A logo for college computing

Description automatically generated

**Assessment Cover Page**

|  |  |
| --- | --- |
| *Student Full Name* | Suresh Chinnasamy |
| *Student Number* | sbs24038 |
| *Module Title* | Machine Learning (10 ETCS) |
| *Assessment Title* | CA1 Project |
| *Assessment Due Date* | April 28, 2024 |
| *Date of Submission* | April 28, 2024 |

**Declaration**

By submitting this assessment, I confirm that I have read the CCT policy on academic misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source.

I declare it to be my own work and that all material from third parties has been appropriately referenced.

I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

Contents

[1. Introduction](#_Toc158384946) 2

[2. Implementation & Validation 2](#_Toc158384947)

[3. Hyperparameter Tuning 4](#_Toc158384947)

[4. Results & Conclusion 5](#_Toc158384947)

5. [References 6](#_Toc158384950)

6. [Github Link 6](#_Toc158384950)

1. Introduction:

The development of approaches over recent years has doubled the enthusiasm for applications of machine learning to further improvement in learning outcomes, adaptation to the needs of each student, and the tailor-making of educational strategies. This work will focus on the most important area, which is the identity of factors influencing academic performance, considering mathematics as part of the core subjects that influence most students' academic life. The **goal of this project would be to develop a predicting model for estimating math scores** based on different features like gender, race/ethnicity, parents' level of education, completion of course, completion of test preparation, and scores obtained in reading and writing.

Understanding and predicting performance in mathematics academic set conditions of the development and establishment of math skills that are vital in most fields of academia and professions. Seeking to unravel patterns and insights that could help come up with effective educational practices and policies by the analysis of the way the whole range of factors, from demographic characteristics to more controllable aspects such as nutritional status and preparation strategies, affects math scores. If, however, the test preparation turns out to yield a very strong relationship with math test scores, then greater access for all would be justified in the light of the context of test preparation resources. Similarly, the type of lunch may bring to light the impact of the type of lunch on the mathematical scores and hence guide the interventions to make them more wholesome, without such influences.

### 2. Implementation & Validation

The dataset (JIKADARA, 2024) consists of **1000 instances**, each being described by **8 unique attributes**. A preview of the first five rows of the data is show in Figure 1. Preliminary steps of data pre-processing included looking through the dataset with care in order to fine-tune the integrity of the dataset for analysis. This mainly involved the crosschecking of data types to be compatible for analytical methods, identification of missing or null values to take the appropriate action and help in maintaining the completeness of the data set, and A black screen with white text

Description automatically generatedchecking of categorical variables for instances with single unique values that may affect the result of the analysis.

Figure 1 Top five rows of the dataset.

Moreover, the distribution was checked for the numerical data to assess its fit for statistical analysis. In the preprocessing step, one more important step in this phase was the correlation among numerical variables. Being that a high correlation was obtained between reading and writing scores, this will be suggestive of possible multicollinearity. This means that its reading score is so high that it masks the effects of other parameters. So, to avoid the harmful effect on model performance, the reading score was omitted from further analysis. On the other hand, multicollinearity has been said to be capable in this regard of obscuring the individual effect of the independent variables on the dependent variable, hence seriously affecting the whole interpretation.

For data preprocessing and model construction, the scikit-learn was used. The numerical features were normalized using the StandardScaler, so as not to make more weight drawn from different scales than from any single feature while building a model. An important note on the preprocessing: scaling of the training data only with the scaler and transforming the test data to avoid a possibility of information leakage from the test set into the training process of the model. The categorical columns were one-hot-encoded to convert to numerical form.

Developed a custom function specifically that would make the operation of dividing and rescaling data automated, keeping the predetermined ratio for the test set. Moreover, five-fold cross-validation was applied to all the predictive models to further increase the robustness and generalizability. I used three different distinct models: **Linear Regression, Random Forest Regression, and Support Vector Regression (SVR)**. Linear regression was selected as a method for the baseline model, assuming the relationship between independent variables and the dependent variable to be linear, and this model was considered owing to its simplicity and interpretability. In these models was added a strong ensemble method, Random Forest regression, due to their ability to handle non-linear relationships, robust to outliers, and hence will provide a flexible model. Therefore, they can capture complex patterns in the data. Finally, the support vector regression has been opted for mostly because it usually works very well in high-dimensional spaces and, indeed, has powerful abilities through the use of kernel functions for modelling linear relationships. All the three combined give a comprehensive approach to understand and predict the students' math scores, each assuming different assumptions on the structure of the underlying data and relationships.

