
UCD Michael Smurfit
Graduate Business School

# Group Assignment – Group 6

| | |
|---|---|
| Assignment Topic | Practical Data Analysis |
| Module Title | Statistical Learning - 2022/23 Summer |
| Module Code | MIS41120 |
| Lecturers | Assoc Professor Sean McGarraghy |
| Submission Deadline | 19th July 2023 |
| Session | 2022/23 Summer |
| Name & Student Number | Abhishek Sabharwal - 22200432<br>Anushka Jain – 22200246<br>Ayushi Gautam - 22200403 |
| Grade/Mark | |

**A SIGNED COPY OF THIS FORM MUST ACCOMPANY ALL SUBMISSIONS FOR ASSESSMENT. STUDENTS SHOULD KEEP A COPY OF ALL WORK SUBMITTED.**

Procedures for Submission and Late Submission
Ensure that you have checked the school's procedures for the submission of assessments.
**Note:** There are penalties for the late submission of assessments. For further information please see the University's ***Policy on Late Submission of Coursework,*** (http://www.ucd.ie/registrar/)

**Plagiarism:** the unacknowledged inclusion of another person's writings or ideas or works, in any formally presented work (including essays, examinations, projects, laboratory reports or presentations). The penalties associated with plagiarism are designed to impose sanctions that reflect the seriousness of the University's commitment to academic integrity. Ensure that you have read the University's ***Briefing for Students on Academic Integrity and Plagiarism*** and the UCD ***Plagiarism Statement, Plagiarism Policy and Procedures***, (http://www.ucd.ie/registrar/)

Declaration of Authorship
I declare that all material in this assessment is my own work except where there is clear acknowledgement and appropriate reference to the work of others.
Name : Abhishek Sabharwal, Anushka Jain, Ayushi Gautam
Date : 19th July 2023

# Contribution

| Team Member Name | % to Modelling | % to Report Preparation |
|---|---|---|
| Abhishek Sabharwal | 33.33 | 33.33 |
| Anushka Jain | 33.33 | 33.33 |
| Ayushi Gautam | 33.33 | 33.33 |

We, as a team, collaborated cohesively to accomplish the task, with each member making equal and valuable contributions to various aspects. Initially, we individually conducted mathematical modeling and later cross-validated the approaches collectively. For coding in Python, we assigned individual responsibilities, focusing on regularization techniques such as ridge regression, lasso, and elastic net. When preparing the report, we all worked together, contributing to content compilation, layout, and formatting decisions.

Throughout the project, we maintained effective communication and held regular meetings to track progress and resolve challenges. We explored regression, support vector machines (SVM), and multilayer perceptrons (MLP) on publicly available datasets with more than 20 predictors. Evaluating each method's performance, accuracy, interpretability, and efficiency, we discussed the results as a team, making informed decisions. Our collaborative efforts and combined skills enabled us to successfully complete the task, demonstrating our strengths and achieving high-quality outcomes.

# Table of Contents

# Introduction

## Linear Regression

Linear regression is defined as an algorithm that provides a linear relationship between an independent variable and a dependent variable to predict the outcome of future events. The independent variable is also the predictor or explanatory variable that remains unchanged due to the change in other variables. However, the dependent variable changes with fluctuations in the independent variable. The regression model predicts the value of the dependent variable, which is the response or outcome variable being analysed or studied (Kanade, 2022).

