling process.

The data quality within the Wine Quality dataset allowed us to proceed quickly to the data construction task (Section 4.3) within this Data Preparation Phase.

## Construct Data

This task usually includes constructive data preparation operations such as the production of new derived attributes, or entire new records.

At this stage in the process we may also find it necessary to transform values for existing attributes.

In the sub-sections below, we describe the various tasks we apply to our Wine Quality dataset as part of the re-construction of the data prior to the commencement of the Modelling Phase.

*Feature Generation*

A business objective of this project was to simplify the output of the predictive model so that employees could quickly assess the quality rating of a new wine in store.

We felt it would be sensible to supplement the 1 – 10 ‘quality’ score with the following categorical descriptions;

* Poor Quality
* Medium Quality
* High Quality

Thus we added an operator in the RapidMiner Data Preparation Phase processes to generate a new attribute called ‘Quality Score’. Our intention was to allow for more meaningful categorization of the data.

The RapidMiner Operator uses the following logic in an attribute generation process;

Explain data range – <show screen shot>

Aside from making the end production more straightforward to use, we have changed our modelling challenge into a Classification problem. We felt that this would generate a more useful and readable set of analysis on each algorithm in our chosen models.

Linear Regression would generate a predicted real number in the range 1 – 10 but the output on accuracy would not be as readily understandable. In addition, our business objective is to quickly provide a meaningful rating of new wine so it made sense to focus our modelling effort on a Classification challenge.

The new ‘Score Quality’ attribute is derived from the ‘Quality’ value and appended as a new quality to the Wine Quality dataset by the RapidMiner operator.

*Normalisation*

Additional good practice in data mining is to normalise the attribute values to prevent bias with larger numerical values.

In our Wine Quality dataset the range of values for *free sulphur dioxide* is 1 to 72 units, and for *total sulphur dioxide* is 6 to 289 units. This is a larger set of absolute numerical values that the other attributes in the dataset.

To avoid values in those features introducing bias we employee a Normalize operator in our RapidMiner Data Preparation Phase process.

<screenshot>

This task ‘smooths’ out the numerical values in the dataset but preserves the relative differences between the features. It is a purely syntactic change to satisfy the requirements for many modelling algorithms to deliver as accurate a result as possible.

A before and after view of a section of the Wine Quality dataset, as represented in the images below, will display the nature of the data transformation;

<screenshot 1>

<screenshot 2>

*Data Balancing*

This was one of the most significant tasks within the Data Preparation Phase, and across this entire data mining project.

It was necessary to iterate backwards and forwards through the phases on Data Preparation, Modelling, and Evaluation in order to find a balance to the Wine Quality dataset that produced the best results. (In practice, we ended up settling on the ‘least bad’ approach).

This sub-section largely described the end-point at which we arrived in terms of balancing the Wine Quality dataset, but it illustrates the type of operations we executed on the data for this task.

The final observation in Section 3.3 of the document (‘Explore Data’) related to the imbalance of wine types in the dataset. There are considerably more ‘5’ and ‘6’ quality wines that wine at either end of the quality spectrum.

Section 4.3 describes how we added an additional attribute to create a classification attribute ‘Quality\_Score’ which groups the 1 – 10 values under three different quality descriptions (‘Poor’, ’Medium’, ‘High’). This simplifies the understanding of the classification of the data but does not help with the balance of the data. There are still 5 to 6 times more wines described as ‘Medium’ than either grouping for ‘Poor’ or ‘High’.

<screenshot>

Working with this Wine Quality dataset is likely to create a predictive model that is biased towards classifying wine as ‘Medium’. Machine learning algorithms have trouble learning when one class dominates the other, or others as in the case of our Wine Quality dataset. There is not enough data to properly model the characteristics of a new ‘Poor’ or ‘High’ quality wine.

What were the potential solutions for the project? The options we considered were;

* Create new ‘synthetic’ row for the ‘Poor’ and ‘High’ quality wines.
* Reduced the number of ‘Medium’ data rows to be introduced into the Modelling Phase.

