

**Postgraduate Certificate in Software Design with Artificial Intelligence**

**Data Mining and Machine Learning**

**Assignment 1**

Student ID: A00267948

Student Name: Daniel Foth

*Brief Description: This paper evaluates different data sets using regression, decision trees and kNN algorithms. Models are created and predictions are made using the test data from the data sets.*

Git: <https://github.com/DanielsHappyWorks/DM-ML-Module-Assignment>

Contents

[0. Introduction 3](#_Toc27426844)

[1. Regression 3](#_Toc27426845)

[1.1. Overview of the Problem 3](#_Toc27426846)

[1.2. Data Exploration (tables and graphs) 3](#_Toc27426847)

[1.3. Definition of Training and Testing Set 5](#_Toc27426848)

[1.4. Model Generation and Information 6](#_Toc27426849)

[1.5. Predictions for the test data 7](#_Toc27426850)

[1.6. Evaluation of the model(s) and conclusion. 8](#_Toc27426851)

[2. Decision Trees 8](#_Toc27426852)

[2.1. Overview of the Problem 8](#_Toc27426853)

[2.2. Data Exploration (tables and graphs) 8](#_Toc27426854)

[2.3. Definition of Training and Testing Set 8](#_Toc27426855)

[2.4. Model Generation and Information 8](#_Toc27426856)

[2.5. Predictions for the test data 8](#_Toc27426857)

[2.6. Evaluation of the model(s) and conclusion. 8](#_Toc27426858)

[3. kNN 8](#_Toc27426859)

[3.1. Overview of the Problem 8](#_Toc27426860)

[3.2. Data Exploration (tables and graphs) 8](#_Toc27426861)

[3.3. Definition of Training and Testing Set 8](#_Toc27426862)

[3.4. Model Generation and Information 8](#_Toc27426863)

[3.5. Predictions for the test data 8](#_Toc27426864)

[3.6. Evaluation of the model(s) and conclusion. 8](#_Toc27426865)

[4. Citations 8](#_Toc27426866)

# Introduction

# Regression

## Overview of the Problem

For regression, a data set that describes characteristics of wine will be used. The data is called Wine Quality and was originally sourced by Paulo Cortez, Antonio Cerdeira, Fernando Almeida, Telmo Matos and Jose Reis.

The data contains two sets:

1. red wine with 1599 rows
2. white wine with 4898 rows

The data set has 12 features which include:

1. fixed acidity
2. volatile acidity
3. citric acid
4. residual sugar
5. chlorides
6. free sulfur dioxide
7. total sulfur dioxide
8. density
9. pH
10. sulphates
11. alcohol
12. quality (score between 0 and 10)

For this specific problem only the white wine data set will be used. Using regression algorithms, all the features will be analysed and correlated together to try and predict the quality of the wine.

## Data Exploration (tables and graphs)

The data has no missing features which was double checked when the csv was loaded into R. No clean up was needed to handle missing fields

To see how all the attributes effect quality, they were plotted against each other and exported.

Plots:

|  |  |  |
| --- | --- | --- |
| Feature | Scatter | Histogram |
| fixed acidity |  |  |
| volatile acidity |  |  |
| citric acid |  |  |
| residual sugar |  |  |
| chlorides |  |  |
| free sulfur dioxide |  |  |
| total sulfur dioxide |  |  |
| density |  |  |
| pH |  |  |
| sulphates |  |  |
| alcohol |  |  |

Larger versions of the plots can be seen within the R project. Please note that for some graphs the Linear Regression line doesn’t show up on some of the plots which is probably due to linear regression not being the best fit for the data.

The quality was also used against all features separately to get a linear regression models so they could be compared. The performance is described in section 1.4 Model Generation and Information

## Definition of Training and Testing Set

The entire data set with 12 features and 4898 was used for Training models.

14 rows were chosen randomly by hand so they can be used to validate how well the models would predict quality. The validation sample has two of each type of quality to be tested against.

Validation Sample:

