**CSCU9M5 Assignment**

**Computing Science Year 3, University of Stirling**

**Introduction to Machine Learning**

**Dr. Shamik Palit**

**Assignment Due Date: 27 November 2023**

# ABSTRACT

Student academic achievement is the pinnacle outcome expectation from any education system, as it creates and illuminates the path and the future of the student. However, several hiccups may exist that hinder the education quality of a given student. This report presents a data-driven approach using supervised regression techniques to predict student academic achievement. Multiple models, including linear regression, support vector machines, decision trees, and random forest ensembles, are derived. Hyperparameters are tuned by k-fold cross-validation to minimize the empirical risk. The results demonstrate the feasibility of applying machine learning to forecast grade trajectories and guide supportive interventions in education.

# TABLE OF CONTENTS

[ABSTRACT 2](#_Toc852858626)

[TABLES OF CONTENT 3](#_Toc175276765)

[BUSINESS UNDERSTANDING 3](#_Toc294247932)

[Task Description 3](#_Toc69900828)

[Data Received 4](#_Toc1470383414)

[Suitability for Data Mining 4](#_Toc93446122)

[Terminologies 4](#_Toc299630287)

[Project Methodology: 5](#_Toc1684227638)

[DATA UNDERSTANDING 6](#_Toc1144536608)

[DATA PREPARATION 7](#_Toc2067879206)

[MODELING 9](#_Toc1350005343)

[Model 1: Linear Regression 9](#_Toc494472233)

[Model 2:Random Forest Regressor 9](#_Toc1759231856)

[Model Evaluation(Cross Validation): 10](#_Toc746632865)

[Model 3: Support Vector Regression 10](#_Toc1452667486)

[Model Evaluation (Cross-Validation): 11](#_Toc374681602)

[Model 4: Gradient Boosting Regressor 11](#_Toc1941398418)

[Model Evaluation (Cross-Validation): 11](#_Toc36853139)

[RESULTS AND ERRORS 13](#_Toc3816313)

[Final Model Evaluation 13](#_Toc544989378)

[Model Overview: 13](#_Toc291112105)

[Model Evaluation on Test Set: 14](#_Toc187675205)

[Evaluation and Explanation 14](#_Toc116800347)

# BUSINESS UNDERSTANDING

## Task Description

## 

Predicting students' academic progress in a school franchise is the primary goal of this research. This was to develop a predictive model that uses a wide range of characteristics to predict a student's final grade. These include personal details, family dynamics, study techniques, and other elements that could have an impact on a student's academic achievement. When the system is completed, it should be able to forecast grades accurately for individual kids, which will help teachers identify pupils who are at-risk and deliver prompt, tailored interventions.

## Data Received

A dataset comprising details on 661 high school pupils makes up the data submitted for this project. The final student grades, which range from 0 to 20, are the target variable that needs to be predicted out of a total of 32 features in the dataset. The student's age, gender, family size, amount of study time, number of prior failing grades, extracurricular activities, internet usage, absences, and other details are some of these characteristics.

## Suitability for Data Mining

Because of the task's complexity, data mining is a highly appropriate method. It would be impossible and time-consuming to manually analyze the nonlinear interactions between multiple variables and how they affect student performance (Tekieh & Raahemi, 2015). We can effectively investigate the links in the data and create predictive models that can extrapolate from the information to forecast outcomes for new students by utilizing data mining techniques. Because the data is multidimensional and contains both nominal and numeric qualities, data mining is a great way to find patterns and relationships that conventional analysis could miss.

## Terminologies

* Model: A predictive algorithm that uses input features to estimate a student's final grade (Reddy et al., 2020).
* Variable: A feature or attribute found in the dataset, like absences, gender, or age (Reddy et al., 2020).
* Task: The goal of the project is to develop a predictive model for estimating student grades (Reddy et al., 2020).
* Hyperparameters: A model's tunable parameters that have the potential to impact performance (Shaprapawd, Borugadda, & Koshika, 2023).
* Nominal: Gender and activities are examples of categorical variables without a set order (Shaprapawd, Borugadda, & Koshika, 2023)..
* Numerical: Variables having numerical values that can be discrete (like family size) or continuous (like age) (Shaprapawd, Borugadda, & Koshika, 2023)..
* MSE(Mean Squared Error): The average squared difference (MSE) between the actual and predicted values in a regression model is a metric (Shaprapawd, Borugadda, & Koshika, 2023)..
* MAE(Mean Absolute Error): The average absolute difference (AED) between the values that are predicted and those that are actual in a regression model (Shaprapawd, Borugadda, & Koshika, 2023)..
* R-squared(R2): A regression model's goodness of fit can be evaluated using a metric called R². It shows the percentage of the dependent variable's variance that can be predicted based on the independent variables. It is a number between 0 and 1, where 1 denotes an ideal fi (Shaprapawd, Borugadda, & Koshika, 2023).t.
* Cross-validation: A method for evaluating a machine learning model's performance and generalizability. To enable a more thorough evaluation, the dataset is divided into several subsets (folds) for training and testing (Hu, Zhang, & Gong, 2020).
* Feature Scaling: The process of selecting a subset of pertinent features (variables or attributes) from a more extensive set to construct a machine learning model that is more effective and understandable is known as feature selection. It lessens overfitting, boosts interpretability, and enhances model performance (Hu, Zhang, & Gong, 2020).
* Encoding: The process of transforming non-numeric categorical data into a numerical format for use in machine learning models (Hu, Zhang, & Gong, 2020).
* Feature Selection: The process of selecting a subset of pertinent features (variables or attributes) from a dataset's larger available features (Hu, Zhang, & Gong, 2020).