Upsampling

Our prior data modelling experience suggested the use of a ‘SMOTE’ (**S**ynthetic **M**inority **O**ver-sampling **Te**chnique) approach to generate new rows in the Wine Quality dataset.

RapidMiner, through an additional extension, provides an operator which can be applied to a dataset to create new ‘artificial’ rows of data, which attempt to mirror the required classification of data.

<screen shot of SMOTE Operator>

Downsampling

Balancing the Wine Quality dataset can also involve reducing the number of row for ‘Medium’ quality wines.

A nested set of operations can be embedded into a RapidMiner process to rebalance the data and down sample the ‘Medium' wines.

We created a high level ‘Unbalance’ process operator in RapidMiner into which to feed out Wine Quality dataset.

<Screen shot of ‘Balance’ operator>

Clicking through to the workings of the operator we can see how the Wine Quality dataset is split to separate out the ‘Medium’ Quality wines and reduce the % of those rows fed into the training process for the model.

<screenshot>

<screenshot>

The Upsampling and Downsampling operators are chained within a separate process so that can be more easily re-used with the RapidMiner modelling processes.

<screenshot of chained operators>

Why two SMOTE operators?

Each SMOTE operator in RapidMiner tackles the minority class, which is the classification with the fewest number of rows.

In our Wine Quality dataset the first SMOTE operator tackles the ‘Poor’ quality subset of data and generates new rows for that classification.

The second SMOTE operator follows in sequence. As the minority class is now ‘High’ quality wines, the operator generates new artificial data for that classification.

Using the SMOTE operators in sequence allows for balancing process to take place on both of the ‘minor’ classifications.

How much artificial data to create?

The best data to use is information collected in the real world. However, our Wine Quality dataset is deficient in this area, with ‘Medium’ quality data rows very clearly dominant.

The default setting in the SMOTE operator in RapidMiner is to generate enough data to balance the minority class with the majority one.

<screenshot of SMOTE setting – default balance>

Applying this sequence in parallel would generate many times more artificial data in the Training Set than ‘real data’, and would not be desirable.

This setting was tested with our Wine Quality dataset but the resultant models performed poorly on Test data.

We looked at the following numerical balance in the data and determined what we felt would be an acceptable level of new ‘synthetic’ data.

The Wine Quality dataset was split into Training and Test date prior to modelling with a 70%/30% ratio. The numbers of rows in each set breakdown as follows;

|  |  |  |
| --- | --- | --- |
| **Overall** | **1599** | **Split** |
|  |  |  |
| Training | 1119 | 0.7 |
|  |  |  |
| Test | 480 | 0.3 |

The distribution of rows, based on quality, within the Training and Test data is;

|  |  |  |  |
| --- | --- | --- | --- |
| *Wine Quality* | **Training (No SMOTE)** |  | **Test** |
|  |  |  |  |
| Poor Quality | 41 |  | 22 |
|  |  |  |  |
| Medium Quality | 924 |  | 395 |
|  |  |  |  |
| High Quality | 154 |  | 63 |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  | **1119** |  | **480** |

The SMOTE Operator allows the user to select an upper maximum number of new artificial rows.

We experimented with the following options;

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *Wine Quality* | **UpSample Opt - 1** | | **UpSample Opt - 2** | | **UpSample Opt - 3** | |
|  |  |  | |  | |
| Poor Quality | 300 | 200 | | 100 | |
|  |  |  | |  | |
| Medium Quality |  |  | |  | |
|  |  |  | |  | |
| High Quality | 300 | 200 | | 100 | |
|  |  |  | |  | |
|  | *Percentage Upsampling in Training Set* | | |  | |
|  | 54% | 36% | | 18% | |

The above numbers – 100, 200, and 300 – represent different settings in the SMOTE operator configuration. They represent an upper limit to the additional number of artificial data rows.

<screenshot of SMOTE operator with row numbers>

The table above shows the effect that the configuration values have on the actual percentage of artificial records in the Training Set, before it is used in the Modelling Phase.