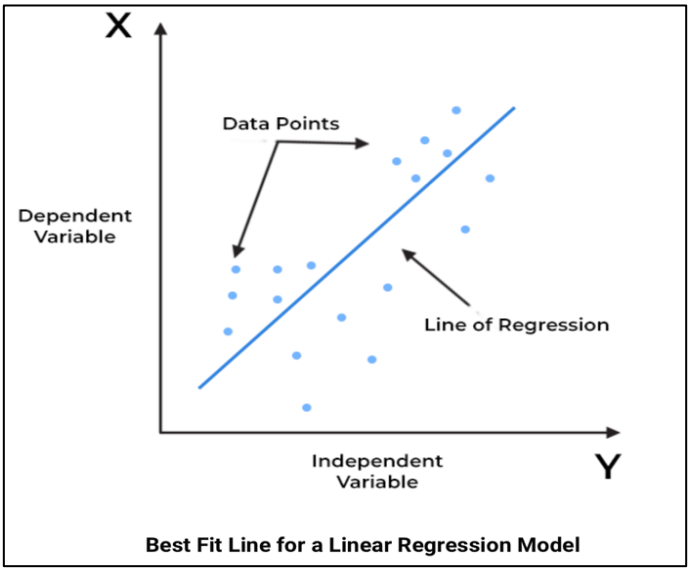


*Figure 1 (Kanade, 2023)*

In the above figure, the X-axis denotes the independent variable, and the Y-axis represents the output or dependent variable. The continuous line depicted in the plot is known as the "best fit line." This line plotted to suitably fit the given data points, making it the most suitable representation of the relationship between the independent and dependent variables. The primary objective of the linear regression algorithm is to determine this best fit line.

## Multi-Layer Perceptron (MLP)

Multi-layer perception is also known as MLP. It is fully connected dense layers, which transform any input dimension to the desired dimension. A multi-layer perception is a neural network that has multiple layers. To create a neural network, we combine neurons together so that the outputs of some neurons are inputs of other neurons.

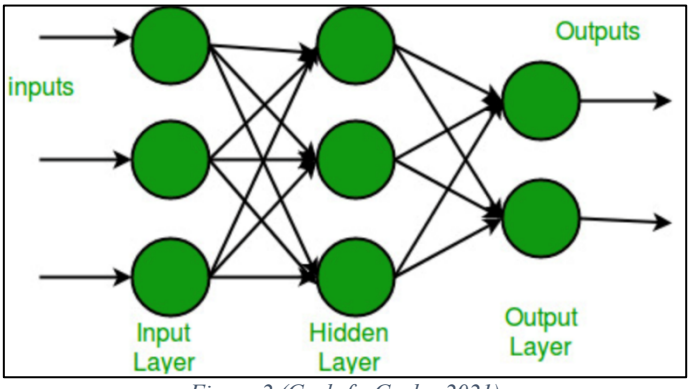


*Figure 2 (GeeksforGeeks, 2021)*

A multi-layer perceptron has one input layer and for each input, there is one neuron (or node), it has one output layer with a single node for each output and it can have any number of hidden layers and each hidden layer can have any number of nodes.

The activation function is represented by the threshold T. The neuron produces the value 1 if the weighted total of the inputs is greater than zero; else, the output value is 0. The perceptron can be utilized as a binary classification model with this discrete output, controlled by the activation function, defining a linear decision boundary. In order to reduce the distance between incorrectly categorized points and the decision boundary, it determines the separation hyperplane. Neurons in a Multilayer Perceptron can utilize any arbitrary activation function, unlike neurons in a Perceptron, which must have an activation function that enforces a threshold, such as ReLU or sigmoid (Bento, 2021).

## Support Vector Machine (SVM)

Support Vector Machine (SVM) is a powerful supervised machine learning algorithm utilized for both classification and regression tasks. Although SVM can be employed for regression, it is particularly well-suited for solving classification problems. The primary objective of the SVM algorithm is to identify the optimal hyperplane within an N-dimensional feature space that effectively separates data points belonging to different classes (Aswathisasidharan, 2021).

