Academic Performance Prediction

Data Set : Student Performance Dataset

Vaishnavi Channakeshava, channake@usc.edu
Sanjana Vasudeva, sanjanav@usc.edu

May 4, 2022

1 Abstract

The ability to predict a student's performance could be beneficial in various ways. Students' demographic information, their prior academic performance, and social and school-related attributes can serve as the data set for a supervised machine learning algorithm. The models can be utilized to forecast fresh student performance, making them advantageous for spotting potential poor performers.

The goal of the project is to compare some of the most advanced learning algorithms currently available for classification and regression. The analysis was carried out on the Student Performance Dataset [1] with six models for regression and seven algorithms for classification. These models were trained with five different feature selection methods along with two-dimensionality reduction methods. The models with the best performance were Gradient Boost and MLP for Mission I for regression and classification respectively. For Mission II, Gradient Boost and SVM were determined to be the best models for regression and classification. For Mission III, ANN and SVM proved to be the best models.

2 Introduction

2.1 Problem Assessment and Goals

The dataset we chose is the student performance dataset. This dataset relates to secondary school student achievement in two Portuguese schools wherein only the performance of the Portuguese language has been considered. The data has been acquired through school reports and questionnaires. It includes student grades(G1, G2, and G3), demographic, social, and school-related factors. All the grade values are numeric and are in the range of 0 - 20.

The goal of the project is to predict first-period academic performance (G1) and final academic performance(G3), which can be broadly categorized into three missions:

MISSION I: Predict first-period academic performance without any prior academic performance data.
MISSION II: Predict final academic performance without any prior academic performance data.
MISSION III: Predict final academic performance with any prior academic performance data.

The numerical values of the grades have been predicted using regression. The prediction of the 5-class categorical value has been achieved through classification. Prior to performing regression and classification, the data was pre-processed as described in section 3.2.

3 Approach and Implementation

3.1 Dataset Usage

The given dataset has two parts - train and test. The training set has 486 points, and the test dataset has 163 points. This preprocessed data is provided as the input to various regression and classification models with and without regularization. Their performance for the test set is tabulated. 5-fold cross-validation is used on all the models and is evaluated against the previously obtained results.

3.2 Preprocessing

Pre-processing is performed to convert the data to more usable form. Initially, the binary categorical data is converted to binary numeric values, then the four categorical non-binary features - Mjob, Fjob, reason

and guardian are removed. The data consists of 13 binary features and 13 numeric features. Then non-binary categorical values were label encoded using sklearn's inbuilt label encoder function. Since the number of features(26) in the dataset is high, label encoding is preferred over one-hot encoding. It also prevents the curse of dimensionality.

The other pre-processing technique was normalization using the min-max scaler function, which ensures that all the values in the dataset lie in the range of 0 - 1. Normalization is preferred over standardization as the latter sometimes results in negative values in the dataset.

For the classification problem, the output variable which is in the range [0,20] is converted into a 5-class categorical value (I, II, III, IV, V) based on the following rule:

- 16-20 - Class I
- 14-15 - Class II
- 12-13 - Class III
- 10-11 - Class IV
- 0-9 - Class V

Feature selection : Feature selection is performed to increase the computational efficiency of the models i.e. it optimizes the model by selecting a subset of the features. Six different feature selection methods have been implemented - Random Forest, Mutual Information, Chi-Squared, Sequential Feature selection, and Recursive Feature Analysis(RFE).

- Random Forest: Random Forests aggregates a specified number of decision trees. Not every tree sees all the features, thereby preventing overfitting. Each tree is a sequence of yes-no questions based on a single or combination of features and hence the importance of each feature is derived. The parameter number of estimators (nestimators) for random forest feature selection has been set to 100.
- Mutual Information: Mutual Information measures the entropy between the feature and the target value. The features with higher entropy are selected. Mutual information for the regression task has been calculated for continuous variables whereas for the classification task it has been calculated for discrete variables. The parameter number of features (k) is chosen to be 10 which is the optimized value.
- Chi-Squared: The Chi-square test is used for categorical features in a dataset. The Chi-square score is calculated between each feature and the target, then the desired number of features with the best Chi-square scores are selected. It considers how dependent the target variable is on the input feature. The parameter number of features (k) is determined to be 16 which is the optimized value.
- Sequential Feature Selection: The algorithm removes features at every stage to form a feature subset. The features to be removed are chosen based on the cross-validation score. The parameter number of features (k) is chosen to be 10 which is the optimized value along with cross-validation (cv) of 6 folds.
- Recursive Feature Analysis: An estimator is trained on the initial set of features, and the importance of each feature is obtained. Then the least important features are removed from the subset. This approach is followed recursively until the desired number of features is achieved. The parameter number of features (k) is determined to be 15 which is the optimized value.

The optimal number of features to be selected was determined by calculating RMSE and macro f1 scores for regression and classification problems respectively, for all the models mentioned in sections 3.5.1 and 3.5.2. The number of features selected ranged from 1 to 26, and the value which gave the least RMSE and the best macro f1 score was chosen.

3.3 Feature engineering (if applicable)

"Not Applicable"

3.4 Feature dimensionality adjustment (if applicable)

Principal Component Analysis (PCA) and Linear Discriminant Analysis(LDA) ars implemented as a part of Feature dimensionality adjustment, where LDA is exclusively applied for the classification problem.

LDA uses Fisher's linear discriminant to reduce the dimensionality of the dataset while maximizing the separation between the classes. This is achieved by maximizing the distance between the class means and minimizing within-class variance.

PCA uses the principal components that are determined by singular value decomposition. Principal components are the directions that maximize variation in the projected data which are the eigenvectors. It drops the eigenvectors that are relatively unimportant thereby reducing the dimension.

LDA takes into account the categories in the data while reducing the dimension, whereas PCA does not. Hence, LDA is utilized for feature dimensionality adjustment in the Classification problem.

3.5 Training, Classification or Regression, and Model Selection

The models are developed with help of sklearn library with parameter hyper tuning. The parameters for regularization were chosen with optimization. 5-fold cross-validation has been performed on all regression and classification models to evaluate the model performance. The training data is split into five folds where four folds are used for training and one fold is used for evaluating the model performance and is performed for 5 runs. Regularization is a technique that discourages learning a complex model to avoid the risk of overfitting.

Reference systems have been designed for both regression and classification tasks to evaluate the performance of the models.

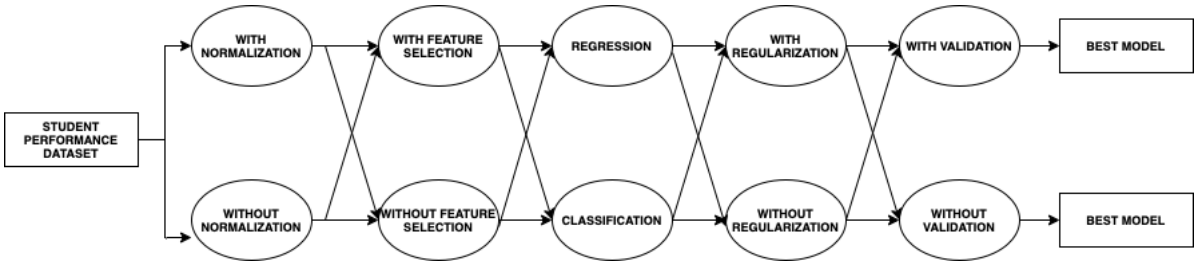

Figure 1: Flow Chart

The flow chart depicts the entire workflow of the approach followed for the regression and classification problem. Initially, the data is pre-processed. The performance of all the machine learning models with this pre-processed data, with and without feature extraction and with feature dimensionality reduction is compared with the performance of the corresponding models with normalized pre-processed data. If the unnormalized data achieved better performance, further evaluations were performed on unnormalized data, otherwise, further evaluations were performed on normalized data.

Regression and Classification models with and without regularization were performed on the pre-processed data chosen. 5-fold cross-validation has been performed on all models as well. We evaluate the following performances:

- Performance without normalization
- Performance with normalization
- Performance of models with best pre-processed data with 5 fold cross validation
- Performance of regularized model with the best pre-processed data
- Performance of regularized model with the best pre-processed data with 5 fold cross validation

The RMSE value in the case of regression and macro F-1 score in the case of classification is chosen as the metrics of comparison. The best performing model among these was considered accordingly.

3.5.1 Regression

The regression task is to predict the numerical value of grade G1 and G3.

Reference System for Regression: The reference systems for the regression problem have been devised as follows:

- Trivial System: System that always outputs the mean output value from the training set.
- 1NN: Output value is the same as the nearest training-set data point in feature data set.
- Linear Regression: Determining the straight line that best fits a set of scattered data points.

Regression Models :

The regression models chosen are listed as follows which were implemented using sklearn.

- SVR : For regression, a margin of tolerance (epsilon) is set in approximation to the SVM. The aim is to minimize error, individualizing the hyperplane which maximizes the margin, keeping in mind that part of the error is tolerated. The regularization parameter has been optimized as well. The hyperplane is determined by implementing different kernels - Linear, RBF, polynomial and sigmoid. The best-performing kernel is chosen. The hyperplane is determined by using Linear Kernel. The regularization parameter has been set to 3 for the regularized model.
- MLP Regressor: Multi-layer Perceptron regressor and optimizes the model using adam optimizer. The L2 penalty (regularization term) has been optimized and set to 0.1. The MLP regressor model is trained over 500 iterations.
- Gradient Boosting Regressor: Gradient boosting is one of the variants of ensemble methods where you create multiple weak models and combine them to get better performance as a whole. In each stage, a regression tree is fit on the negative gradient of the given loss function.
- KNN Regressor : The KNN algorithm predicts the values based on feature similarity. The value is determined by calculating the average of the numerical target of the K nearest neighbors. The best value of K has been chosen after experimenting with values of K ranging from 1 to 80 and choosing the one which gave the least RMSE value and highest macro F1-score for regression and classification respectively.
- Ridge Regression : Ridge regression is a regression model where the loss function is the linear least-squares function and regularization is given by the l2-norm. The L2 penalty for ridge regression has been set to 0.1.
- Lasso Regression : Lasso regression is a regression model where the loss function is the linear least-squares function and regularization is given by the l1-norm. The L1 penalty for lasso regression has been set to 0.1.