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ID | quality | volatile acidity | citric acid | residual sugar | chlorides | free sulfur dioxide | total sulfur dioxide | density | pH | sulphates | alcohol | fixed acidity |
| 1 | 3 | 0.26 | 0.21 | 16.2 | 0.074 | 41 | 197 | 0.998 | 3.02 | 0.5 | 9.8 | 8.5 |
| 2 | 3 | 0.17 | 0.47 | 1.4 | 0.037 | 5 | 33 | 0.9939 | 2.89 | 0.28 | 9.6 | 10.3 |
| 3 | 4 | 0.485 | 0 | 1.5 | 0.065 | 8 | 103 | 0.994 | 3.63 | 0.4 | 9.7 | 5.5 |
| 4 | 4 | 0.31 | 0.31 | 9.9 | 0.04 | 10 | 175 | 0.9953 | 3.46 | 0.55 | 11.4 | 6.7 |
| 5 | 5 | 0.36 | 0.04 | 5.7 | 0.046 | 21 | 87 | 0.9934 | 3.22 | 0.51 | 10.2 | 5.9 |
| 6 | 5 | 0.37 | 0.51 | 11.8 | 0.044 | 62 | 163 | 0.9976 | 3.19 | 0.44 | 8.8 | 6.8 |
| 7 | 6 | 0.34 | 0.39 | 7.6 | 0.04 | 45 | 215 | 0.9965 | 3.11 | 0.53 | 9.2 | 7.6 |
| 8 | 6 | 0.3 | 0.27 | 11.6 | 0.028 | 22 | 97 | 0.99314 | 2.96 | 0.38 | 11.7 | 6.8 |
| 9 | 7 | 0.23 | 0.39 | 2.3 | 0.033 | 29 | 102 | 0.9908 | 3.26 | 0.54 | 12.3 | 7.2 |
| 10 | 7 | 0.41 | 0.37 | 4.5 | 0.03 | 40 | 114 | 0.992 | 3.17 | 0.54 | 12.4 | 7.9 |
| 11 | 8 | 0.19 | 0.27 | 13.9 | 0.057 | 45 | 155 | 0.99807 | 2.94 | 0.41 | 8.8 | 7.3 |
| 12 | 8 | 0.28 | 0.34 | 2.2 | 0.037 | 24 | 125 | 0.98986 | 3.36 | 0.33 | 12.8 | 5.8 |
| 13 | 9 | 0.27 | 0.45 | 10.6 | 0.035 | 28 | 124 | 0.997 | 3.2 | 0.46 | 10.4 | 9.1 |
| 14 | 9 | 0.36 | 0.34 | 4.2 | 0.018 | 57 | 119 | 0.9898 | 3.28 | 0.36 | 12.7 | 6.9 |

## Model Generation and Information

There were 14 models created. 11 with only one feature each. 3 with all the features of which two were polynomial regression models.

Single Feature models:

|  |  |
| --- | --- |
| Model | R-squared |
| Model 1: quality ~ fixed.acidity | Multiple R-squared: 0.01292  Adjusted R-squared: 0.01272 |
| Model 2: quality ~ volatile.acidity | Multiple R-squared: 0.03792  Adjusted R-squared: 0.03772 |
| Model 3: quality ~ citric.acid | Multiple R-squared: 8.481e-05  Adjusted R-squared: -0.0001194 |
| Model 4: quality ~ residual.sugar | Multiple R-squared: 0.009521  Adjusted R-squared: 0.009319 |
| Model 5: quality ~ chlorides | Multiple R-squared: 0.04407  Adjusted R-squared: 0.04388 |
| Model 6: quality ~ free.sulfur.dioxide | Multiple R-squared: 6.655e-05  Adjusted R-squared: -0.0001377 |
| Model 7: quality ~ total.sulfur.dioxide | Multiple R-squared: 0.03053  Adjusted R-squared: 0.03034 |
| Model 8: quality ~ density | Multiple R-squared: 0.09432  Adjusted R-squared: 0.09414 |
| Model 9: quality ~ pH | Multiple R-squared: 0.009886  R-squared: 0.009684 |
| Model 10: quality ~ sulphates | Multiple R-squared: 0.002881  Adjusted R-squared: 0.002678 |
| Model 11: quality ~ alcohol | Multiple R-squared: 0.1897  Adjusted R-squared: 0.1896 |

Multiple Feature Models:

|  |  |
| --- | --- |
| Model | R-squared |
| All Features | Multiple R-squared: 0.2819  Adjusted R-squared: 0.2803 |
| All Features to degree 2 | Multiple R-squared: 0.3679  Adjusted R-squared: 0.3578 |
| All Features to degree 3 | Multiple R-squared: 0.4612  Adjusted R-squared: 0.4181 |

A few other models with just a select few parameters were tried but they were usually worse than the three listed above.

## Predictions for the test data

For predictions I chose to do them on the models with all features and the model with all features to the 3rd polynomial degree using the 14 rows listed in section 1.3. Anything beyond polynomial degree 3 caused R-Studio to freeze as it required too much memory so testing for overfitting couldn’t be performed.

To this point the models weren’t great based on the r values. So, when the predictions are made, they will be rounded to the nearest whole number to see how accurate they are since regression can predict values between whole numbers.

Output from predictions:

## Evaluation of the model(s) and conclusion.

# Decision Trees

## Overview of the Problem

## Data Exploration (tables and graphs)

## Definition of Training and Testing Set

## Model Generation and Information

## Predictions for the test data

## Evaluation of the model(s) and conclusion.

# kNN

## Overview of the Problem

## Data Exploration (tables and graphs)

## Definition of Training and Testing Set

## Model Generation and Information

## Predictions for the test data

## Evaluation of the model(s) and conclusion.

# Citations

* P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553. ISSN: 0167-9236.