**Table of Contents**

[**1.** **Regression Algorithms** 1](#_Toc187006860)

[**1.2.** **Showing accuracies of each model.** 1](#_Toc187006861)

[**1.3.** **Explanation of how the two models work.** 3](#_Toc187006862)

[1.3.1. Linear Regression Model: 3](#_Toc187006863)

[1.3.2. Support Vector Regressor (SVR): 3](#_Toc187006864)

[**1.4.** **Reason for the better performance of Linear Regression:** 4](#_Toc187006865)

[**1.5.** **Suggest different methods to improve your models.** 4](#_Toc187006866)

[**1.6.** **Conclusion** 5](#_Toc187006867)

[**References** 6](#_Toc187006868)

# **Regression Algorithms**

The assignment task was to run two Regression models or algorithms on a dataset and compare their results to see which one performs better and why. So, first, a dataset (H, 2022) was selected, which detects the quality of wine on a scale of 0 to 10. The dataset consists of 4898 rows and 12 columns. The dataset has the following attributes (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
12. Output variable (based on sensory data):
    * quality (score between 0 and 10)

After confirming the dataset (H, 2022), some data exploratory analysis was done, such as checking whether the data has any missing values or outliers. After removing the outliers and filling in the missing values, a new dataset was formed. The distribution of data among different qualities of wine is not equal which means the dataset is imbalanced which is also a cause of not so good results by the models. Now, it turns to choosing algorithms/models to perform regression, out of many two algorithms were selected based on their specialty for the regression task such as **Linear Regression (LR)** and **Support Vector Regressor (SVR).** After choosing the models, the model was trained and tested with the dataset. In the below section, results are discussed briefly along with the working of the algorithms selected for the task. A section is added to examine the accuracy of both models which performs better and why it performs better. Finally, a detailed section is added to discuss the measures taken to increase the efficiency of the models.

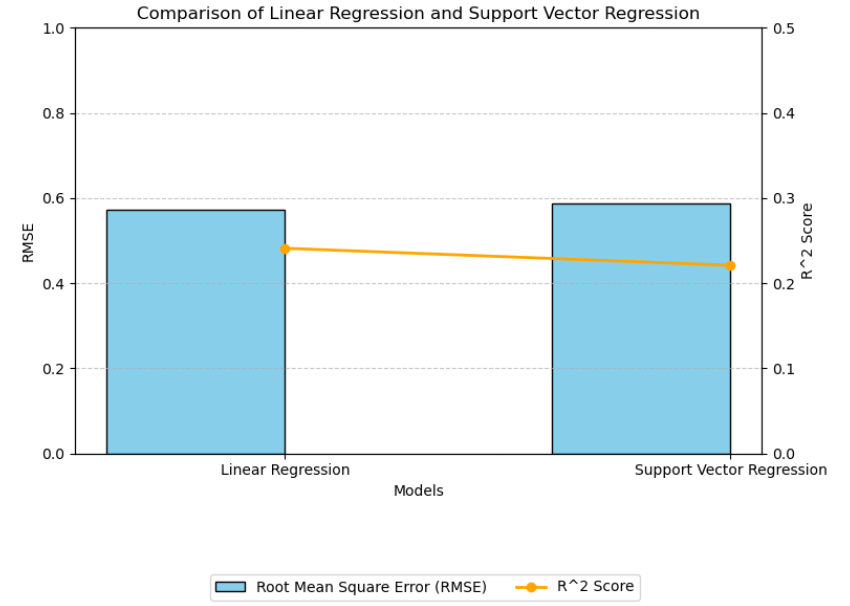
## **Showing accuracies of each model.**

The **Table 1: Linear Regression and Support Vector Regressor Results** is added that contains the RMSE and R2 of the both models, the results highlights that the **Linear Regression (LR)** outperforms **Support Vector Regressor (SVR)** on the wine quality white dataset (H, 2022). **Linear Regression** achieved a slightly lower **Root Mean Square Error (RMSE)** of 0.572 compared to Support Vector Regression's 0.588, showing that its predictions are closer to the true values on average. Additionally, **Linear Regression (LR)** achieved an **R2** score of 0.241, apprehending nearly 24% of the variance in the target variable, while **Support Vector Regressor (SVR)** apprehended only 22% (**R2**=0.221).