3.5.2 Classification

The classification task is to predict the 5 class categorical values of G1 and G3.

Reference System for Classification : The reference systems for the classification problem have been devised as follows:

- Trivial System: A system that randomly outputs class labels with probability based on class priors (priors calculated from the training set)
- Nearest Means: Output value is the same as the nearest training-set data point in feature data set.

Classification Models:

- Logistic Regression : The logistic regression model takes a linear equation as input and uses logistic function and log odds to perform a classification task. The models have been evaluated with L1 and L2 regularization as well.
- SVM : Support Vector Classification is implemented with a "one-versus-one" approach for multiclass classification. The hyperplane is determined by implementing different kernels - Linear, RBF, third-degree polynomial, and sigmoid. The best-performing kernel is chosen based on the model. The regularization parameter has been set to 3 for the regularized model.
- Gaussian Classifier: It is based on Laplace approximation and Bayesian methodology. For multi-class classification, several binary one-versus rest classifiers are fitted. The Gaussian Classifier has been trained on an RBF kernel.

- Random Forest Classifier : Random forest is a supervised learning algorithm[7]. The forest is an ensemble of decision trees and uses averaging to improve accuracy and reduce over-fitting. The maximum depth (maxdepth) of the forest has been chosen to be 2.
- MLP : MLPClassifier stands for Multi-layer Perceptron classifier which is a Neural Network. The model has been trained with adam optimizer. The model has been evaluated with L2 regularization as well. The MLP classifier model is trained over 500 iterations. The L2 penalty for the MLP regressor has been set to 0.1.
- Naive Bayes Classifier: It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors[8]. It assumes a feature in a class is unrelated to the presence of any other feature.
- KNN: KNN determines the distances between a new data point and all the data points in the training set, selecting the specified number of neighbors closest to the given data point, then predicts the most frequent label as the classification output. The best value of K has been chosen after experimenting with values of K ranging from 1 to 80 and choosing the one which gave the least RMSE value and highest macro F1-score for regression and classification respectively.

4 Analysis: Comparison of Results, Interpretation

To compare the performance of different regression models the evaluation metric chosen is the RMSE which tells how well a regression model can predict the value while the r2 score tells how well the predictive variables can explain the variation in the output variable. Similarly, the Macro F1 Score is chosen as the evaluation metric for classification models over accuracy as it is a better metric when there are imbalanced classes.

4.1 Mission I

The goal of Mission I is to predict the grade G1 without any prior academic performance The regularization parameter tuning was performed by choosing a range of values and selecting the optimal model which gives the best performance for the corresponding regularization parameter.

4.1.1 Regression

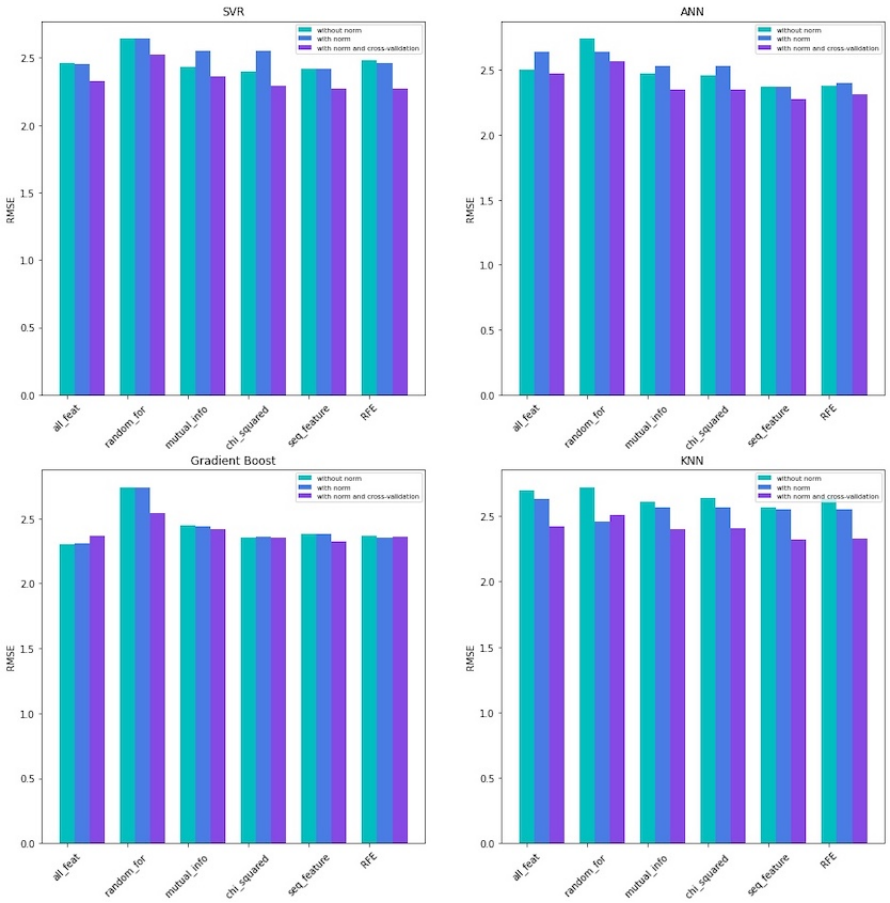

Figure 2: Comparison of regression models based on normalisation and validation for Mission I.

Regression was performed over six models with 5 feature selection methods each along with PCA for dimensionality reduction. Figure 2 shows the performance of the models with and without feature selection, as well as with and without normalizing the dataset. The performance of the models with normalization is better, as can be seen in the graph. As a result, the normalized data is used in subsequent experiments. Figure 2 also includes the performance of the models on normalized data with cross-validation (k = 5) which performs significantly better than the rest.

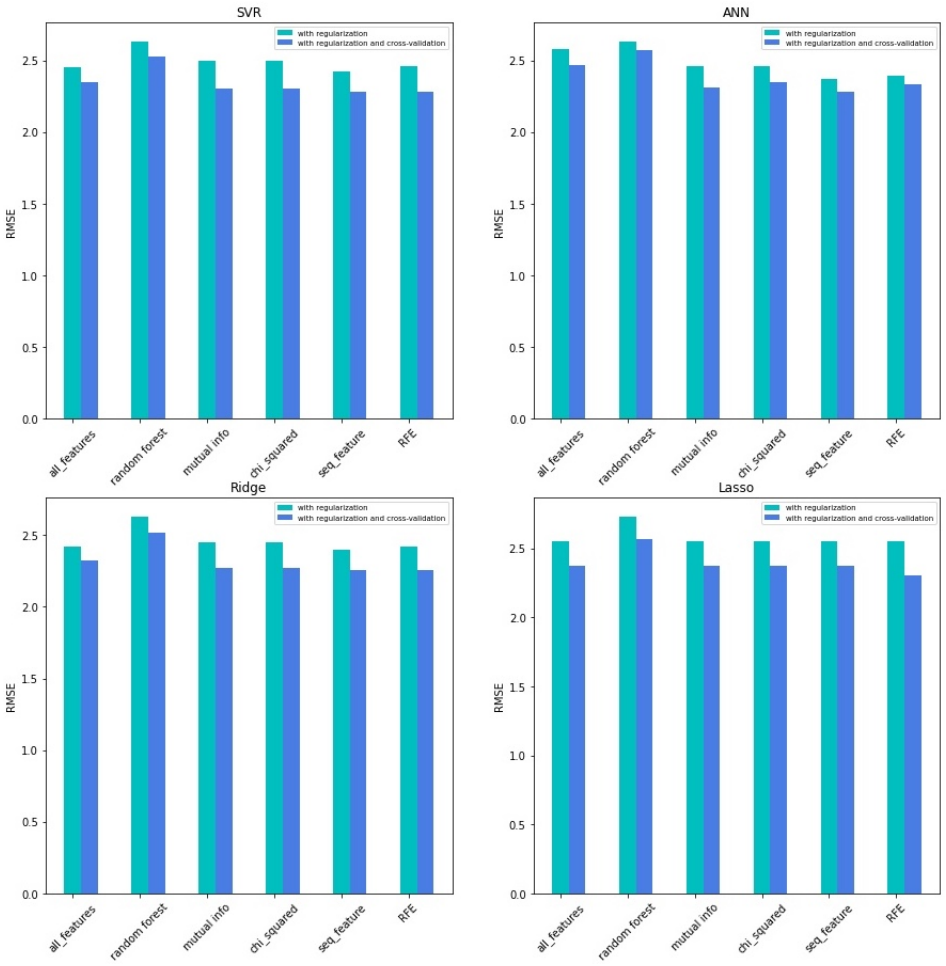

Figure 3: Comparison of regression models based regularization for Mission I.

As inferred from Figure 3 we have performed Regularisation along with cross-validation (k=5) on the models and observed the performance is much better.

Reference Systems	**RMSE**	**MSE**	**R2 Score**
Trivial System	2.84	8.07	-2.02e^{-8}
Linear Regression	2.42	5.86	2.74e^{-1}
One NN	3.22	10.40	-2.90e^{-1}

Table 1: Mission I - Performance of reference system for regression on test dataset.

Models	SVR			ANN			Gradient Boost Reg			KNN		
	RMSE	MSE	r2 score	RMSE	MSE	r2 score	RMSE	MSE	r2 score	RMSE	MSE	r2 score
Without Feature Selection	2.45	6.02	0.25	2.64	6.98	0.13	2.31	5.33	0.34	2.63	6.91	0.14
Random Forest	2.64	6.94	0.14	2.64	6.97	0.14	2.74	7.54	0.06	2.46	7.46	0.07
Mutual Information	2.55	5.49	0.19	2.53	6.44	0.20	2.44	5.96	0.26	2.57	6.62	0.18
Chi Squared	2.55	6.50	0.20	2.53	6.44	0.20	2.36	5.60	0.31	2.57	6.62	0.17
Sequential Feature	2.42	5.87	0.27	2.37	5.65	0.30	2.38	5.67	0.30	2.55	6.54	0.19
RFE	2.46	6.07	0.25	2.40	5.72	0.29	2.35	5.53	0.31	2.55	6.54	0.19
PCA	2.46	6.04	0.25	2.40	5.80	0.28	2.48	6.16	0.24	2.55	6.52	0.19

Table 2: Mission I - Performance of different regression systems with and without feature selection, with normalisation.