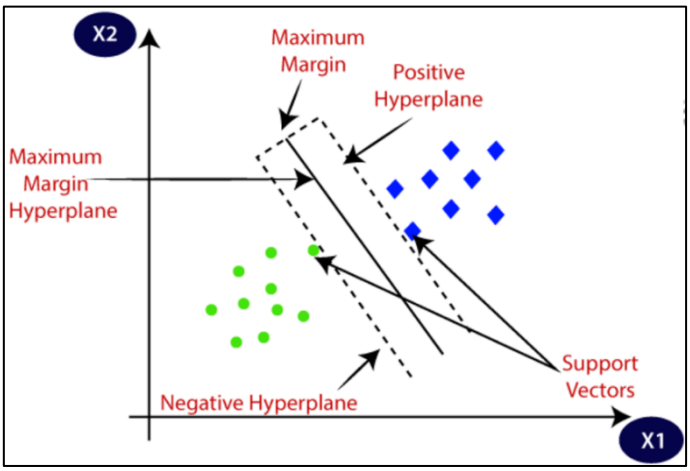


*Figure 3 (Saini, 2023)*

### SVM for Classification

In classification problems, SVM aims to create a hyperplane that maximizes the margin between the closest data points of different classes. This margin is known as the "decision boundary," and the hyperplane acts as a discriminatory surface, classifying data points into distinct classes. SVM strives to find the hyperplane with the largest margin to ensure robust generalization and better handling of unseen data.

### SVM for Regression

Though SVM can be used for regression, it is predominantly renowned for its classification capabilities. In regression tasks, SVM attempts to find a hyperplane that best fits the data points, optimizing the trade-off between fitting the data closely and maintaining a smooth decision boundary. However, other regression algorithms may be more commonly used for tasks where continuous prediction is the primary objective.

## Regularization

Regularization is a technique used in machine learning and deep learning to prevent overfitting and improve the generalization performance of a model. It involves adding a penalty term to the loss function during training. This penalty discourages the model from becoming too complex or having large parameter values, which helps in controlling the model's ability to fit noise in the training data (Jain, 2018).
In this report, we will explore three common regularization approaches: Ridge, Lasso, and Elastic Net, along with three widely used models: Regression, Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP).

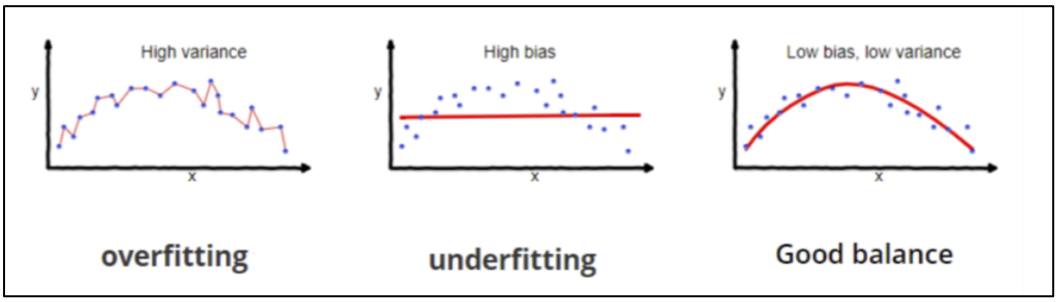


*Figure 2 (GeeksforGeeks, 2019)*

## Regularization Approaches

### Lasso Regression

A regression model which uses the L1 Regularization technique is called LASSO (Least Absolute Shrinkage and Selection Operator) regression. Lasso Regression adds the "absolute value of magnitude" of the coefficient as a penalty term to the loss function (L) (GeeksforGeeks, 2019). This penalty term induces sparsity in the coefficients, allowing some coefficients to be exactly zero. This feature selection capability makes Lasso regression useful when dealing with datasets containing many irrelevant features. By eliminating irrelevant features, Lasso creates a more interpretable and simplified model. However, it may struggle with correlated features as it tends to pick only one of them and set the others to zero.

