Introduction to Data Science - HW10 - First Steps of the Project

Task 1. Setting up

Project repository: https://github.com/kristiina-h/SAT

Task 2. Business understanding

Project title: Wine quality analysis

Team members: Kaja Jakobson, Kistiina Hakk

Identifying the business goals

Background: In recent history, there has been an increasing focus on analyzing and rating wines. The quality of a wine is essential for the wine industry as well as the consumers. The traditional (expert) way of measuring wine quality is expensive and time-consuming. As wine is a complex product obtained by biological and biochemical transformations by wine microorganisms and during wine aging, machine learning models are important tools to help with these tasks.

Business goals: Help wine businesses improve wine quality by better understanding each attribute contribution and predicting the wines' rate based on the physicochemical properties.

Business success criteria: Results of analysis can be used in everyday business.

Assessing the situation

Inventory of resources

Resources available for the project:

- People: two project team members.
- Data: the dataset contains two files (red wine and white wine variants) of the Portuguese "Vinho Verde" wine.
- Hardware: A traditional laptop is sufficient for this machine learning task.
- Software: The data analysis will be carried out using Python.

Requirements, assumptions, and constraints

Schedule for completion:

- Choose the topic and collect the data - November 8
- A detailed plan of the project - November 29
- Project completed - December 13
- Final presentation – December 16

Risks and contingencies

Legal and security obligations: not relevant in our project as public datasets are used.

Requirements for acceptable finished work:

- data analysis completed and main results formulated;
- project video recorded;
- project poster designed.

Risks and contingencies:

Human resources:

- two team members must on time cover the work of three people on;
- find suitable collaboration opportunities for team members with different schedules.

Terminology:

Features included in the dataset

Alcohol - the amount of alcohol in wine

Chlorides - the amount of salt in the wine

Citric acid - acts as a preservative to increase acidity (small quantities add freshness and flavor to wines)

Density - sweeter wines have a higher density

Fixed acidity - non-volatile acids that do not evaporate readily

Free sulfur dioxide - prevents microbial growth and the oxidation of wine

pH - the level of acidity

Residual sugar: is the amount of sugar remaining after fermentation stops. The key is to have a perfect balance between — sweetness and sourness (wines > 45g/ltrs are sweet)

Sulphates - a wine additive that contributes to SO_2 levels and acts as an antimicrobial and antioxidant

Total sulfur dioxide: is the amount of free + bound forms of SO_2

Volatile acidity - high acetic acid in wine which leads to an unpleasant vinegar taste

Costs and benefits: not relevant in our project

Defining the data-mining goals

Data-mining goals

- Determine which features are the best quality wine indicators.
- Build a machine learning model that predicts the rating of wines.

Data-mining success criteria

- Create a model which helps to improve wine quality by 1-2 points with an accuracy rate 80 – 90%.

Task 3. Data understanding

Data gathering:

For predicting wine quality, it is needed to have datasets that have many different quantative measurements of the wine. One dataset is needed for the white wine and one for the red wine.

The measurements should include assessed wine quality and measurements of many quality affecting properties. For example, acidity, residual sugar, chlorides, sulfur oxide, pH, density, suplhates, alcohol.

Through data search from kaggle, such datasets were found (link: https://www.kaggle.com/vishalyo990/prediction-of-quality-of-wine/data). The data is accessible and usable for the set purposes and the two csv files from the website (winequality-red.csv and winequality-white.csv) will be used for the analysis through machine learning,

Data description 1 (attributes):

Both of the datasets have cleaned data with float and int values for the properties. The properties and their units in the dataset are as follows:

Fixed Acidity (tartaric acid - g / dm^3)

Volatile acidity (acetic acid - g / dm^3)

Citric acid (g / dm^3)

Residual sugar (g / dm^3)

Chlorides (sodium chloride - g / dm^3

Free sulfur dioxide (mg / dm^3)

Total sulfur dioxide (mg / dm^3)