Models	SVR			ANN			Ridge Regression			Lasso Regression		
	MSE	RMSE	r2 score	MSE	RMSE	r2 score	MSE	RMSE	r2 score	MSE	RMSE	r2 score
Without Feature Selection	5.59	2.35	0.24	6.13	2.47	0.15	5.42	2.32	0.26	2.37	5.70	0.22
Random Forest	6.46	2.53	0.12	6.67	2.57	0.08	6.38	2.52	0.12	2.57	6.68	0.08
Mutual Information	5.35	2.30	0.26	5.37	2.31	0.26	5.20	2.27	0.29	2.37	5.67	0.22
Chi Squared	5.32	2.30	0.27	5.56	2.35	0.23	5.20	2.27	0.29	2.37	5.69	0.22
Sequential Feature	5.25	2.28	0.28	5.23	2.28	0.28	5.15	2.26	0.30	2.37	5.68	0.22
RFE	5.26	2.28	0.28	5.46	2.33	0.25	5.15	2.26	0.30	2.37	5.69	0.22

Table 3: Mission I - Performance of different regression systems with and without feature selection, with normalisation, with validation and with regularisation.

Reference Systems	**RMSE**	**MSE**	**R2 Score**
Trivial System	2.84	8.07	-2.02e^{-8}
Linear Regression	2.42	5.86	2.74e^{-1}
One NN	3.22	10.40	-2.90e^{-1}
Gradient Boost Regression	2.30	4.60	0.34

Table 4: Mission I - Comparison of best regression model with reference systems.

Initially, the best performing model was found to be Gradient Boost Regressor with Normalised Data Set along with the feature selection method i.e. recursive feature elimination (RFE) which has an r2 score of 0.31 and RMSE of 2.35. The number of features selected by RFE is 15. The parameter, number of boosting stages in Gradient Boosting was set to default as 100. Later hyper-parameter tuning was performed and the number of boosting stages was decreased from 100 to 67 and an r2 score of 0.34 and an RMSE value of 2.30 were achieved.

4.1.2 Classification

The classification was performed over seven different models with 5 feature selection methods each along with two-dimensionality reduction methods PCA and LDA. The performance of the models with and without feature selection was performed with and without normalizing the dataset as shown in Figure 4. It can be observed from the graph that the performance of the models without normalization is much better compared to the other. Thus, for further experiments, the data without normalization was considered. Figure 4 also shows the performance of the models on un-normalized data with cross-validation (k = 5) which did not yield any improvement in performance.

We can extrapolate from Figure 5 that the classification models with regularisation performed better without cross-validation (k=5).

Reference Systems	**Accuracy**	**Macro F1 Score**
Trivial System	0.24	0.21
Baseline Classification	0.27	0.24

Table 5: Mission I - Performance of reference systems for classification on test dataset

Models	Logistic Regression		SVM		Gaussian Classifier		Random Forest Classifier		MLP Classifier		Naive Bayes Classifier		KNN	
	Accuracy	Macro F1 Score	Accuracy	Macro F1 Score	Accuracy	Macro F1 Score	Accuracy	Macro F1 Score	Accuracy	Macro F1 Score	Accuracy	Macro F1 Score	Accuracy	Macro F1 Score
Without Feature Selection	0.35	0.30	0.34	0.28	0.34	0.18	0.37	0.22	0.35	0.29	0.23	0.21	0.33	0.24
Random Forest	0.41	0.32	0.38	0.27	0.41	0.30	0.36	0.22	0.40	0.32	0.20	0.17	0.36	0.25
Mutual Information	0.37	0.27	0.39	0.27	0.33	0.18	0.39	0.27	0.40	0.30	0.19	0.16	0.35	0.28
Chi Squared	0.36	0.30	0.37	0.31	0.31	0.17	0.38	0.24	0.38	0.31	0.22	0.21	0.31	0.25
Sequential Feature	0.36	0.30	0.40	0.32	0.37	0.26	0.37	0.23	0.42	0.35	0.24	0.22	0.39	0.30
RFE	0.41	0.34	0.40	0.32	0.39	0.27	0.36	0.22	0.40	0.34	0.22	0.20	0.37	0.28
PCA	0.35	0.30	0.39	0.30	0.34	0.18	0.32	0.18	0.37	0.30	0.34	0.31	0.33	0.24
FLD	0.32	0.27	0.38	0.30	0.36	0.20	0.36	0.29	0.34	0.28	0.35	0.30	0.34	0.28

Table 6: Mission I - Performance of different classification systems with and without feature selection, without normalisation.

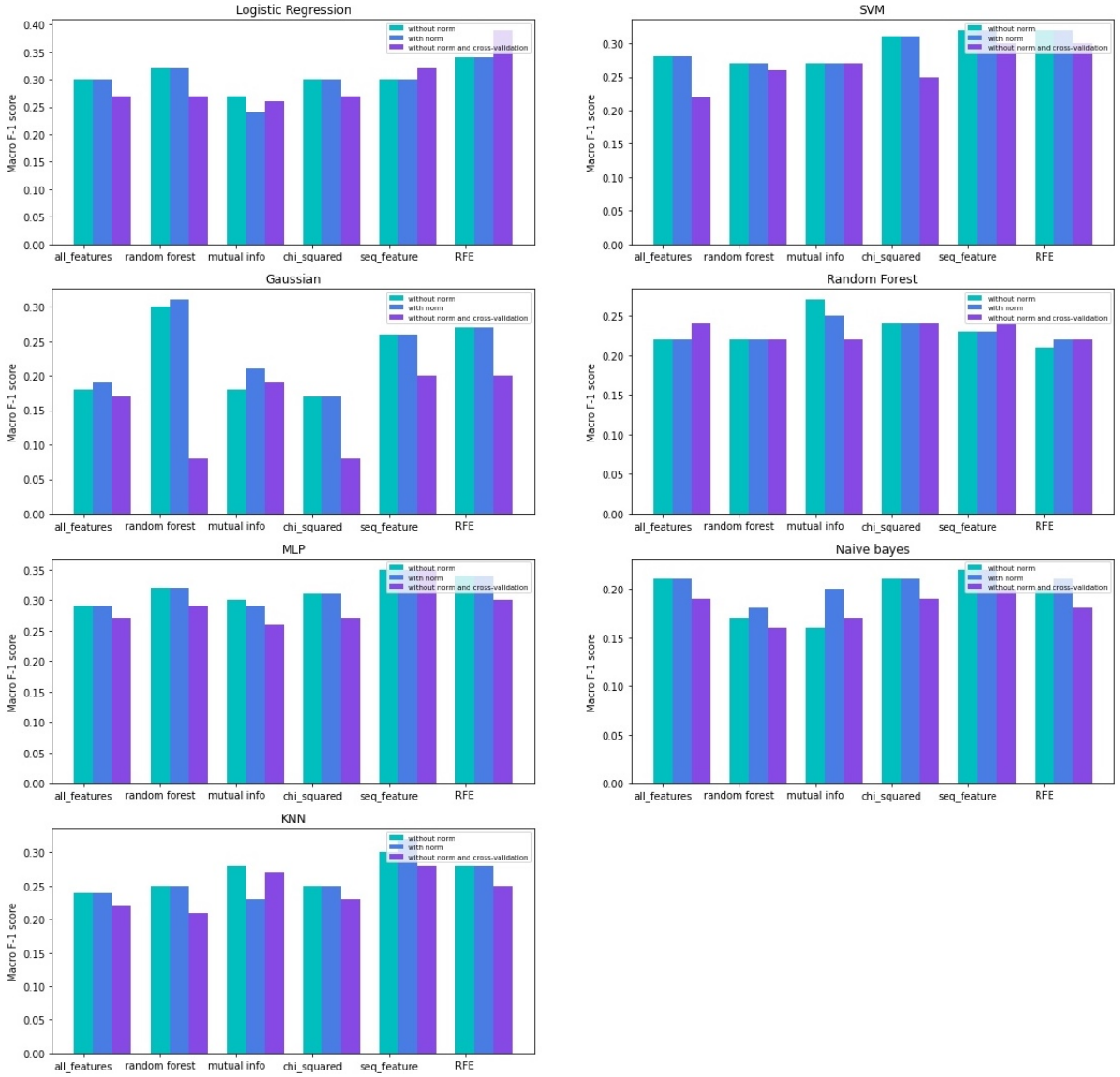

Figure 4: Comparison of classification models based on normalisation and validation for Mission I.

Models	Logistic Regression - L1		Logistic Regression - L2		SVM		MLP Classifier	
	Accuracy	Macro F1 Score	Accuracy	Macro F1 Score	Accuracy	Macro F1 Score	Accuracy	Macro F1 Score
Without Feature Selection	0.36	0.30	0.34	0.29	0.37	0.29	0.32	0.27
Random Forest	0.39	0.31	0.38	0.30	0.39	0.29	0.32	0.26
Mutual Information	0.39	0.32	0.36	0.30	0.37	0.30	0.36	0.30
Chi Squared	0.40	0.33	0.37	0.31	0.36	0.30	0.33	0.30
Sequential Feature	0.39	0.32	0.39	0.32	0.40	0.37	0.41	0.37
RFE	0.38	0.31	0.38	0.31	0.37	0.31	0.37	0.30
PCA	0.37	0.31	0.34	0.29	0.36	0.30	0.33	0.30
FLD	0.34	0.28	0.33	0.27	0.35	0.29	0.36	0.29

Table 7: Mission I - Performance of different classification systems with and without feature selection, without normalisation, without validation and with regularisation.