The result in terms of actual new ‘synthetic’ rows is given in the table below.

|  |  |  |  |
| --- | --- | --- | --- |
| *Wine Quality* | **Training (with SMOTE) Opt 1** | **Training (with SMOTE) Opt 2** | **Training (with SMOTE) Opt 3** |
|  | *+300 minority rows* | *+200 minority rows* | *+200 minority rows* |
| Poor Quality | 341 | 241 | 241 |
|  |  |  |  |
| Medium Quality | 924 | 924 | 924 |
|  |  |  |  |
| High Quality | 454 | 354 | 154 |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
| Total | **1719** | **1519** | **1319** |

Our initial assumption was that 50%+ of artificial data in the Training Set was too high a value, and that 18% of an increase still generated too few ‘Poor’ and ’High’ quality data rows for the Modelling.

We choose to set the SMOTE Operator parameter in RapidMiner at ‘200’, generating just over 35% of new records in advance of Modelling.

During the actual project we iterated a number of times with the value in the SMOTE operators and the ‘200’ value performed the best on the Test set. (However, as stated earlier in this report this was really just the ‘least worst’ setting in terms of generating predictive values for ‘Poor’ and ‘High’ quality wines).

How much to remove?

The scale of Downsampling was much more of a trial and error process as we iterated through the Data Preparation, Modelling, and Evaluation phases to tune the models to try and maximise accuracy.

In practice we found that Downsampling the ‘Medium’ wines in the Training set to 40% of the original number produced marginally better accuracy in the Modelling / Evaluation phase.

<screenshot of Filter operator>

The Upsampling and Downsampling routines are built and refined as part of the Data Preparation Phase but implemented during the training of the Model.

It is important to note that the SMOTE and down sampling operations are **only** applied to the Training dataset for Wine Quality. This is the subset of the overall Wine Quality dataset used to train the actual model.

The application of these over and undersampling routines within the Modelling Phase is described in detail in Section 5.2 and 5.3 of this document.

## Integrate Data

Our Wine Quality dataset from Kaggle is a complete repository of information for our data mining purposes.

We determined that there is no need to merge additional data sources, although we do carry out a supplementary ‘White Wine’ analysis after the main modelling, evaluation and deployment.

See Appendix 9.2 for details.

## Format Data

Formatting transformations in this task primarily refer to syntactic changes made to the data that do not change its meaning, but may be required by the modelling tool and/or choice of modelling algorithm.

The major tasks in reworking the Wine Quality dataset are described in Section 4.3, but there are some other minor updates that we make to the data

*Data Ordering*

The Kaggle page describes the Wine Quality dataset as ‘ordered’.

In data mining it is often important to change the order of the records in the dataset. Many modelling algorithms will need datasets to be in a fairly random order. For example, when using neural networks it is general best for the records to be presented in a random order.

We do not employee the use of neural networks in this project but, following good practice, we deployed a Shuffle operator in our RapidMiner Data Preparation Phase processes.

<Screenshot>

By shuffling the data, the intention is to remove potential bias that could be introduced by the sequence with which the rows in our Wine Quality dataset were added.

# Modelling

## Select Modelling Technique

The first step in the Modelling Phase is to select the actual modelling technique to be used.

This is a Classification problem so there are many academic guidelines. We also choose to focus on using RapidMiner exclusively as our tool of choice from this point on in the project and this influenced our approach to the Modelling Phase.

The RapidMiner website provides resources (<https://mod.rapidminer.com/>) to suggest modelling techniques based on the characteristics of a dataset. Based on the nature of the Wine Quality dataset this web based RapidMiner application suggested the following models (this is a partial screenshot with the highest rated models on the right in **red**):

