Predicting Wine Quality

he two datasets are related to red and white variants of the Portuguese "Vinho Verde" wine. Only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.).

These datasets can be viewed as classification or regression tasks. The classes are ordered and not balanced (e.g. there are munch more normal wines than excellent or poor ones). Outlier detection algorithms could be used to detect the few excellent or poor wines. Also, we are not sure if all input variables are relevant.

Our goal in the project is to perform a data analysis on the dataset and figure out the features for the model that predicts the wine quality index(score from 0-9).
You can use either regression or classification for the model prediction.

Prerequisites:

We would highly recommend that before the hack night you have some kind of toolchain and development environment already installed and ready. If you have no idea where to start with this, try a combination like:

- Python
- scikit-learn / sklearn
- Pandas
- NumPy
- matplotlib
- An environment to work in - something like Jupyter or Spyder

For Linux people, your package manager should be able to handle all of this. If it somehow can't, see if you can at least install Python and pip and then use pip to install the above packages.

Objectives in this project:

- Perform data cleaning on the dataset
- Make a EDA report
- Visualize the distributions of various features and correlations between them

- Feature engineering to extract the correct features for the model
- Build a classification or regression model to predict the wine quality index

Dataset:

The dataset is in the form of a csv file and the link to download is given below:

https://drive.google.com/file/d/1pW85WoyJnWyLo8FlNEL4UDtkvoaxFI1Z/view?usp=sharing\

Dataset description:

The dataset contains 6498 rows and 14 columns

Input variables (based on physicochemical tests):

- 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

Output variable (based on sensory data):

- 12 - quality (score between 0 and 10)

13- good(1/0)

14-Color(red/white)

WorkFlow:

The workflow for the project is described in steps given below:

- Perform data cleaning using pandas library. Which includes replacing dropping useless information and filling missing values
- Make a Exploratory Data Analysis on the data using pandas.
- Visualize distributions and correlation of features using seaborn and pandas
- Build a regression or classification model to predict the wine quality
- Use various other standard regression or classification models and compre them using appropriate metrics.