The best performing model was found to be MLP Classifier whose confusion matrix is as shown in Figure 6. A regularisation parameter of 0.1 and maximum iteration of 300 with a sequential feature selection method that selected 10 features. The best model is compared with the reference system and the model achieved an accuracy of 0.41 and a macro f1 score of 0.37 as shown in Table 9.

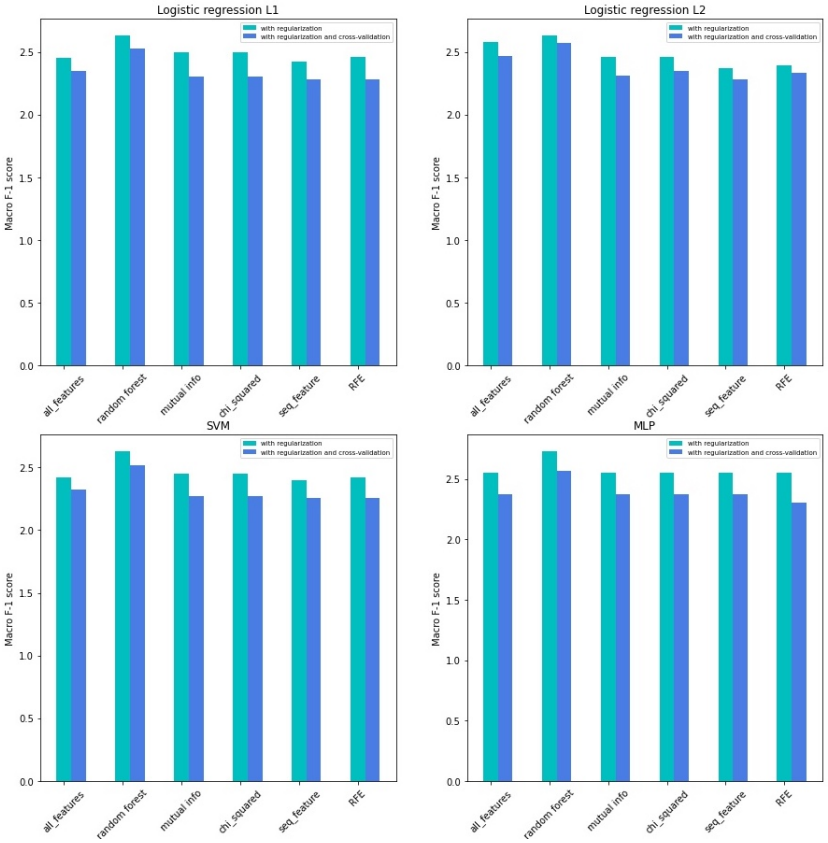

Figure 5: Comparision of classification models based on regularization for Mission I.

Reference Systems	**Accuracy**	**Macro F1 Score**
Trivial System	0.24	0.21
Baseline Classification	0.27	0.24
MLP Classifier	0.41	0.37

Table 8: Mission I - Comparison of best classification model with reference systems.

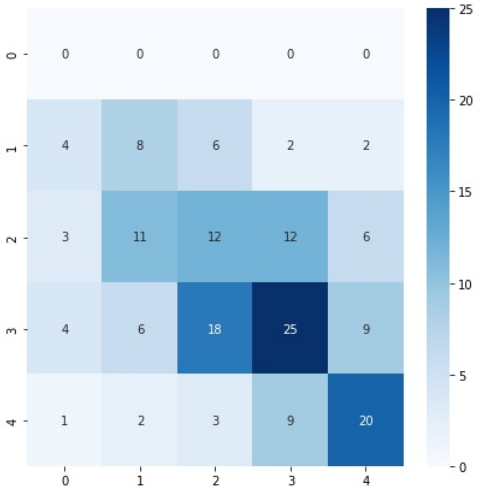

Figure 6: Confusion Matrix of Mission-I Classification.

4.2 Mission II

The goal of Mission II is to predict the grade G1 without any prior academic performance.

4.2.1 Regression

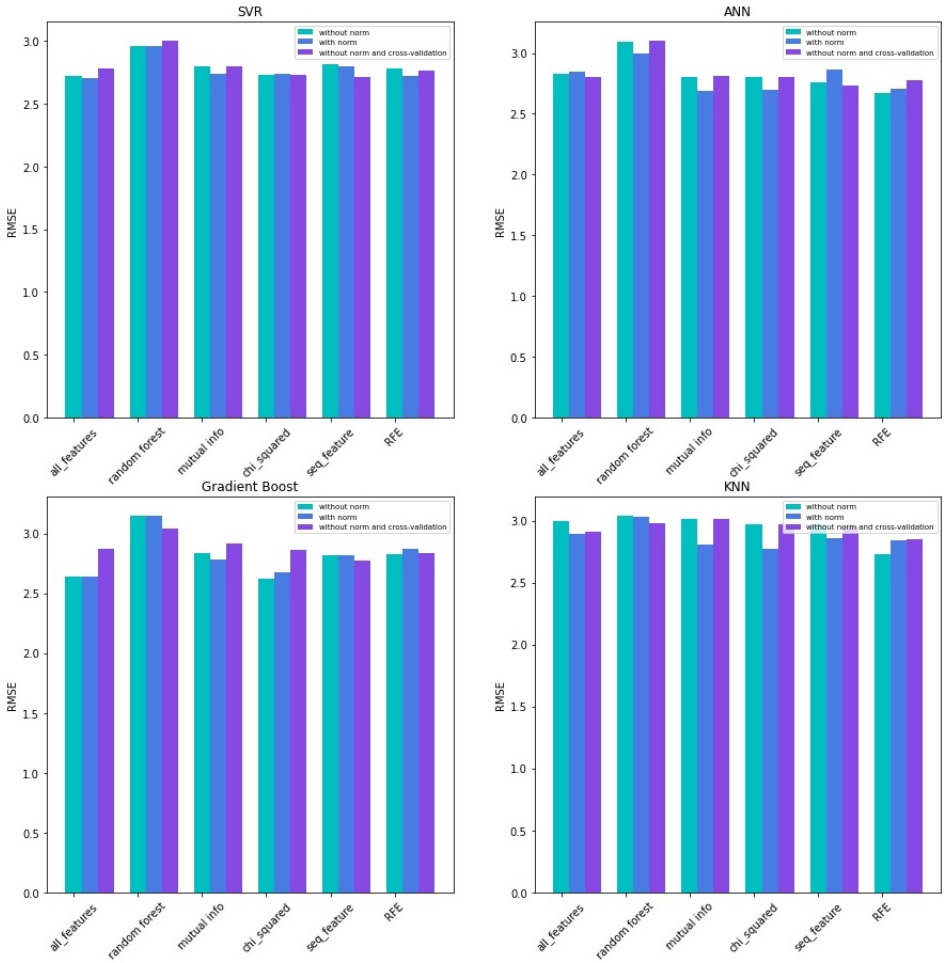

Figure 7: Comparison of regression models based on normalisation and validation for Mission II.

Regression was performed over six models with 5 feature selection methods each along with PCA for dimensionality reduction. Figure 7 shows the performance of the models with and without feature selection, as well as with and without normalizing the dataset. The performance of the models using unnormalized data is better, as can be seen in the graph. Thus, for further experiments, unnormalized data is used. Figure 5 also includes the performance of the models on unnormalized data with cross-validation (k = 5) which performs significantly better than the rest.

As seen in Figure 8 we have performed experiments on the models with Regularisation and cross-validation (k=5) and it can be inferred that the model performs better with only regularisation.

Reference Systems	**RMSE**	**MSE**	**R2 Score**
Trivial System	3.17	10.07	-2.6e^{-4}
Linear Regression	2.71	7.36	0.27
One NN	3.63	13.23	-0.31

Table 9: Mission II - Performance of reference systems for regression on test dataset.

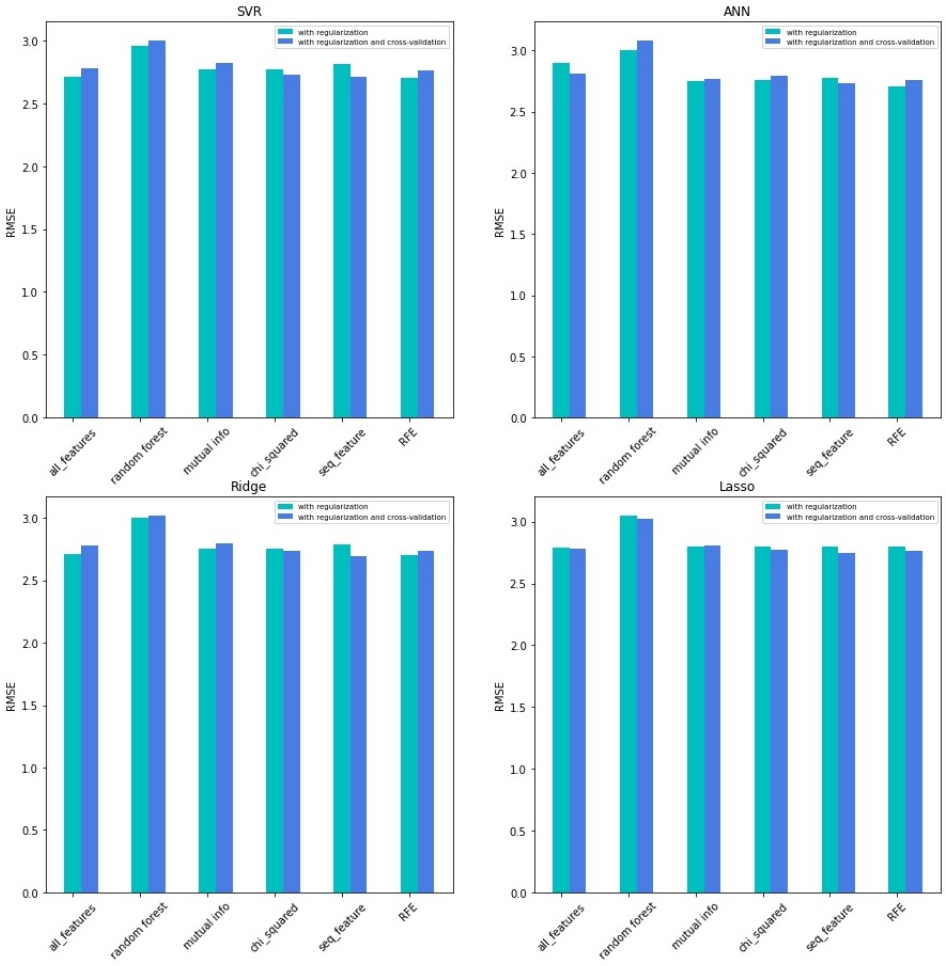

Figure 8: Comparison of regression models based on regularization for Mission II.

