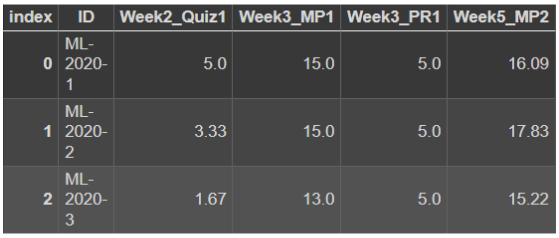
Mini Project II – Students Performance

Introduction

This assignment aims to predict students' final grades in an online machine learning course using different supervised machine learning models. The dataset from Moodle contained anonymized information relating to 107 enrolled students. This project explores three different models and proceeds to compare the performance of the best and the worst model. Additionally, It Identifies the most important features contributing to final grade prediction. Through this analysis, the aim is to determine which models best capture student behavior and performance patterns.

Step 1: Data Processing

The dataset contains records of 107 students enrolled in a course, with features representing student performance and engagement.



index	ID	Week2_Quiz1	Week3_MP1	Week3_PR1	Week5_MP2
0	ML-2020-1	5.0	15.0	5.0	16.09
1	ML-2020-2	3.33	15.0	5.0	17.83
2	ML-2020-3	1.67	13.0	5.0	15.22

Figure: DataFrame

The data includes:

- **Grades**: Weekly scores (Week2_Quiz1, Week3_MP1, Week3_PR1...) from quizzes, mini-projects, peer reviews, and an overall total grade.
- **Activity Logs**: Engagement activities (Week1_Stat0, Week1_Stat1..) categorized by week and type (content-related, assignment-related, grade-related, and forum-related).

Missing Values

The dataset was complete and ready for analysis as no missing values were found.

Feature Selection

We utilized all available features in the dataset except for the student ID column. The key features included:

- Grades, which indicated a student's performance in assessments, were essential in predicting the final grade.
- Activity logs gave insight into student engagement and behavior.

Given that student performance could be influenced by various factors, all features were kept to capture the full breadth of student behavior and performance.

Step 2: Data Split

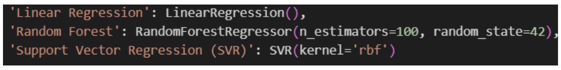
The data was split into training and testing sets:

- **Training Set**: 80% of the data used to train the models.
- **Test Set**: 20% of the data reserved for testing and evaluating the models' performance on unseen data.

Step 3: Model Training

Three supervised learning models were chosen to predict students' final grades:

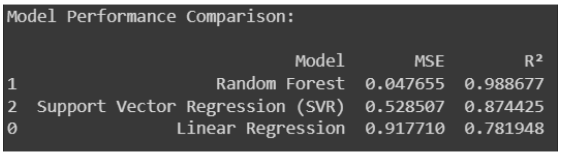
1. Linear Regression
2. Random Forest
3. Support Vector Regression (SVR)



```
'Linear Regression': LinearRegression(),
'Random Forest': RandomForestRegressor(n_estimators=100, random_state=42),
'Support Vector Regression (SVR)': SVR(kernel='rbf')
```

Figure: Chosen Models

Performance Metrics



```
Model Performance Comparison:

                             Model       MSE        R²
1                    Random Forest  0.047655  0.988677
2  Support Vector Regression (SVR)  0.528507  0.874425
0                Linear Regression  0.917710  0.781948
```

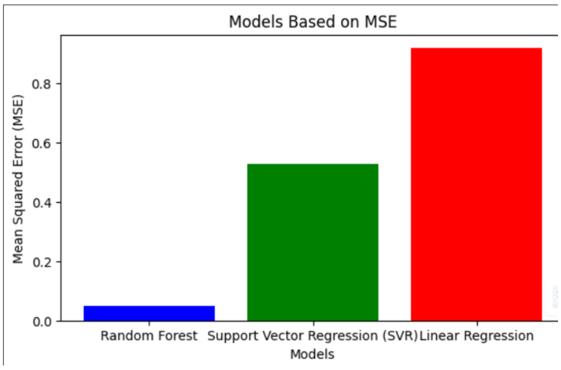


Figure: Model Comparison

The performance of each model was evaluated using:

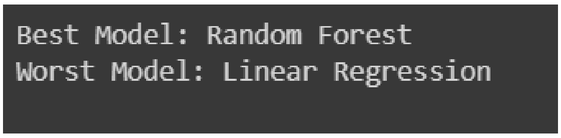
- **Mean Squared Error (MSE)**: Measures the average squared difference between predicted and true values.
- **R-squared (R²)**: Indicates the proportion of the variance in the target variable explained by the model.