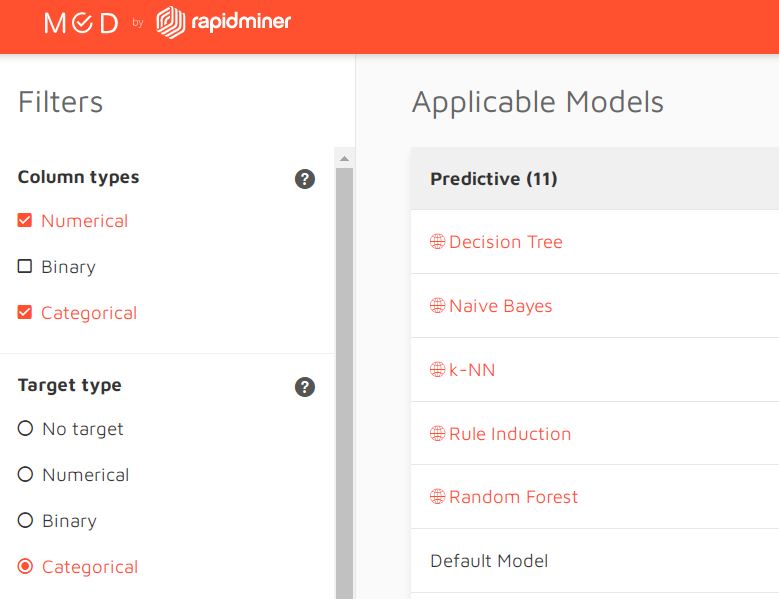


Figure n

We also used the AutoModel function in RapidMiner on our prepared Training Set data to look at the recommended selection of models, their relative accuracy, along with other information such as recommended feature sets, and correlations.

<Screenshot of AutoModel on Training Data>

The AutoModel output re-confirmed that we would focus our Modelling Phase to look at the recommended list of five algorithms, as listed in the RapidMiner website screenshot above. K-NN and Rule Induction did not feature in the AutoModel output but we choose to go with the recommendations from the RapidMiner web application.

The Wine data quality has already been determined to be very good, based on our analysis and work described in Section 3 and Section 4 of this document. Therefore it was expected that our modelling techniques with the above algorithms would not encounter any issues with the data structure or format.

## 

## Generate Test Design

Before building an actual model, it is general practice in this task within the Modelling Phase to generate a procedure to test the model’s quality and validity.

In a supervised data mining task, such as the classification of wine quality in this project, it would be common to use error rates as quality measures for the models.

*Training and Test Sets*

We therefore separated our Wine Quality dataset into Training and Test sets.

The Training Set is used to build the model, and the quality of the model is subsequently estimated using the separate Test Set.

We build a Data Preparation Process in RapidMiner to prepare the Wine Quality csv data for use in Modelling. The final steps in that process were to split the overall dataset in a 70/30 ratio into the Training and Testing sets respectively.

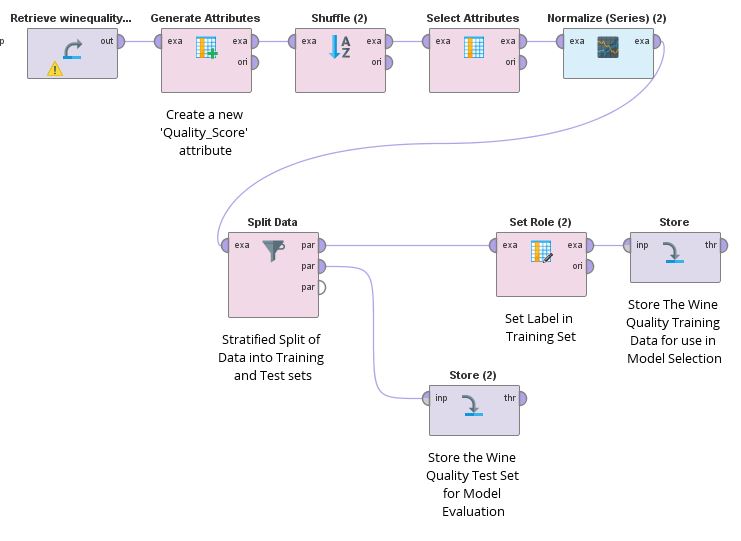


Figure n

Storing the Training and Test Sets at the end of this process allows us to investigate the data stores independently before we enter the Modelling Phase.

