**MACHINE LEARNING WINE QUALITY**

1. **Purpose of the problem**

We have built a machine learning model capable of predicting the quality of a type of wine. The main objective is to predict the quality of a wine based on its technical specifications. The model has significant implications in developing an application to predict the wine's quality before it is produced and to prepare for import. The model provides businesses with a prediction of wine quality before testing, saving time and costs in the quality assurance process.

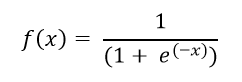
1. **Data description**

This dataset is related to red variants of the Portuguese "Vinho Verde" wine. The dataset describes the amount of various chemicals present in wine and their effect on its quality. Amount of various chemicals presents in wine and their effect on its quality.

The dataset includes the following fields: fixed acidity, Volatile acidity, Citric acid, Residual sugar, Chlorides, Free sulfur dioxide, Total sulfur dioxide, Density, pH, Sulphates, Alcohol, Quality

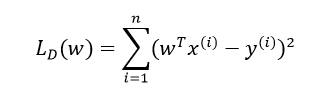
1. **Methodology used in the problem.**
   1. **Logistic Regression**

Logistic Regression is a popular machine learning algorithm used for binary classification problems. It is used to predict the probability that a given input belongs to a particular class, making it suitable for problems where the dependent variable is categorical, typically binary, such as 0 or 1, yes or no, spam or not spam, etc.



\* The reason for choosing the Logistic Regression method:

* Logistic Regression is a popular and effective classification model widely used in classification problems.
* It allows predicting the probability of a wine sample belonging to a specific class.
* This provides essential information for consumers and producers about the "quality" level of the wine.
  1. **Linear Regression**

Linear Regression is a fundamental machine learning model used to model the linear relationship between input variables and output variables. It is used to predict continuous values based on input variables. It allows for easy interpretation of the relationship between input variables and output variables. This enables a clear understanding of the influence of each input variable on the prediction outcome. 

\* The reason for choosing the Linear Regression method:

* This method is typically suitable when there is a linear relationship between the input variables and the target variable.
* Linear Regression is a simple, interpretable model.
* Linear Regression usually has a fast-training time, and the results from Linear Regression can be easily interpreted.
  1. **K-Nearest Neighbors (K-NN)**

A black and white image of a mathematical equation

Description automatically generatedK-Nearest Neighbors (K-NN) is a simple, yet effective supervised machine learning algorithm used for both classification and regression tasks. In K-NN, the "K" stands for the number of nearest neighbors to consider when making a prediction. It is often categorized as an instance-based or memory-based learning algorithm.

\* The reason for choosing the K-NN method:

* K-NN has the ability to handle complex relationships between input variables and the target variable without requiring assumptions about linearity.
* K-NN is based on the idea that similar samples will have similar outcomes.
* K-NN is a simple and easy-to-understand model.

1. **Evaluating Experimental Results**
   1. **Dataset Size**

The dataset consists of 1143 rows and 13 attributes, which can be considered a small dataset for many modern machine learning models. Therefore, we need to build a model that is not overly complex to avoid overfitting. Simple models such as Logistic Regression, Decision Trees, Linear Regression, and K-Nearest Neighbors (KNN) are good choices to consider.

* 1. **Train-Test Split**
* When comparing the data, given that there are 1143 data points, we will divide it into two parts: one for model training (the training set) and one for evaluating model performance (the test set). A common approach is to split the dataset in an 80:20 ratio, meaning 80% of the 1143 rows are used for training, and 20% for testing.
* The 80:20 split is often used because it provides a good balance between model training and evaluation. This ratio is widely employed in practice as it has proven to be effective in many scenarios.
* Further splitting the training set into a train-validation split at an 80:20 ratio can help evaluate the model more effectively.

=> Train-Validation-Test Split: Train: Val: Test = 64:16:20

* 1. **Choice of Evaluation Metrics**

For specific algorithms, different evaluation metrics can be used to assess the results. For example, for Linear Regression or Logistic Regression, metrics like Mean Squared Error (MSE), Mean Absolute Error (MAE), or accuracy can be used. For K-Nearest Neighbors or Decision Trees, accuracy is often used.

1. **Work assignment plan**