**Table 1: Linear Regression and Support Vector Regressor Results**

|  |  |  |
| --- | --- | --- |
|  | **Linear Regression** | **Support Vector Regression** |
| **Root Mean Square Error (RMSE)** | 0.572 | 0.588 |
| **R^2 Score** | 0.241 | 0.221 |
| **Remarks** | Performs better with approximately 24% of the variance | Performs worse as it only has approximately 22% of the variance |

From this, it can be deduced that **Linear Regression (LR)** is slightly more operational in modeling the primary relationship between the features and the target variable in the wine quality white dataset. The borderline difference in performance likely stems from the dataset exhibiting a mainly linear structure, which Linear Regression (LR) is better suited to handle, whereas Support Vector Regressor (SVR) may require additional hyper parameter tuning or kernel adjustments to improve its effectiveness. Different hyper parameter tuning were tried to improve the results of Linear Regression (LR) and Support Vector Regressor (SVR).



**Figure 1: Comparison of Linear Regression and Support Vector Machine**

## **Explanation of how the two models work.**

### Linear Regression Model:

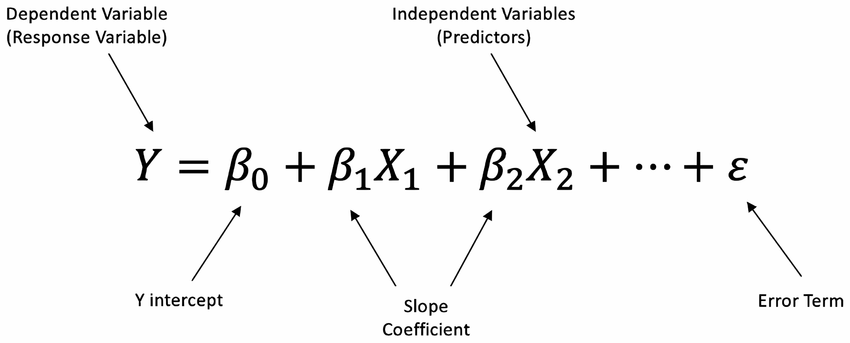
**Main Idea:**

Linear Regression is a model that uses statistical methods and generates a relationship between a dependent variable (also known as the target) which is the quality in our dataset and the independent variables (also known as attributes present in our dataset) by fitting (Kavita, 2024) a linear equation which is given in **Figure 2: Linear Regression Equation** to experimental data.

**Working:**

Linear Regression (LR) predicts a continuous (a variable over a non-empty range of the real numbers is continuous if it can take on any value in that range) target variable by modeling a linear relationship between input features and the target. The process involves defining the research question to implement the model, applying preprocessing on the data, choosing a hypothesis function, and minimizing the error using a loss function, usually **Root Mean Squared Error (RMSE)**. Coefficients (beta β) are optimized using Gradient Descent or the Normal Equation given in **Figure 2: Linear Regression Equation**. After training the model on the data, predictions are made using the learned coefficients, and the model is evaluated using metrics like **RMSE, or R2.** While simple and interpretable, Linear Regression (LR) assumes a linear relation between columns and is sensitive to outliers means outliers affect the result of the model, which can limit its performance on complex datasets (IBM, 2024).

**Equation**:



**Figure 2: Linear Regression Equation**

**Training**: The model is trained by finding the best-fit line that minimizes the sum of the squared differences (residuals) between observed values and predicted values (Least Squares Method).