### Ridge Regression

A regression model that uses the L2 regularization technique is called Ridge regression. Ridge regression adds the "squared magnitude" of the coefficient as a penalty term to the loss function(L), effectively shrinking them towards zero (GeeksforGeeks, 2019). Ridge regression is particularly useful in handling multicollinearity issues when dealing with highly correlated features. It ensures a more stable and robust model by reducing the variance of the estimates. However, it does not perform feature selection as it keeps all predictors in the model.

### Elastic Net

Elastic Net is a combination of Ridge and Lasso regularization techniques. It adds both L1 and L2 penalty terms to the loss function. Elastic Net leverages the strengths of both Ridge and Lasso, providing a balance between feature selection and handling multicollinearity. It is a more flexible approach that can be favourable in situations where both Ridge and Lasso exhibit limitations.

## Performance Metrics for Regression

### Root Mean Squared Error (RMSE)

RMSE is a widely used metric to measure the accuracy of regression models. It represents the square root of the average of the squared differences between predicted and actual target values. RMSE assesses how closely the predicted values align with the true target values. A lower RMSE indicates better model performance, as it signifies smaller errors between predictions and actual values (Glen, 2022).
We use RMSE as it provides a straightforward interpretation of the average prediction error, allowing us to assess the model's precision in predicting housing prices.

### R-squared ($R^2$) Score

R-Squared (coefficient of determination) is a statistical measure in a regression model that determines the proportion of variance in the dependent variable that can be explained by the independent variable. In other words, r-squared shows how well the data fit the regression model (the goodness of fit) (Taylor, 2020). $R^2$ score ranges from 0 to 1, with 1 indicating a perfect fit and 0 indicating a model that performs no better than predicting the mean of the target variable. Higher $R^2$ scores indicate better model performance in capturing the variance in the data.

## Performance Metrics for Classification

### Recall (Sensitivity)

The recall is calculated as the ratio between the number of Positive samples correctly classified as Positive to the total number of Positive samples. The recall measures the model's ability to detect Positive samples. The higher the recall, the more positive samples detected (Gad, 2020).

$$Recall = \frac{TP}{TP + FN}$$

### F1 Score

The F-score, also called the F1-score, is a measure of a model's accuracy on a dataset. It is used to evaluate binary classification systems, which classify examples into 'positive' or 'negative'. The F-score is a way of combining the precision and recall of the model, and it is defined as the harmonic mean of the model's precision and recall. It is possible to adjust the F-score to give more importance to precision over recall, or vice-versa. Common adjusted F-scores are the F0.5-score and the F2-score, as well as the standard F1-score (Wood, 2019).

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

### Accuracy

Accuracy is a metric that generally describes how the model performs across all classes. It is useful when all classes are of equal importance. It is calculated as the ratio between the number of correct predictions to the total number of predictions (Gad, 2020). It provides a general measure of the model's performance but can be misleading in imbalanced datasets.

$$Accuracy = \frac{TN + TP}{TN + FP + TP + FN}$$

# Boston Housing Dataset: Results & Analysis

## Task 1: Apply multiple regression on the Boston Dataset and list the predictors for which the null hypothesis H0: βj = 0 can be rejected?

In the analysis of the Boston Dataset using a multiple regression model, the coefficients and p-values were calculated for each predictor as shown below.

| | coef | std err | t | P>\|t\| |
|---|---|---|---|---|
| const | 9.9967 | 6.979 | 1.432 | 0.153 |
| ZN | 0.0364 | 0.019 | 1.943 | 0.053 |
| INDUS | -0.0694 | 0.084 | -0.825 | 0.410 |
| CHAS | -1.3117 | 1.179 | -1.112 | 0.267 |
| NOX | -6.9288 | 5.225 | -1.326 | 0.185 |
| RM | -0.3348 | 0.573 | -0.585 | 0.559 |
| AGE | 0.0013 | 0.018 | 0.074 | 0.941 |
| DIS | -0.7089 | 0.271 | -2.612 | 0.009 |
| RAD | 0.5389 | 0.088 | 6.151 | 0.000 |
| TAX | -0.0014 | 0.005 | -0.263 | 0.793 |
| PTRATIO | -0.0834 | 0.179 | -0.465 | 0.642 |
| B | -0.0096 | 0.004 | -2.625 | 0.009 |
| LSTAT | 0.2356 | 0.069 | 3.431 | 0.001 |

The null hypothesis in this context typically states that there is no relationship between the predictor variable and the response variable, while the alternative hypothesis suggests that there is a significant relationship between them. A ***p-value*** less than **0.05** is used as a threshold to determine statistical significance.