Models	SVR			ANN			Gradient Boost Reg			KNN		
	RMSE	MSE	r2 score	RMSE	MSE	r2 score	RMSE	MSE	r2 score	RMSE	MSE	r2 score
Without Feature Selection	2.72	7.40	0.266	2.83	8.04.	0.20	2.64	6.97	0.31	3.00	8.96	0.11
Random Forest	2.96	8.77	0.13	3.09	9.59	0.05	3.15	9.92	0.02	3.04	9.26	0.08
Mutual Information	2.80	7.83	0.22	2.80	7.86	0.22	2.84	8.06	0.20	3.01	9.09	0.10
Chi Squared	2.73	7.46	0.26	2.8	7.85	0.22	2.62	6.85	0.32	2.97	8.82	0.12
Sequential Feature	2.81	7.90	0.21	2.76	7.62	0.24	2.82	7.98	0.20	2.97	8.84	0.12
RFE	2.78	7.77	0.22	2.67	7.16	0.29	2.83	8.04	0.20	2.73	7.48	0.26
PCA	2.72	7.40	0.26	3.20	10.24	-0.01	2.97	8.83	0.12	3.00	8.93	0.11

Table 10: Mission II - Performance of different regression systems with and without feature selection and without normalisation.

Models	SVR			ANN			Ridge Regression			Lasso Regression		
	MSE	RMSE	r2 score	MSE	RMSE	r2 score	MSE	RMSE	r2 score	MSE	RMSE	r2 score
Without Feature Selection	2.78	7.86	0.25	2.81	7.96	0.24	2.78	7.84	0.25	2.78	7.85	0.26
Random Forest	3.00	9.16	0.13	3.08	9.59	0.08	3.02	9.21	0.12	3.02	9.22	0.12
Mutual Information	2.82	8.08	0.23	2.77	7.84	0.26	2.80	7.98	0.24	2.81	8.03	0.24
Chi Squared	2.73	7.55	0.28	2.79	7.81	0.25	2.74	7.59	0.28	2.77	7.80	0.26
Sequential Feature Selection	2.71	7.47	0.29	2.73	7.54	0.28	2.69	7.35	0.30	2.75	7.71	0.27
RFE	2.76	7.74	0.26	2.76	7.70	0.26	2.74	7.58	0.27	2.76	7.75	0.26

Table 11: Mission II - Performance of different regression systems with and without feature selection, without normalisation, with validation and with regularisation.

Reference Systems	RMSE	MSE	R2 Score
Trivial System	3.17	10.07	-2.6e^{-4}
Linear Regression	2.71	7.36	0.27
One NN	3.63	13.23	-0.31
Gradient Boost Regression	2.62	6.85	0.32

Table 12: Mission II - Comparison of best regression model with reference systems.

The best performing model was found to be Gradient Boost Regressor with unnormalized Data Set along with the chi-squared feature selection method which gave an r2 score of 0.32 and an RMSE value of 2.62. The number of features selected by chi-squared is 16. The parameter number of boosting stages in Gradient Boosting was set to default as 100. Later hyper-parameter tuning was performed and this was found to be the best model.

4.2.2 Classification

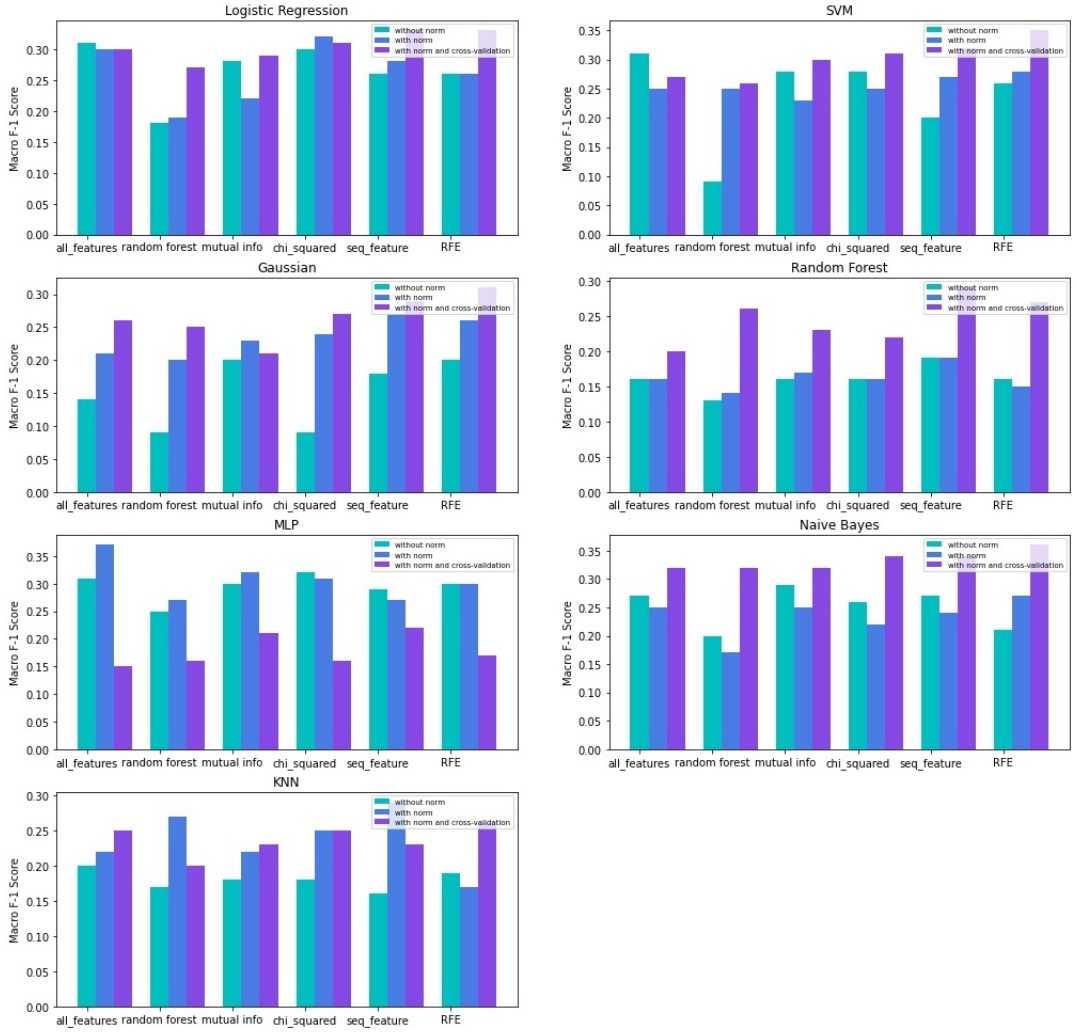

Figure 9: Comparision of classification models based on normalisation and validation for Mission II.

Figure 9 shows that the performance of different classification models gives a higher macro f1 score with normalization as compared to without normalization. 5-fold cross-validation was performed on the models with normalized data and its macro f1 score is almost comparable with that of the normalized dataset.

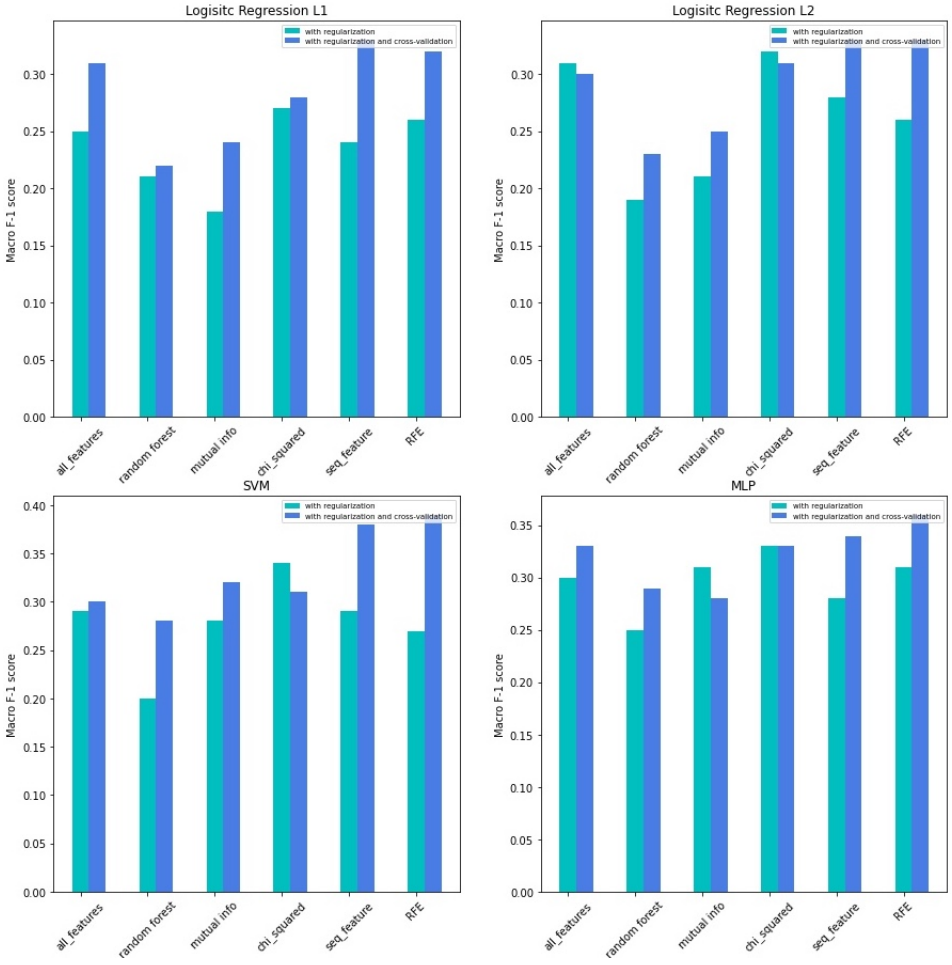

Figure 10: Comparision of classification models based on regularization for Mission II.