The two models with the best and worst performance were selected based on these metrics.

Step 4: Performance Evaluation

Model Comparison

After evaluating all three models, the **Random Forest (Best)** and **Linear Regression (Worst)** models were selected by comparing MSE.

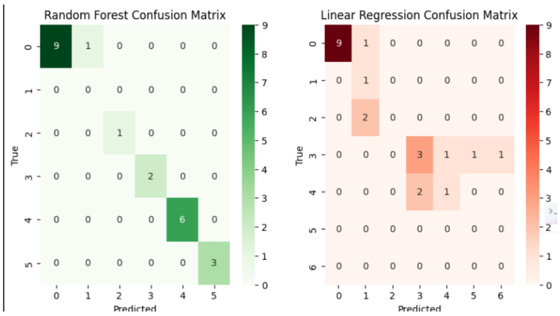


```
Best Model: Random Forest
Worst Model: Linear Regression
```

Figure: Two models

Visualization of Model Performance

Confusion matrix and classification reports were generated. Distict variation can be observed from the learning curve. Residual plots (error distribution) were also generated to identify prediction errors.



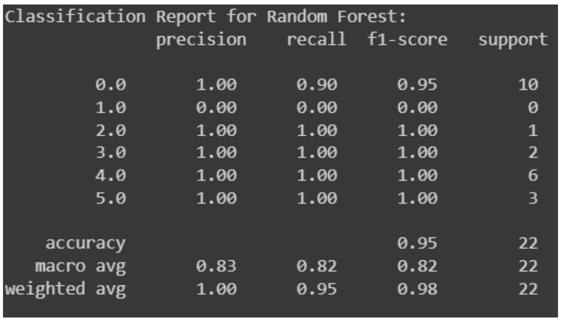
Random Forest Confusion Matrix

	0	1	2	3	4	5
0	9	1	0	0	0	0
1	0	0	0	0	0	0
2	0	0	1	0	0	0
3	0	0	0	2	0	0
4	0	0	0	0	6	0
5	0	0	0	0	0	3

Linear Regression Confusion Matrix

	0	1	2	3	4	5	6
0	9	1	0	0	0	0	0
1	0	1	0	0	0	0	0
2	0	2	0	0	0	0	0
3	0	0	0	3	1	1	1
4	0	0	0	2	1	0	0
5	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0

Figure: Confusion Matrix

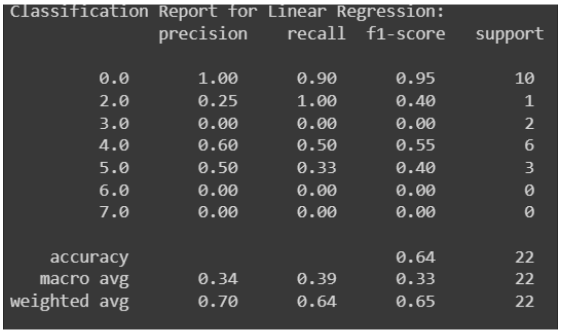


```
Classification Report for Random Forest:
              precision    recall  f1-score   support

         0.0       1.00      0.90      0.95        10
         1.0       0.00      0.00      0.00         0
         2.0       1.00      1.00      1.00         1
         3.0       1.00      1.00      1.00         2
         4.0       1.00      1.00      1.00         6
         5.0       1.00      1.00      1.00         3

    accuracy                           0.95        22
   macro avg       0.83      0.82      0.82        22
weighted avg       1.00      0.95      0.98        22
```

Figure: Classification of RF



```
Classification Report for Linear Regression:
                    precision    recall  f1-score   support

         0.0           1.00      0.90      0.95        10
         2.0           0.25      1.00      0.40         1
         3.0           0.00      0.00      0.00         2
         4.0           0.60      0.50      0.55         6
         5.0           0.50      0.33      0.40         3
         6.0           0.00      0.00      0.00         0
         7.0           0.00      0.00      0.00         0

    accuracy                               0.64        22
   macro avg           0.34      0.39      0.33        22
weighted avg           0.70      0.64      0.65        22
```

Figure: Classification of LR

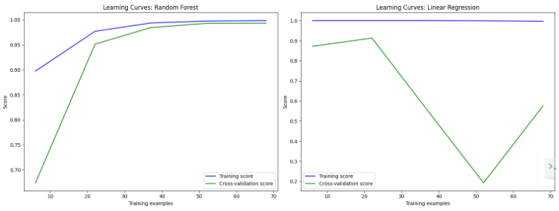


Figure: Learning Curves

Step 5: Important Features

For the **Random Forest** and **Linnear Regression** models, the feature importance scores were evaluated. A visualization of all the Feature Importance was depicted.