The results indicated that the null hypothesis can be rejected for some predictors but not for others. Based on the results, ***'DIS', 'RAD', 'B'***, and ***'LSTAT'*** are statistically significant and have a significant relationship with the response variable.

## Task 2: Apply Regression with Ridge, Lasso, and Elastic Net regularisation techniques.

| Regularisation | Root Mean Square Error | R Squared Value |
|---|---|---|
| Ridge | [1.91, 2.73, 1.72, 10.02, 11.30] | [-23.78, -6.73, -64.22, 0.27, 0.14] |
| Lasso | [1.30, 1.60, 1.38, 10.49, 11.41] | [-10.45, -1.64, -41.02, 0.20, 0.13] |
| Elastic Net | [1.29, 3.01, 1.14, 10.49, 11.67] | [-10.25, -8.36, -27.67, 0.20, 0.08] |

**Inferences:**

- **RMSE Comparison:** It can be observed that Elastic Net regularization has the lowest RMSE value of **1.29** across all runs compared to Ridge and Lasso. This suggests that the Elastic Net model has the best overall predictive performance in this scenario.
- **R-squared Comparison:** For all regularization techniques, R-squared values are negative for the first three instances. This indicates that the models perform poorly and explain less variance than a horizontal line ($R^2$=0). These models might not be appropriate or useful for prediction in these cases. However, for the last two instances, the R-squared values are positive, indicating some improvement, but the values are still relatively low.

## Task 3: Apply MLP with Ridge, Lasso, and Elastic Net regularisation techniques.

| Regularisation | Root Mean squared error | R Squared Value |
|---|---|---|
| Ridge | [0.57, 3.23, 0.47, 09.81, 11.32] | [-1.19, -09.79, -03.91, 0.30, 0.14] |
| Lasso | [0.71, 4.06, 0.62, 09.93, 11.27] | [-2.39, -16.07, -07.44, 0.29, 0.15] |
| Elastic Net | [1.88, 2.61, 1.66, 10.01, 11.29] | [-23.0, -06.06, -59.82, 0.28, 0.14] |

**Inferences:**

- **RMSE Comparison:** Among the three regularization techniques, Ridge regularization consistently achieves the lowest RMSE values across all instances. It results in the smallest prediction errors on average compared to Lasso and Elastic Net.
- **R-squared Comparison:** Similar to regression analysis, all regularization techniques have negative R-squared values for the first three instances and positive values for the last two instances. Hence, these models might not be appropriate or useful for prediction in these cases.

## Task 4: MLP v/s Regression

- **Performance Metrics:** On comparing the results across regression and MLP, the Multilayer Perceptron (MLP) model generally outperforms the regression model across all three regularization techniques due to its non-linear capabilities and feature representation.
- **Model Complexity and Efficiency:** MLP models' non-linear capabilities and ability to learn complex patterns lead to superior performance but requires more computational time and resources resulting in lower efficiency compared to the regression model.