It can be inferred from Figure 10 that 5 fold cross-validation significantly improves the performance of the regularised models.

Reference Systems	**Accuracy**	**Macro F1 Score**
Trivial System	0.22	0.19
Baseline Classification	0.28	0.27

Table 13: Mission II - Performance of reference systems for classification on test dataset.

Models	Logistic Regression		SVM		Gaussian Classifier		Random Forest Classifier		MLP Classifier		Naive Bayes Classifier		KNN	
	Accuracy	Macro F1 Score	Accuracy	Macro F1 Score	Accuracy	Macro F1 Score	Accuracy	Macro F1 Score	Accuracy	Macro F1 Score	Accuracy	Macro F1 Score	Accuracy	Macro F1 Score
Without Feature Selection	0.34	0.30	0.36	0.25	0.31	0.21	0.32	0.16	0.40	0.37	0.25	0.25	0.34	0.22
Random Forest	0.27	0.19	0.30	0.25	0.28	0.21	0.29	0.16	0.26	0.24	0.24	0.17	0.28	0.15
Mutual Information	0.28	0.19	0.30	0.20	0.28	0.19	0.28	0.14	0.30	0.27	0.25	0.23	0.31	0.20
Chi Squared	0.37	0.32	0.32	0.25	0.36	0.24	0.31	0.16	0.31	0.31	0.23	0.22	0.32	0.25
Sequential Feature Selection	0.35	0.28	0.31	0.27	0.33	0.27	0.34	0.20	0.30	0.27	0.26	0.24	0.31	0.30
RFE	0.31	0.26	0.32	0.28	0.32	0.26	0.32	0.15	0.33	0.30	0.28	0.27	0.27	0.17
PCA	0.33	0.30	0.36	0.25	0.31	0.21	0.33	0.15	0.32	0.31	0.34	0.34	0.34	0.22
FLD	0.32	0.27	0.33	0.23	0.33	0.24	0.34	0.29	0.31	0.29	0.36	0.34	0.29	0.26

Table 14: Mission II : Performance of different classification systems with and without feature selection and with normalisation.

Models	Logistic Regression		SVM		Gaussian Classifier		Random Forest Classifier		MLP Classifier		Naive Bayes Classifier		KNN	
	Accuracy	Macro F1 Score	Accuracy	Macro F1 Score	Accuracy	Macro F1 Score	Accuracy	Macro F1 Score	Accuracy	Macro F1 Score	Accuracy	Macro F1 Score	Accuracy	Macro F1 Score
Without Feature Selection	0.34	0.30	0.37	0.27	0.34	0.26	0.33	0.20	0.31	0.15	0.35	0.32	0.26	0.25
Random Forest	0.37	0.27	0.38	0.26	0.37	0.25	0.36	0.26	0.31	0.16	0.35	0.32	0.27	0.20
Mutual Information	0.35	0.29	0.37	0.30	0.33	0.21	0.33	0.23	0.34	0.21	0.37	0.32	0.24	0.23
Chi Squared	0.36	0.31	0.36	0.31	0.36	0.27	0.33	0.22	0.31	0.16	0.26	0.34	0.27	0.25
Sequential Feature Selection	0.39	0.33	0.40	0.32	0.38	0.29	0.37	0.29	0.35	0.22	0.36	0.34	0.25	0.23
RFE	0.37	0.33	0.40	0.35	0.40	0.31	0.37	0.27	0.34	0.17	0.38	0.36	0.27	0.26

Table 15: Mission II - Performance of different classification systems with and without feature selection, with normalisation and with validation.

Reference Systems	**Accuracy**	**Macro F1 Score**
Trivial System	0.22	0.19
Baseline Classification	0.28	0.27
SVM Classifier	0.40	0.39

Table 16: Mission II - Comparison of best classification model with the reference systems.

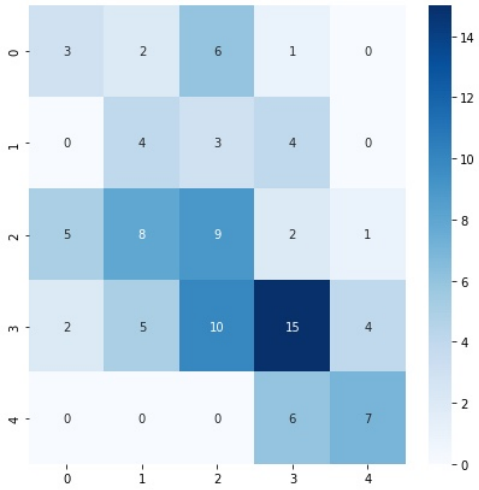

Figure 11: Confusion Matrix of Mission-II Classification.

The best performing model for mission II classification is normalized data with the RFE feature selection method whose confusion matrix is shown in Figure 11. This model was given as the input to SVM along with a 5-fold cross-validation which was also regularised. The SVM model experimented with different kernels - Linear, RBF, Polynomial, and Sigmoid. Among these Linear gave the highest accuracy. As seen in Table 16. the model gave an accuracy of 0.40 and a macro f1 score of 0.39 which was the highest.

4.3 Mission III

The goal of Mission III is to predict the grade G3 with prior academic performance G1 and G2.

4.3.1 Regression

Regression for mission III was performed over six different models. Figure 12 shows that the performance of different regression models gives lesser RMSE values without normalization as compared to normalization. Five-fold cross-validation was performed on the models with unnormalized data and as seen in the figure there was an increase in the RMSE values.

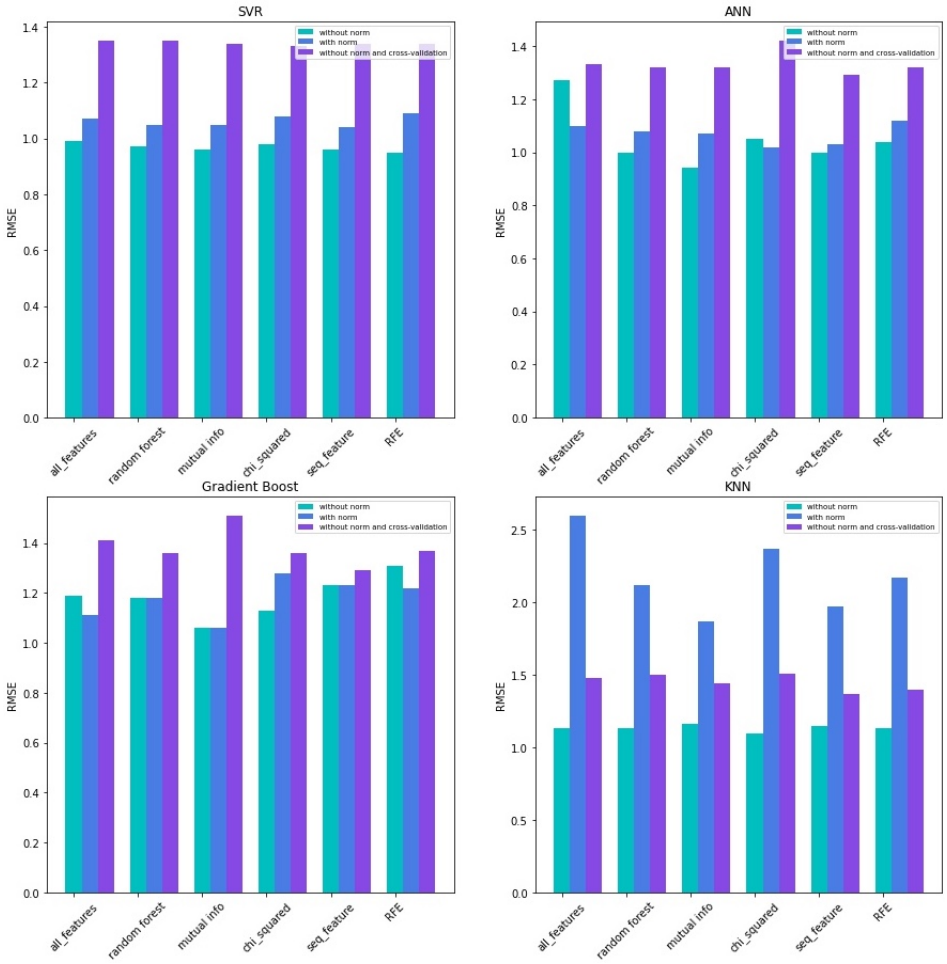

Figure 12: Comparison of regression models based on normalisation and validation for Mission III.

As observed in Figure 10 five-fold cross-validation increases the RMSE value on the regularised regression models.

Reference Systems	**RMSE**	**MSE**	**R2 Score**
Trivial System	3.17	10.08	-2.6e^{-4}
Linear Regression	1.01	1.02	0.90
One NN	1.57	2.46	0.76

Table 17: Mission III - Performance of reference systems for regression on test dataset.

Models	SVR			ANN			Gradient Boost Regressor			KNN		
	RMSE	MSE	r2 score	RMSE	MSE	r2 score	RMSE	MSE	r2 score	RMSE	MSE	r2 score
Without Feature Selection	0.99	0.99	0.9	1.27	1.63	0.84	1.19	1.42	0.86	1.13	1.27	0.87
Random Forest	0.97	0.95	0.90	1.00	1.02	0.90	1.18	1.39	0.86	1.13	1.28	0.87
Mutual Information	0.96	0.96	0.90	0.94	0.88	0.91	1.06	1.13	0.89	1.16	1.36	0.86
Chi Squared	0.98	0.96	0.90	1.05	1.10	0.89	1.13	1.28	0.87	1.10	1.22	0.88
Sequential Feature Selection	0.96	0.93	0.91	1.00	1.01	0.90	1.23	1.51	0.85	1.15	1.32	0.86
RFE	0.95	0.91	0.91	1.04	1.09	0.89	1.31	1.72	0.83	1.13	1.27	0.87
PCA	0.99	0.99	0.90	1.62	2.63	0.73	1.15	1.32	0.87	1.13	1.28	0.87

Table 18: Mission III - Performance of different regression systems with and without feature selection and without normalisation.