The Modelling Phase process in RapidMiner begins with the Training and Test Sets as the initial inputs.

## Build Model

The purpose of this task is to run the modelling tool (RapidMiner) on the prepared dataset (Wine Quality) to create one or more models.

In our project, we choose to focus initially on five models and then focused on the one that was more consistently generating more accurate results with the Test Set data.

*Cross Validation*

To recap, for training our models we initially split the model into two sections which are ‘**Training data**’ and ‘**Testing data**’.

In our RapidMiner modelling process we incorporated a Cross Validation operator, which dynamically splits the Training Set into **Validation Sets**.

Our modelling technique is to train the classifier using ‘**training data set**’, tune the parameters using ‘**validation set**’ and then test the performance of your classifier on unseen ‘**test data set**’. An important point to note is that during training the classifier only the training and/or validation set is available. The test data set is not used during the training of our Wine Quality classifier. The test set will only be available during testing the classifier.

To elaborate:

**Training set:** The training set is the material through which the computer learns how to process information. In our Wine Quality Data Preparation Phase we generated this Training Set so that our Machine Learning approach will use algorithms to perform the training part for a predictive model.

**Validation set:** Cross-validation is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data. A set of unseen data is used from the training data to tune the parameters of a classifier.

**Test set:** A set of unseen data used only to assess the performance of a fully-specified classifier. The Test Set for the Wine Quality model building task is one of the outputs from our Data Preparation Phase.

The layout of the RapidMiner process, with Cross Validation can be see in this image.

<screenshot>

Drilling into the Cross Validation operator we can see the following set up in RapidMiner.

<screenshot>

The left-hand side of the Cross Validation process runs multiple iterations using different slices of ‘Training’ and ‘Validation’ data.

Our Cross Validation operator uses a 10-fold parameter. For simplicity, we have included below a diagram that represents the same process but for a 5-Fold Cross Validation split.