Density (g / cm^3)

PH

Sulphates (potassium sulphate - g / dm3)

Alcohol (% by volume)

Quality (score between 0 and 10)

Data description 2 (meaning of attributes):

Fixed acidity: Most acids involved with wine or fixed or nonvolatile

Volatile acidity: The amount of acetic acid in wine

Citric acid: citric acid can add ‘freshness’ and flavor to wines

Residual sugar: The amount of sugar remaining after fermentation stops

Chlorides: the amount of salt in the wine

Free sulfur dioxide: The free form of SO_2 exists in equilibrium between molecular SO_2 and bisulfite ion; it prevents microbial growth and the oxidation of wine

Total sulfur dioxide: Amount of free and bound forms of S02; at free $SO_2$2 concentrations over 50 ppm, SO_2 becomes evident in the nose and taste of wine

Density: The density of wine

PH: Describes how acidic or basic a substance is on a scale from 0 (very acidic) to 14 (very basic)

Sulphates: a wine additive which can contribute to sulfur dioxide gas (SO_2) levels, wich acts as an antimicrobial and antioxidant

Alcohol: The percent alcohol content of the wine

Quality: Score between 0 and 10

Data describtion 3 (attribute range):

White wine:

Datatset white wine there has 4898 rows × 12 columns

1 - fixed acidity- values between 3.8-14.2

2 - volatile acidity- values between 0.08-1.1

3 - citric acid- values between 0.0-1.66

4 - residual sugar- values between 0.6-65..8

5 - chlorides- values between 0.009000000000000001-0.34600000000000003

6 - free sulfur dioxide-values between 2.0-289.0

7 - total sulfur dioxide- values between 6.0-289.0

8 - density- values between 0.98711-1.03898

9 - pH- data type:values between 2.72-3.82

10 - sulphates- values between 0.22- 1.08

11 - alcohol-values between 8.0- 14.1

12- quality-values between 3-9

Data gathering 4 (attribute range):

Red wine:

Datatset red wine there has 1599 rows × 12 columns

1 - fixed acidity- values between 4.6-15.9

2 - volatile acidity- values between 0.12-1.158

3 - citric acid- values between 0.0-1.0

4 - residual sugar- values between 0.9-15.5

5 - chlorides- values between 0.009-0.611

6 - free sulfur dioxide- values between 1.0-72.0

7 - total sulfur dioxide- values between 9.0-440.0

8 - density- values between 0.9900700000000001 -1.00369

9 - pH- values between 2.74-4.01

10 - sulphates- values between 0.33-2.0

11 - alcohol- values between 8.4-14.9

12- quality- values between 3-8

Data exploration and quality assessment:

I feel that there are close to no problems with the dataset, as everything is cleaned, there are no anomalies and the dataset size is big enough for modelling (1600 and 4900 rows for the datasets). Data quality is thus very good in my opinion and no editing of the datasets is needed. In addition, the datasets can be used for the purposes and come from reliable sources.

Task 4. Planning the project

Tasks	**Team member/Hours**
Choose the topic, collect the data, create a project repository	Kristiina 4h
Plan of the project (tasks 1, 3, 4)	Kristiina 6h
Plan of the project (task 2)	Kaja 6h
Selecting, integrating, cleaning data	Kaja 2h, Kristiina 2h
Visualizing data	Kaja 2h, Kristiina 2h
Selecting and applying various classification models (k-nearest-neighbors, random forests, support vector machines) to get aim which physicochemical properties affect the wine quality most	Kaja 6h, Kristiina 6h
Selecting and applying various 5regression models (support vector machine, multiple regression) to predict wine quality	Kaja 6h, Kristiina 6h
Assessing used models	Kaja 4h, Kristiina 4h
Summarising and formulating results	Kaja 2h, Kristiina 2h
Project video	Kaja 1h, Kristiina 1h
Project poster	Kaja 3h, Kristiina 3h