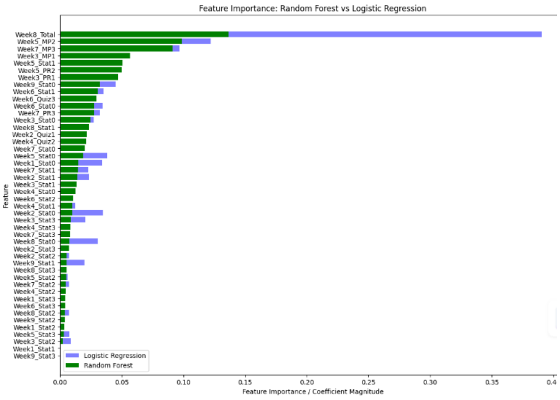
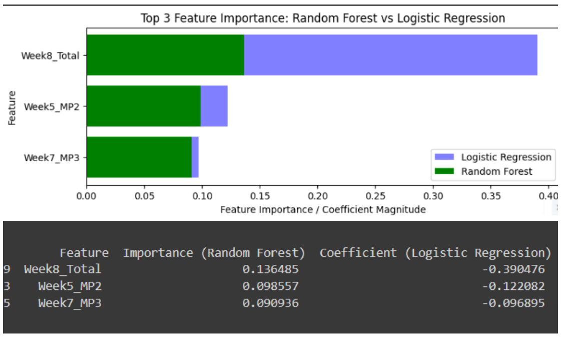


Figure: Feature Importance (All)

The top three most important features for predicting the final grade were:



```
       Feature  Importance (Random Forest)  Coefficient (Logistic Regression)
9  Week8_Total                    0.136485                          -0.390476
3    Week5_MP2                    0.098557                          -0.122082
5    Week7_MP3                    0.090936                          -0.096895
```

Figure: Feature Importance (All)

Data Analysis

Visualization and Interpretation

- **Residual Plots**: The Random Forest residuals are generally small, and points are close to the 0.0 line, indicating more accuracy. The LR points are more spread out, showing larger residuals, which means more errors in prediction.

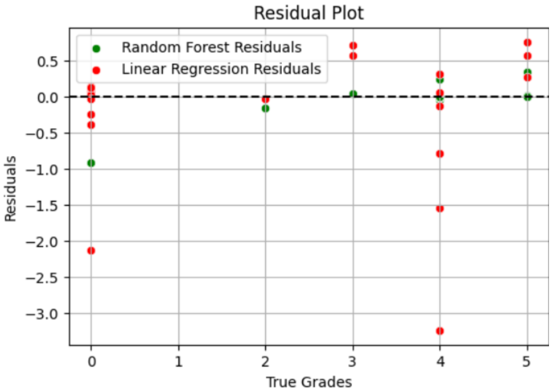


Figure: Residual Plot

- **Error Plots:** RF shows accuracy and LR shows significant deviations.

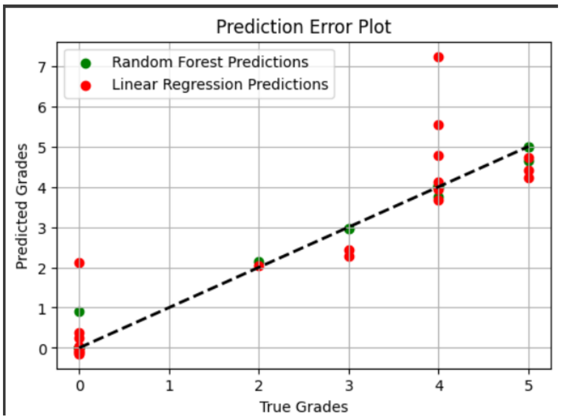


Figure: Error Plot

- **Feature Importance**: A bar chart of feature importance confirmed that total grades and key assignments were the most significant predictors of student performance.

Interesting Observations

- **Activity Logs**: While features related to activity logs were less significant compared to grades, they still contributed to predicting the final grade. This suggests that consistent engagement with course materials and activities plays a role in student success, though it's less influential than direct assessments.

Conclusion

Scientific Bottlenecks

- **High Dimensionality**: With 46 features, managing and interpreting the data was challenging. Feature importance analysis helped narrow down the most influential variables.
- **Overfitting Risk**: The complexity of Random Forest and Gradient Boosting posed a risk of overfitting. We mitigated this through hyperparameter tuning and cross-validation.

Overcoming Bottlenecks

- **Feature Importance Analysis**: By leveraging Random Forest's feature importance scores, we identified the most critical predictors of student performance, reducing the need to include all features in future models.