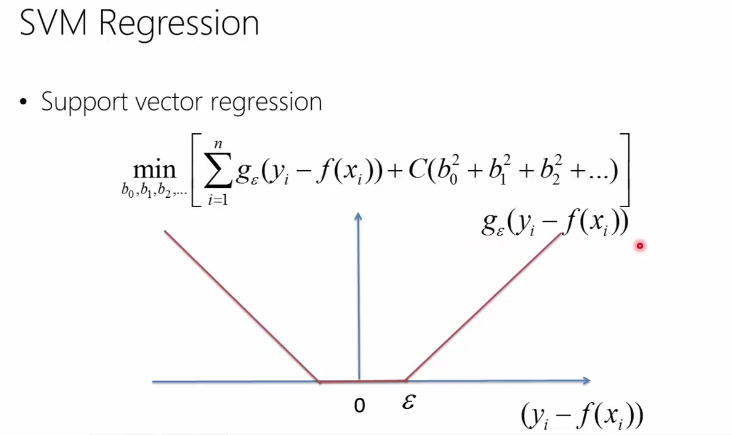
### Support Vector Regressor (SVR):

**Main Idea:**

Support Vector Regressor (SVR) (Sethi, 2024) is an extended form or originated from a Support Vector Machine (SVM), whose main idea is to fit the line in a range of specified threshold (also known as the epsilon) considering all the data points in the training set of the dataset.

**Working:**

Support Vector Regression (SVR) aims to predict values while maintaining a margin of tolerance, ϵ\epsilon, within which deviations from the actual values are not penalized. It identifies a function f(x)f(x) that minimizes errors outside this margin while balancing model complexity using a regularization parameter CC. The model focuses only on data points outside the ϵ\epsilon-margin, called support vectors, to define the regression function. For non-linear relationships, SVR applies kernel functions (e.g., RBF, polynomial) to transform the data into a higher-dimensional space, enabling it to capture complex patterns. Proper scaling and hyper parameter tuning are essential for its effectiveness (Sharp, 2023).

**Equation**:

**Figure 3: Equation used in SVR**

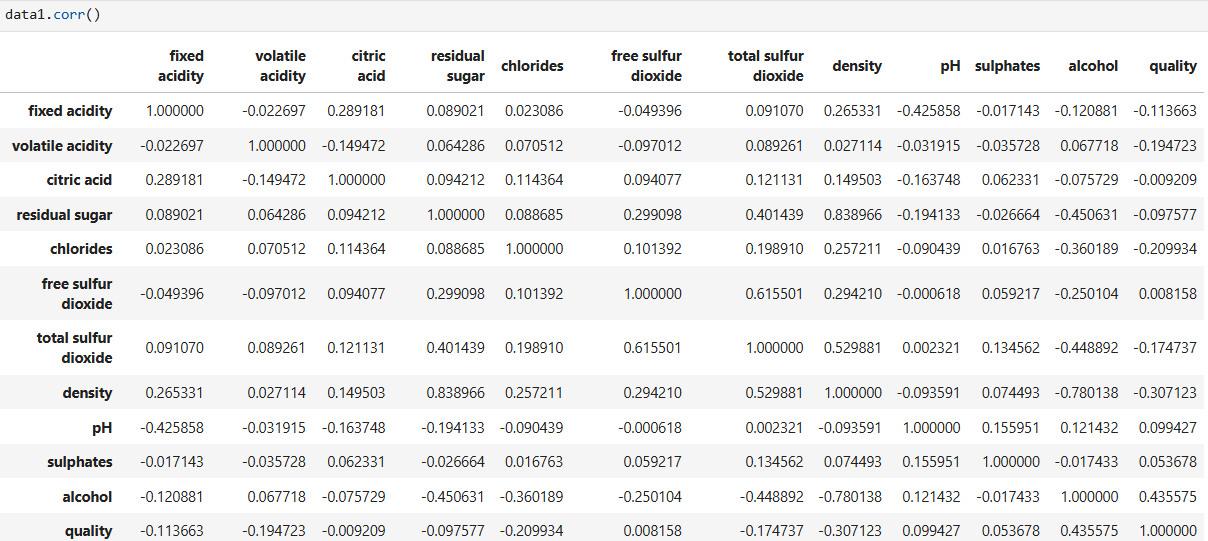
**Training**: SVR works by transforming the original problem using a kernel function into a higher-dimensional space where a linear regression problem is easier to solve. The commonly used kernels include linear, polynomial, and radial basis functions (RBF) (Sethi, 2024).