## Task 5: Apply SVM with Ridge, Lasso, and Elastic Net regularisation techniques.

| Regularisation | Accuracy | F1 Score | Recall |
|---|---|---|---|
| Ridge | 0.8922 | 0.8920 | 0.8922 |
| Lasso | 0.8726 | 0.8721 | 0.8726 |
| Elastic Net | 0.8725 | 0.8721 | 0.8726 |

**Inferences:**

- **Accuracy Comparison:** Ridge regularization achieved the highest accuracy and correctly predicted approximately **89.22%** of the data instances. Lasso and Elastic Net have slightly lower accuracies. However, the differences in accuracy are not significant.
- **F1 Score and Recall Comparison:** all regularization techniques have similar values. This indicates that the three regularization techniques perform similarly in terms of correctly identifying positive samples and achieving a balance between precision and recall.
- In this case, the impact of regularization might not be substantial, as the results are quite close for the three techniques.
- Overall, the SVM model appears to perform reasonably well with all three regularization techniques (Ridge, Lasso, and Elastic Net), achieving high accuracy and comparable F1 scores and recall values. Further optimization and tuning of hyperparameters could potentially lead to even better results and help in selecting the most appropriate regularization technique for this specific dataset and problem.

# Student Performance Dataset

## Overview

The dataset focusses on student achievement in secondary education from two Portuguese schools, encompassing student grades, demographic, social, and school-related features. The data was collected through school reports and questionnaires, specifically focusing on the Portuguese language subject. The objective of this dataset is to predict the third term results, represented by the target attribute, G3, based on the various available information (Cortez, 2014).

## Data Transformation

To render the dataset suitable for regression, data transformation was undertaken to convert textual categorical attributes into numerical variables. This conversion proved vital for facilitating diverse analytical and modeling tasks, as machine learning algorithms necessitate numerical inputs for predictors. By encoding categorical attributes into numeric values, we can proficiently analyse the correlations between various factors and student performance.

## Results and Analysis

### Task 1: Apply Regression with Ridge, Lasso, and Elastic Net regularisation techniques.

| Regularisation | Root Mean Square Error | R Squared Value |
|---|---|---|
| Ridge | [3.25, 2.68, 2.13, 2.23, 3.36] | [ 0.09, 0.39, 0.20, 0.23, 0.15 ] |
| Lasso | [3.51, 3.40, 2.26, 2.59, 4.08] | [-0.07, 0.02, 0.09, -0.03, -0.25] |
| Elastic Net | [3.37, 3.13, 2.09, 2.44, 3.74] | [ 0.02, 0.17, 0.23, 0.08, -0.05 ] |

**Inferences:**

- **RMSE Comparison:** Elastic Net achieved the lowest RMSE values consistently across all test cases, making it the best-performing technique in terms of predictive accuracy in this comparison.
- **R-squared Comparison:** Ridge had the highest $R^2$ values on average, indicating that it provided the best overall fit to the data among the three techniques in this comparison.
- In summary, Ridge appears to be the most favourable choice in this comparison due to its balanced performance in terms of both RMSE and R-squared. However, Elastic Net may also be suitable when focusing on specific aspects of the model's performance.

### Task 2: Apply MLP with Ridge, Lasso, and Elastic Net regularisation techniques.

| Regularisation | Root Mean squared error | R Squared Value |
|---|---|---|
| Ridge | [3.49, 2.84, 2.58, 2.65, 4.03] | [-0.05, 0.32, -0.18, -0.08, -0.23] |
| Lasso | [3.58, 3.70, 2.70, 2.84, 4.13] | [-0.11, -0.16, -0.30, -0.25, -0.28] |
| Elastic Net | [3.24, 2.69, 2.12, 2.23, 3.34] | [ 0.09, 0.39, 0.20, 0.23, 0.16] |

**Inferences:**

- **RMSE Comparison:** Elastic Net achieved the lowest RMSE values consistently across all test cases, making it the best-performing technique in terms of predictive accuracy in this comparison.
- **R-squared Comparison:** Elastic Net had the highest $R^2$ values on average, indicating that it provided the best overall fit to the data among the three techniques in this comparison.
- Based on both metrics, Elastic Net appears to be the most favourable regularisation technique as it provides a better balance between prediction accuracy and model fit.