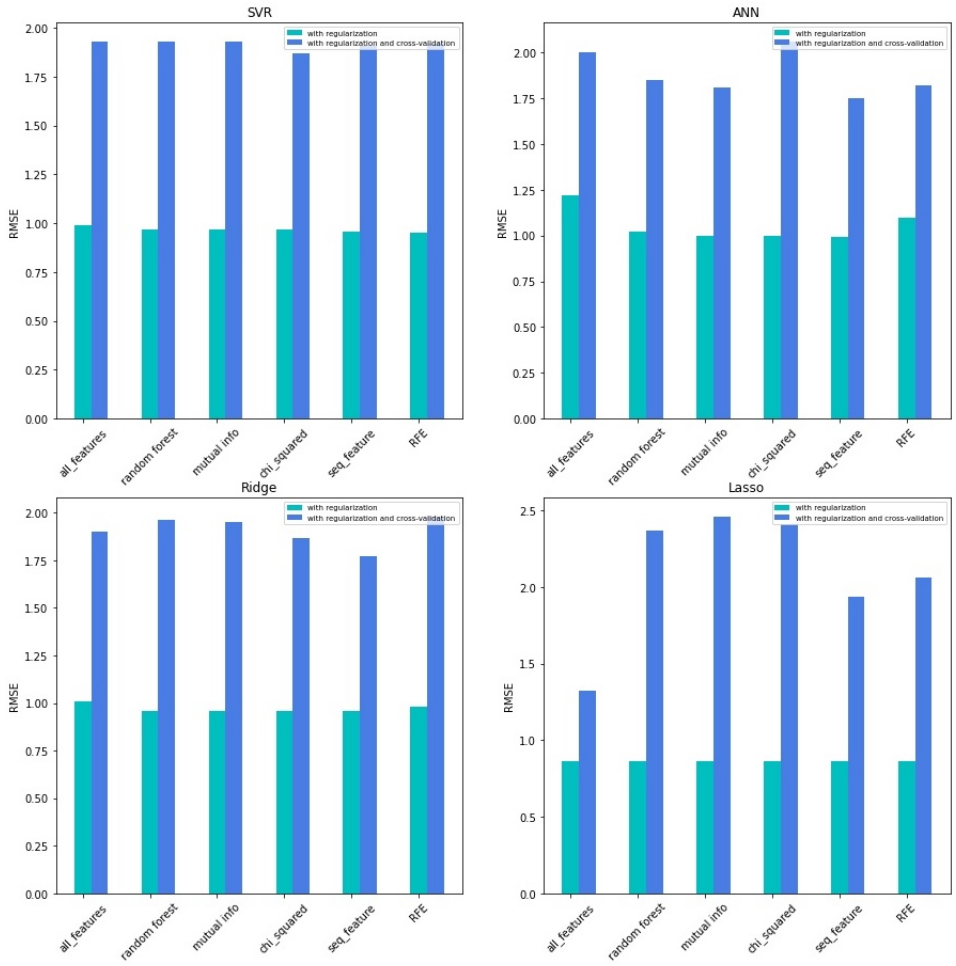

Figure 13: Comparision of regression models based on regularization for Mission III.

Models	SVR			ANN			Ridge Regression			Lasso Regression		
	RMSE	MSE	r2 score	RMSE	MSE	r2 score	RMSE	MSE	r2 score	RMSE	MSE	r2 score
Without Feature Selection	0.99	1.00	0.90	1.22	1.50	0.85	1.01	1.02	0.90	0.86	0.93	0.91
Random Forest	0.97	0.95	0.90	1.02	1.05	0.89	0.96	0.93	0.90	0.86	0.92	0.91
Mutual Information	0.97	0.94	0.90	1.00	0.99	0.90	0.96	0.92	0.91	0.86	0.93	0.91
Chi Squared	0.97	0.94	0.91	1.00	1.00	0.90	0.96	0.92	0.91	0.86	0.93	0.91
Sequential Feature Selection	0.96	0.93	0.91	0.99	0.99	0.90	0.96	0.92	0.91	0.86	0.93	0.91
RFE	0.95	0.90	0.91	1.10	1.22	0.88	0.98	0.97	0.90	0.86	0.93	0.91
PCA	1.00	1.00	0.90	1.60	2.60	0.74	1.01	1.03	0.90	0.90	0.95	0.91

Table 19: Mission III - Performance of different regression systems with and without feature selection, without normalisation, with regularisation and without validation.

Table 20: Comparison of Best Regression Model with Reference System Mission III.

Reference Systems	**RMSE**	**MSE**	**R2 Score**
Trivial System	3.17	10.08	$-2.6e^{-4}$
Linear Regression	1.01	1.02	0.90
One NN	1.57	2.46	0.76
ANN	0.94	0.88	0.91

The best performing model was found to be MLP Regressor with an unnormalized dataset along with the Mutual Information feature selection method where 10 features were selected. The number of iterations for the MLP Regressor was set to 500. As seen in Table 20 the model gave an r2 score of 0.91 and an RMSE value of 0.94 which is the highest. The above mentioned hyperparameters of the model were found to be the best after hyper tuning as well.

4.3.2 Classification

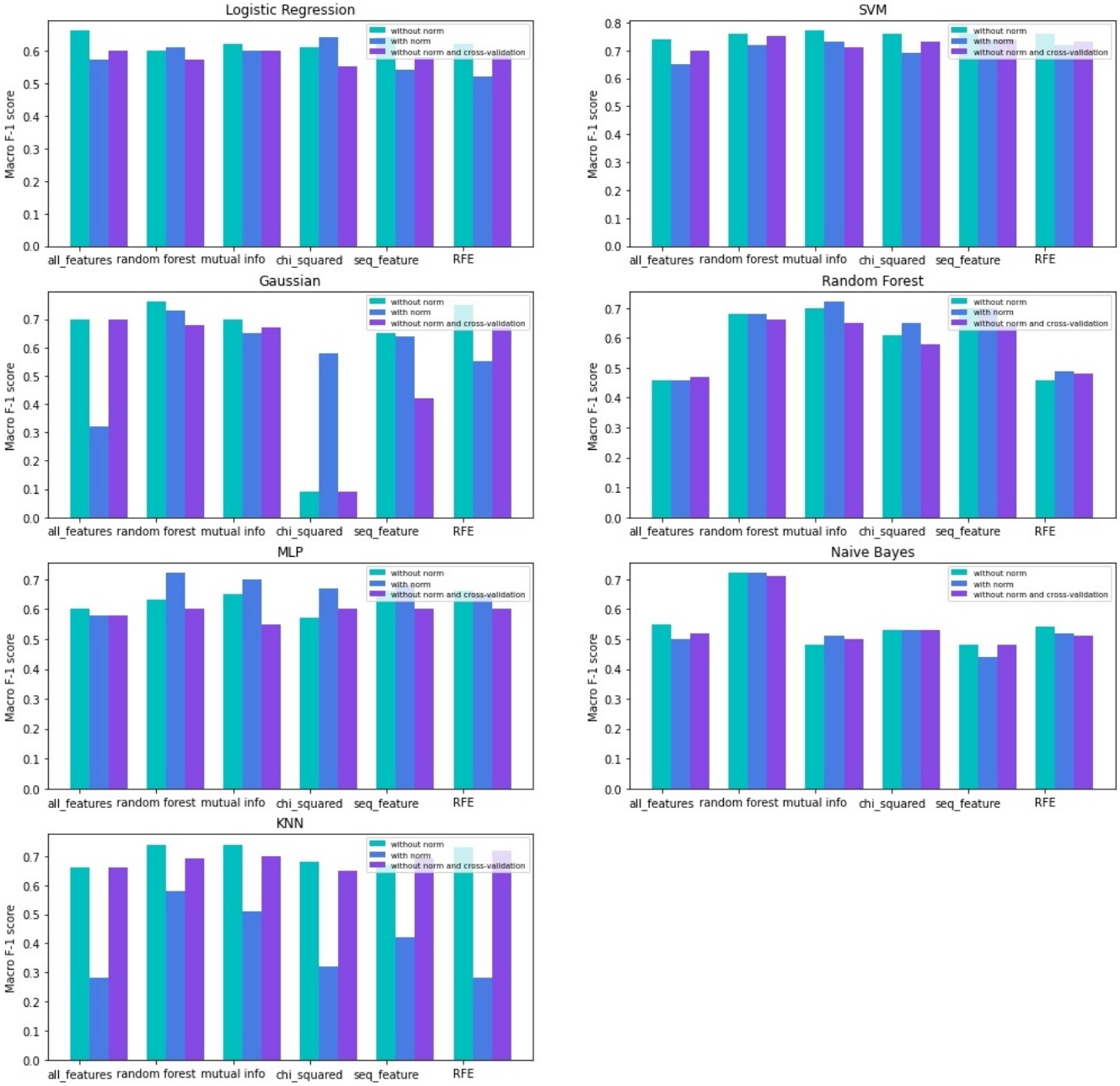

Figure 14: Comparison of classification models based on normalisation and validation for Mission III.

It can be observed from Figure 14 that the performance of the models without normalization is slightly better compared to the normalized data. Thus, for further experiments, the data without normalization was considered. Figure 11 also shows the performance of the models on un-normalized data with cross-validation (k = 5) and its performance is almost comparable with the un-normalized data.