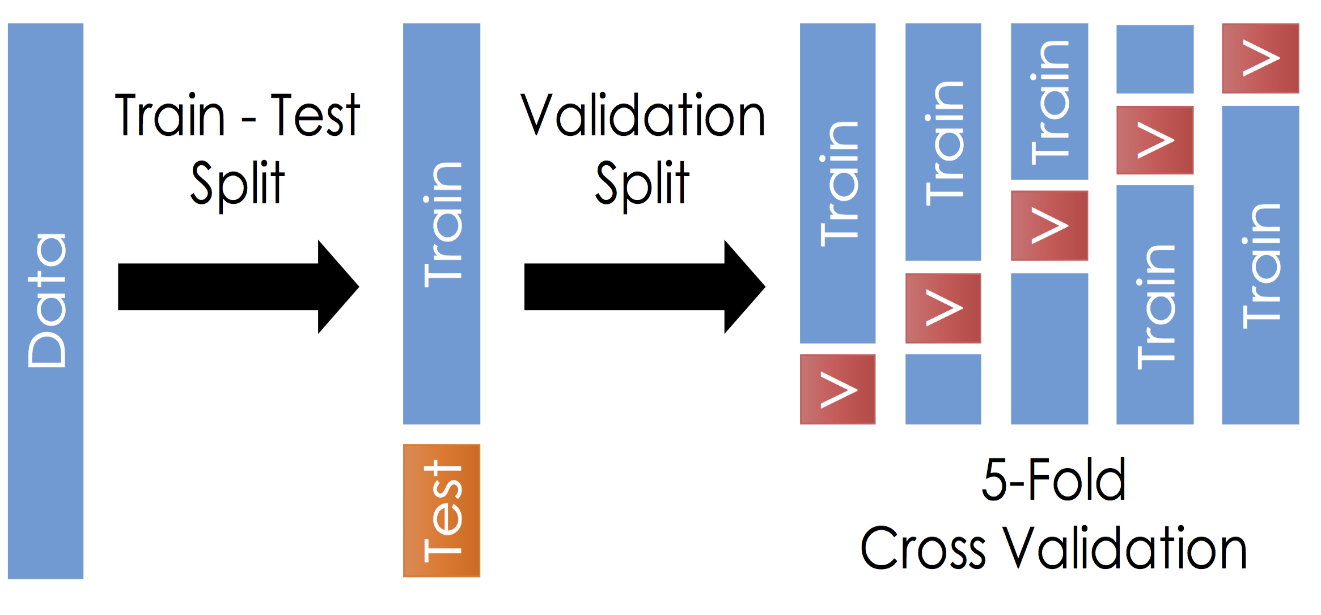


Figure n

In a data set, a training set is implemented to build up a model, while a test (or validation) set is to validate the model built. Data points in the training set are excluded from the test (validation) set. Usually, a data set is divided into a training set, a validation set (some people use ‘test set’ instead) in each iteration, or divided into a training set, a validation set and a test set in each iteration.

.

Once the Wine Quality data is divided into the 3 given segments in our RapidMiner processes, we could commence the training process.

RapidMiner provides operators to ensure the model is applied to the Training/Validation data using the Apply Model Operator. The Performance operator is used to evaluate how accurate the aggregate of the various validations were across the data ‘splits’.

<screenshot with just right hand side>

The output of the Performance operator in the above screenshots is related to the performance of the model on the Training data. Or primary interest is in the performance on the ‘unseen’ Wine Quality Test data, which is discussed later in this section of the document.

*Embedded Over and UnderSampling*

The..

*Models*

The..

*Tracking Performance – Training and Test Sets*

The RapidMiner operator..

*Parameter Settings*

The..

## Assess Model

In general, the data mining engineer will interpret the models according to his domain knowledge, the data mining success criteria, and the desired test design.

*Business Domain Criteria*

In our business domain for this project, we want a predictive model that will help employees provide a high level recommendation on wine quality based on its known chemical composition (as opposed to actually sampling every new type of red wine purchased from the wholesaler!)

The model assessment task in the Modelling Phase…

*Ranking Models with Confusion Matrices*

We attempted to rank the models based on the outputs of the RapidMiner Performance operator when the model is applied to the Test Set.

*Model Assessments*

Models are assessed based on evaluation criteria, looking for general accuracy but also taking business success criteria into account.

*The Resultant Model of Choice for Wine Quality*

The..

*Revised Parameter Settings*

The..

# Evaluation

## Evaluate Results

Previous evaluation steps dealt with factors such as the accuracy and generality of the model. This task in the Evaluation Phase assesses the degree to which the model meets the business objectives and determines if there is some business reason why this model is deficient.

*Assessment With Respect To Business Success Criteria*

At best, we derived a model that will be success enough for our business purposes to correctly predict a ‘Medium’ wine in most circumstances.

The..

## Review Process

Have we overlooked any factors in our data mining process? Can we determine any flaw in our modelling approach?

*Data Structure*

The..

## Determine Next Steps

For this task in the Evaluation Phase the project needs to decide how to proceed.

*Determining Next Steps*

In the ‘real world’, our wine Quality predictive model would probably have too many gaps and deficiencies to actually roll out into a production setting.

However, for the purposes of the CA, and for academic curiosity at least, we will move the Deployment Phase.

# Deployment

## Plan Deployment

This task takes the evaluation results and determines a strategy for deployment.

*Deployment Strategy*

Our analysis of..

* Step 1 : ..
* Step 2 : ..
* Step 3 : ..

## Plan Monitoring and Maintenance

Monitoring and maintenance are important issues if the data mining result becomes part of the day-to-day business and its environment.

*Random Manual Sampling to Verify Predicted Results*

The..

*Obtain More Data in Future and Re-Model*

The..

## Produce Final Report

At the end of the project, he project team writes up a final report.

This is that report.

## Review Project

This is the task within the Evaluation Phase when assessments are made about what went right and what needs to be improved.