## Task 3: Apply SVM with Ridge, Lasso, and Elastic Net regularisation techniques.

| Regularisation | Accuracy | F1 Score | Recall |
|---|---|---|---|
| Ridge | 0.6423 | 0.5858 | 0.6423 |
| Lasso | 0.6385 | 0.5893 | 0.6385 |
| Elastic Net | 0.6346 | 0.5820 | 0.6346 |

**Inferences:**

- **Accuracy Comparison:** The SVM model with Ridge regularization achieved the highest accuracy of **64.23%**, followed closely by Lasso and Elastic Net regularization. However, the differences in accuracy between Ridge, Lasso, and Elastic Net are relatively small.
- **F1 Score Comparison:** The highest F1 score was obtained by the Lasso regularization with a value of **0.5893**. However, the differences between the techniques' F1 scores are also relatively minor
- **Recall Comparison:** Ridge regularization achieved the highest recall of **0.6423**. Like the other metrics, the variation in recall values between the techniques is not substantial.
- Overall, the performance differences between the three regularization techniques for SVM are relatively small. Ridge regularization slightly outperformed Lasso and Elastic Net in accuracy and recall, while Lasso performed best in terms of F1 score.
- The minor performance differences between the regularization techniques can be attributed to the nature of the dataset and the underlying characteristics of the regularization methods. Ridge regularization, by penalizing the sum of squared coefficients, can handle multicollinearity effectively and maintain all relevant features in the model, which might have led to its slightly better accuracy and recall. On the other hand, Lasso regularization's tendency to induce sparsity and retain only a subset of important features likely contributed to its superior F1 score as it strikes a balance between precision and recall. However, since the dataset did not have an overwhelming amount of multicollinearity, the differences in performance between the techniques remained relatively subtle.

# References:


1. Aswathisasidharan. (2021). Support Vector Machine Algorithm, GeeksforGeeks. Available at: https://www.geeksforgeeks.org/support-vector-machine-algorithm/
2. Bento, C. (2021). Multilayer Perceptron Explained with a Real-Life Example and Python Code: Sentiment Analysis. Medium. Available at: https://towardsdatascience.com/multilayer-perceptron-explained-with-a-real-life-example-and-python-code-sentiment-analysis-cb408ee93141
3. Cortez, P. (2014). Student Performance. UCI Machine Learning Repository. Available at: https://doi.org/10.24432/C5TG7T
4. Gad, A.F. (2020). Accuracy, Precision, and Recall in Deep Learning, Paperspace Blog. Available at: https://blog.paperspace.com/deep-learning-metrics-precision-recall-accuracy/
5. GeeksforGeeks. (2019). Regularization in Machine Learning. Available at: https://www.geeksforgeeks.org/regularization-in-machine-learning/
6. GeeksforGeeks. (2021). Multi-Layer Perceptron Learning in Tensorflow. Available at: https://www.geeksforgeeks.org/multi-layer-perceptron-learning-in-tensorflow/
7. Glen, S. (2022). RMSE: Root Mean Square Error, Statistics How To. Available at: https://www.statisticshowto.com/probability-and-statistics/regression-analysis/rmse-root-mean-square-error/
8. Jain, S. jain (2018). An Overview of Regularization Techniques in Deep Learning, Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2018/04/fundamentals-deep-learning-regularization-techniques/
9. Kanade, V.A. (2023) 'What is linear regression?', Spiceworks. Available at: https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-linear-regression/
10. Saini, A. (2023) Guide on Support Vector Machine (SVM) algorithm, Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners/
11. Taylor, S. (2020). 'R-Squared: A statistical measure that determines the proportion of variance in the dependent variable that can be explained by the independent variable', Corporate Finance Institute, 17 March. Available at: https://corporatefinanceinstitute.com/resources/data-science/r-squared/
12. Wood, T. (2019). F-Score, DeepAI. Available at: https://deepai.org/machine-learning-glossary-and-terms/f-score