## **Reason for the better performance of Linear Regression:**

Linear Regression (LR) performs better because dataset features are linearly correlated with the target variable, making linear regression more effective. As the quality which was our target is strongly correlated with other features like density, alcohol, and total sulfur dioxide. However, its restrained performance specifies that linearity doesn't fully explain the variance. Whereas the reason for the Support Vector Regressor's (SVR) not so good performance might be kernel or hyper-parameters not according to the data or because the Support Vector Regressor (SVR) relies on more accurate data and some of the columns might need the scaling to be accurate.

## **Suggest different methods to improve your models.**

There are some methods that can be adopted to increase or make the performance better such as **feature engineering, hyper-parameter tuning, data scaling, and outlier detection.** Some of them are used in this assignment to improve the results.

The result of the Support Vector Regressor (SVR) was not good at first and showed approximately 15% of the variance. To improve the performance, First, the corr() function of pandas is used to check the correlation between the features as shown in **Figure 4: Correlation between the attributes of the Dataset** and remove the features with the least correlation between the result column and the other features. **Hyper-parameter tuning**, for Support Vector Regressor (SVR), tested different kernels ('rbf', 'poly') and tuned parameters like CC, ϵ\epsilon, and γ\gamma. **Data Scaling**, Normalize or standardize the dataset to improve Support Vector Regressor (SVR) performance. **Outlier Detection**, Use boxplots or scatterplots to identify and handle outliers, which can disproportionately affect all models.

**Figure 4: Correlation between the attributes of the Dataset**

## **Conclusion**

In this report, performance of two regression was compared, **Linear Regression (LR)** and **Support Vector Regressor (SVR),** on the white wine quality dataset. Some preprocessing techniques were performed on the datasets to address the missing values, outliers, and imbalanced classes. In the conclusion: **Linear Regression (LR)** outperformed **Support Vector Regressor (SVR)**, achieving a lower RMSE (0.572 vs. 0.588) and higher R² score (0.241 vs. 0.221) respectively. LR's better results can be deduced due to the dataset's linear structure and the strong correlation between quality and features like density, alcohol, and total sulfur dioxide. On the other hand, SVR's lower results highpoint its dependence on accurate scaling, proper hyper parameter tuning, and kernel selection. Various improvement methods, including feature engineering, hyper parameter tuning, data scaling, and outlier detection, were employed to enhance model performance. Ultimately, Linear Regression proved more effective due to its simplicity and alignment with the dataset's linear nature. However, SVR could benefit from further optimization and data preprocessing for better results.

# **References**

H, M. Y., 2022. *Wine Quality Dataset.* [Online]   
Available at: https://www.kaggle.com/datasets/yasserh/wine-quality-dataset  
[Accessed 29 12 2024].

IBM, 2024. *What Is Linear Regression?.* [Online]   
Available at: https://www.ibm.com/think/topics/linear-regression  
[Accessed 19 12 2024].

Kavita, 2024. *Linear Regression: A Comprehensive guide.* s.l.:s.n.

Sethi, A., 2024. *Support Vector Regression Tutorial for Machine Learning.* s.l.:s.n.

Sharp, T., 2023. *An Introduction to Support Vector Regression (SVR).* [Online]   
Available at: https://towardsdatascience.com/an-introduction-to-support-vector-regression-svr-a3ebc1672c2  
[Accessed 2024].