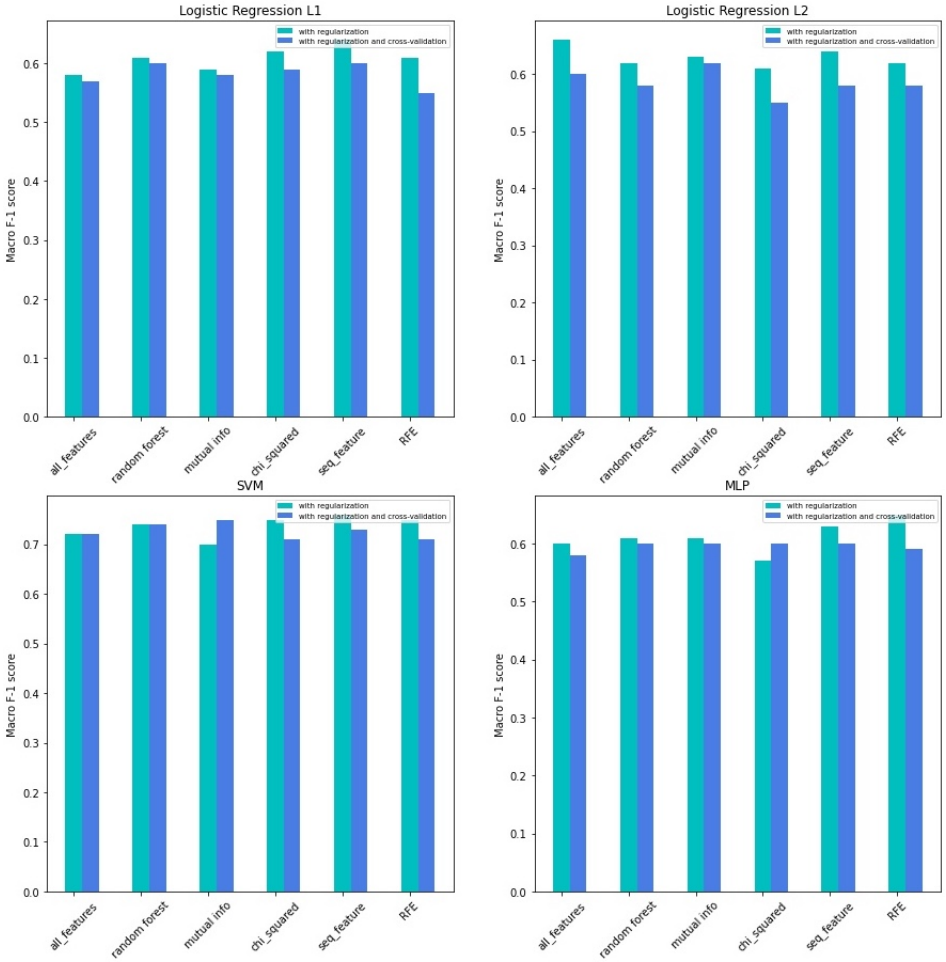

Figure 15: Comparison of classification models based on regularization for Mission III.

Figure 15 shows that the classification models with regularisation perform better compared to classification models with regularisation and five-fold cross-validation.

Reference Systems	**Accuracy**	**Macro F1 Score**
Trivial System	0.23	0.20
Baseline Classification	0.60	0.62

Table 21: Mission III - Performance of the classification reference systems on test data set.

Models	Logistic Regression		SVM		Gaussian Classifier		Random Forest Classifier		MLP Classifier		Naive Bayes Classifier		KNN	
	Accuracy	Macro F1 Score	Accuracy	Macro F1 Score	Accuracy	Macro F1 Score	Accuracy	Macro F1 Score	Accuracy	Macro F1 Score	Accuracy	Macro F1 Score	Accuracy	Macro F1 Score
Without Feature Selection	0.66	0.66	0.75	0.74	0.70	0.70	0.58	0.46	0.58	0.60	0.53	0.55	0.67	0.66
Random Forest	0.60	0.60	0.76	0.76	0.75	0.76	0.72	0.68	0.62	0.63	0.72	0.72	0.74	0.74
Mutual Information	0.62	0.62	0.77	0.77	0.71	0.70	0.73	0.70	0.66	0.65	0.48	0.48	0.75	0.74
Chi Squared	0.61	0.61	0.75	0.76	0.31	0.09	0.67	0.61	0.57	0.57	0.51	0.53	0.70	0.68
Sequential Feature Selection	0.64	0.64	0.76	0.76	0.67	0.65	0.73	0.70	0.66	0.66	0.55	0.48	0.67	0.67
RFE	0.61	0.62	0.77	0.76	0.75	0.75	0.58	0.46	0.64	0.66	0.51	0.54	0.75	0.73
PCA	0.72	0.71	0.73	0.72	0.70	0.70	0.45	0.36	0.63	0.63	0.52	0.52	0.65	0.64
FLD	0.77	0.75	0.75	0.74	0.75	0.73	0.64	0.61	0.72	0.71	0.72	0.71	0.74	0.72

Table 22: Mission III - Performance of different classification systems with and without feature selection and without normalisation.

Models	Logistic Regression - L1		Logistic Regression - L1		SVM		MLP Classifier	
	Accuracy	Macro F1 Score	Accuracy	Macro F1 Score	Accuracy	Macro F1 Score	Accuracy	Macro F1 Score
Without Feature Selection	0.56	0.57	0.61	0.60	0.73	0.72	0.60	0.58
Random Forest	0.62	0.60	0.59	0.58	0.75	0.74	0.61	0.60
Mutual Information	0.60	0.58	0.63	0.62	0.75	0.75	0.61	0.60
Chi Squared	0.59	0.59	0.56	0.55	0.72	0.71	0.61	0.60
Sequential Feature Selection	0.61	0.60	0.59	0.58	0.74	0.73	0.61	0.60
RFE	0.57	0.55	0.59	0.58	0.73	0.71	0.60	0.59

Table 23: Mission III - Performance of different classification systems with and without feature selection, without normalisation, with validation and without regularisation.

Reference Systems	**Accuracy**	**Macro F1 Score**
Trivial System	0.23	0.20
Baseline Classification	0.60	0.62
SVM	0.77	0.77

Table 24: Mission III - Comparison of best classification model with reference systems.

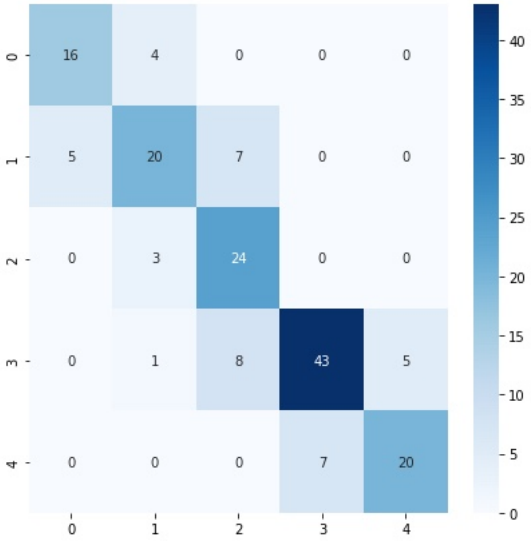

Figure 16: Confusion Matrix of Mission-III Classification.

The best model for Mission III as inferred from Table 24 is the Support Vector Machine (SVM). Unnormalized data with the best 10 features using mutual information was given as the input to SVM. The SVM model was experimented with different kernels - Linear, RBF, Polynomial, and Sigmoid. Among these Linear gave the highest accuracy. As seen in Table 24. the model gave an accuracy of 0.77 and a macro f1 score of 0.77 which was the highest.

5 Libraries used and what you coded yourself

The coding of the project is done in Python. The libraries used are pandas[2], NumPy[3], Matplotlib[4], seaborn[5], and sklearn [6]. The pandas library is used to read the data frame. Matplotilb and seaborn are used to plot the graphs and confusion matrices. During the data preprocessing stage, label encoding of the dataset is performed using the sklearn label encoder.The regression and classification models are implemented using sklearn modules.

6 Contributions of each team member

The contributions of team member Sanjana Vasudeva are as follows:

- Label Encoding and normalization
- Feature Selection methods
- Feature dimensionality reduction methods
- Classification models
- Classification models with regularization
- Cross validation for Classification models

The contributions of team member Vaishnavi Channakeshava are as follows:

- Label Encoding and normalization
- Feature Selection methods
- Feature dimensionality reduction methods
- Regression models
- Regression models with regularization
- Cross validation for Regression models

7 Summary and conclusions

The goal of the project was to predict first-period academic performance and final-period academic performance with and without prior academic performance.

For the Mission I regression problem, the best model was Gradient Boost Regressor with Normalised Data Set with recursive feature elimination (RFE) for features selection and an r2 score of 0.34 and RMSE of 2.30 were achieved.

For the Mission I classification problem, the best model has regularized MLP Classifier with Sequential feature eliminated dataset and an accuracy of 0.41 and a macro F1 score of 0.37 was achieved.

For the Mission II regression problem, the best model was Gradient Boost Regressor with features selected using chi-squared, and an r2 score of 0.32 and RMSE of 2.62 were achieved.

For the Mission II classification problem, the best model was SVM with a linear kernel with features selected using RFE, and an accuracy of 0.40 and a macro F1 score of 0.39 were achieved.

For the Mission III regression problem, the best model was MLP Regressor with mutual information for features selection, and an r2 score of 0.91 and RMSE of 0.94 were achieved.

For the Mission III classification problem, the best model was SVM with a linear kernel with features selected using mutual information and an accuracy of 0.77, and a macro F1 score of 0.77 was achieved.

From the above results, we can conclude that the feature set containing prior academic performance(G1 and G2) improved the performance of both regression and classification problems. The number of features required for mission 3 was lesser compared to the other two missions. After analyzing the features selected, the following features were found to have more significance in the prediction - school, schoolsup, sex, higher, Medu, and failures.

References

[1] *Student Performance Dataset* , available at `https://archive.ics.uci.edu/ml/datasets/student+performance`

[2] *Pandas library in python* `https://pandas.pydata.org`

[3] *NumPy library in python* `https://numpy.org`

[4] *Matplotlib library in python* `https://matplotlib.org`

[5] *Seaborn library in python* https://seaborn.pydata.org

[6] *Scikit-learn library in python* https://scikit-learn.org/stable/

[7] *Naive Bayes Classifier*, available at https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

[8] *Scikit-learn library in python* https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/