<Look at slide deck>

# Conclusion

## Conclusion..

# Appendices and References

## Appendix A Understanding Wine and Types

Wine is an alcoholic beverage made from grapes which is fermented without the addition of sugars, acids, enzymes, water, or other nutrients

Red wine is made from dark red and black grapes. The colour usually ranges from various shades of red, brown and violet. This is produced with whole grapes including the skin which adds to the colour and flavour of red wines, giving it a rich flavour.

White wine is made from white grapes with no skins or seeds. The colour is usually straw-yellow, yellow-green, or yellow-gold. Most white wines have a light and fruity flavour as compared to richer red wines.

**Understanding Wine Attributes and Properties**

* **Fixed acidity:** Acids are one of the fundamental properties of wine and contribute greatly to the taste of the wine. Reducing acids significantly might lead to wines tasting flat. Fixed acids include tartaric, malic, citric, and succinic acids which are found in grapes (except succinic).
* **Volatile acidity:** These acids are to be distilled out from the wine before completing the production process. It is primarily constituted of acetic acid though other acids like lactic, formic and butyric acids might also be present. Excess of volatile acids are undesirable and lead to unpleasant flavour. In the US, the legal limits of volatile acidity are 1.2 g/L for red table wine and 1.1 g/L for white table wine.
* **Citric acid:** This is one of the fixed acids which give a wine its freshness. Usually most of it is consumed during the fermentation process and sometimes it is added separately to give the wine more freshness.
* **Residual sugar:** This typically refers to the natural sugar from grapes which remains after the fermentation process stops, or is stopped.
* **Chlorides:** This is usually a major contributor to saltiness in wine.
* **Free sulphur dioxide:** This is the part of the sulphur dioxide that when added to a wine is said to be free after the remaining part binds. Winemakers will always try to get the highest proportion of free sulphur to bind. They are also known as sulphites and too much of it are undesirable and give a pungent odour.
* **Total sulphur dioxide:** This is the sum total of the bound and the free sulphur dioxide. This is mainly added to kill harmful bacteria and preserve quality and freshness. There are usually legal limits for sulphur levels in wines and excess of it can even kill good yeast and give out undesirable odour.
* **Density:** This can be represented as a comparison of the weight of a specific volume of wine to an equivalent volume of water. It is generally used as a measure of the conversion of sugar to alcohol.
* **pH:** Also known as the potential of hydrogen, this is a numeric scale to specify the acidity or basicity the wine. Fixed acidity contributes the most towards the pH of wines. You might know, solutions with a pH less than 7 are acidic, while solutions with a pH greater than 7 are basic. With a pH of 7, pure water is neutral. Most wines have a pH between 2.9 and 3.9 and are therefore acidic.
* **Sulphates:** These are mineral salts containing sulphur. Sulphates are to wine as gluten is to food. They are a regular part of the winemaking around the world and are considered essential. They are connected to the fermentation process and affects the wine aroma and flavour.
* **Alcohol:** Wine is an alcoholic beverage. Alcohol is formed as a result of yeast converting sugar during the fermentation process. The percentage of alcohol can vary from wine to wine. Hence it is not a surprise for this attribute to be a part of this dataset. It's usually measured in % vol or alcohol by volume (ABV).
* **Quality:** Wine experts graded the wine quality between 0 (very bad) and 10 (very excellent). The eventual quality score is the median of at least three evaluations made by the same wine experts.
* **wine\_type:** Since we originally had two datasets for red and white wine, we introduced this attribute in the final merged dataset which indicates the type of wine for each data point. A wine can either be a 'red' or a 'white' wine. One of the predictive models we will build in this chapter would be such that we can predict the type of wine by looking at other wine attributes.
* **quality\_label:** This is a derived attribute from the quality attribute. We bucket or group wine quality scores into three qualitative buckets namely low, medium and high. Wines with a quality score of 3, 4 & 5 are low quality, scores of 6 & 7 are medium quality and scores of 8 & 9 are high quality wines. We will also build another model in this chapter to predict this wine quality label based on other wine attributes.

## The White Wine Dataset

The ..

## References

The use of this Wine Quality dataset in this CA acknowledges the source publication:

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modelling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.