The negative mean squared error of the training with different splits and with cross validation is shown in Table 1.

Table 1 Negative Mean Squared Error for different train splits.

|  |  |  |  |
| --- | --- | --- | --- |
| Split Ratio/Model | Linear Regression | Random Forest | SVM |
| 20% | -30.4 | -41.8 | -31.1 |
| 25% | -30 | -41.1 | -30.4 |
| 30% | -30.3 | -41.9 | -30.7 |

### 3. Hyperparameter Tuning

The primary objective of hyperparameter tuning is to maximize the model’s performance. Wherein the training process, it's a bit different, the model parameters are learned, while the hyperparameters are set before the training process and have a very big effect on model performance. Grid Search Cross-Validation (GridSearchCV) is one of the most used hyperparameter tuning techniques. Only this way, it's possible to work systematically through many possible combinations of hyperparameter values, assessing the combination using cross-validation. In other words, GridSearchCV will ensure the data is folded into a specified number of k-folds such that each combination will be tested over separate parts of the data to ensure good estimates of generalization on the unseen data for each set of hyperparameters. The outcome is the hyperparameters' set that gives the best performance, as quantified by a given score or metric.

Following shows the set of hyperparameters used for the models. I have created a dictionary which is in line with format required for GridSearchCV. It will try all possible combinations from the given list and provides the best combination.

**rf\_param\_grid = {**

**'n\_estimators': [100, 200],**

**'max\_depth': [None, 5, 10],**

**'min\_samples\_split': [20, 50],**

**'min\_samples\_leaf': [10, 20]**

**}**

**svm\_param\_grid = {**

**'C': [0.1, 1, 10],**

**'gamma': ['scale', 'auto'],**

**'kernel': ['rbf', 'linear']**

**}**

4. Results & Conclusion

Table 2 shows the MSE for train and test set (75%-25% split). Linear Regression yielded an MSE of 28.99 on the training set and 31.37 on the test set. This very close performance between training and testing set shows that Linear Regression has offered a good balance between bias and variance, meaning it has generalized well without overfitting or underfitting seriously. The increase in MSE from training to the test set is expected to be marginal and it still generalizes well towards unseen data.

In the Random Forest model, the training MSE was 32.73, and MSE was 35.71. Random Forest model's MSE increases from training to test, an indication that mild overfitting problems are exhibited by the model to the training data. Overfitting is present when the model learns from the training data even its noise and includes the outliers, hence learning it better, so making poor generalization to new, unseen data. However, the difference in MSE is not far too high. In other words, overfitting is present but not severe.

The training MSE for the SVM model was observed at 29.22, and in testing, it was 31.1. This model's performance on both the training set and the testing set is close, indicating there is not a very significant level of overfitting or underfitting by the SVM model. This balance indicates that the model has captured the underlying pattern in the data without being too responsive to noise in the training data.

Table 2 MSE for train and test set.

|  |  |  |  |
| --- | --- | --- | --- |
| Data/Model | Linear Regression | Random Forest | SVM |
| Train | 28.99 | 32.73 | 29.22 |
| Test | 31.37 | 35.71 | 31.1 |

Summarizing the generalization capability of all the three models: Linear Regression and **SVM did fairly good, with SVM making a better** showing in general, as its training and test MSE values are pretty close in balance. Figure 2 shows the feature importance obtained from the coefficient values of the trained SVM model. It can be observed that a student with **high writing score tend to get good marks in maths**. The math score is also gender biased, i.e. if the student is male then he is likely to secure more marks in maths(just as per the data).

# 5. References

JIKADARA, B. (2024). *Student Study Performance*. [online] www.kaggle.com. Available at: https://www.kaggle.com/datasets/bhavikjikadara/student-study-performance.

# 6. Github Link

[SureshChinnasamy123/MachineLearning (github.com)](https://github.com/SureshChinnasamy123/MachineLearning)