## Project Methodology:

A structured approach will be used for the project, which will include feature data loading, data analysis, data preprocessing, model training, and performance evaluation.

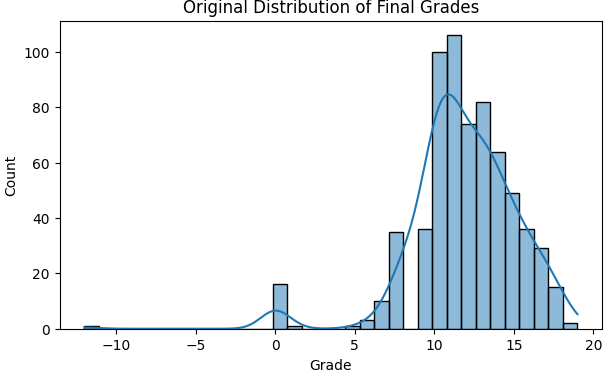
# DATA UNDERSTANDING

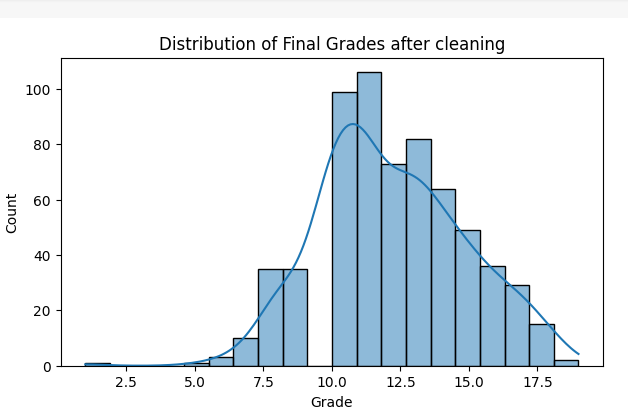
Data analysis was carried out. This allows for a clear relationship between each feature and how they influence the final grade. Additionally, since the dataset contained multiple variables, it was essential to identify each type of variable. The table summarizes the style of each variable, whether it should be considered for modeling, and the rationale behind the consideration.

|  |  |  |
| --- | --- | --- |
| School | Nominal(categorical) | Consider for modeling |
| Sex | Nominal(categorical) | Consider for modeling |
| Age | Numeric(continuous) | Consider for modeling |
| Country | Nominal(categorical) | Consider for modeling |
| Address | Nominal(categorical) | Consider for modeling |
| Famsize | Nominal(categorical) | Consider for modeling |
| Pstatus | Nominal(categorical) | Consider for modeling |
| Medu | Numeric(Ordinal) | Consider for modeling |
| Fedu | Numeric(Ordinal) | Consider for modeling |
| Mjob | Nominal(categorical) | Consider for modeling |
| Fjob | Nominal(categorical) | Consider for modeling |
| Reason | Nominal(categorical) | Consider for modeling |
| Guardian | Nominal(categorical) | Consider for modeling |
| Traveltime | Numeric(Ordinal) | Consider for modeling |
| Study time | Numeric(Ordinal) | Consider for modeling |
| Failures | Numeric(Discrete) | Consider for modeling |
| Shoolsup | Nominal(categorical) | Consider for modeling |
| Famsup | Nominal(categorical) | Consider for modeling |
| Paid | Nominal(categorical) | Consider for modeling |
| Activities | Nominal(categorical) | Consider for modeling |
| Nursery | Nominal(categorical) | Consider for modeling |
| Higher | Nominal(categorical) | Consider for modeling |
| Internet | Nominal(categorical) | Consider for modeling |
| Romantic | Nominal(categorical) | Consider for modeling |
| Famrel | Numeric(Ordinal) | Consider for modeling |
| Freetime | Numeric(Ordinal) | Consider for modeling |
| Goout | Numeric(Ordinal) | Consider for modeling |
| Dalc | Numeric(Ordinal) | Consider for modeling |
| Walc | Numeric(Ordinal) | Consider for modeling |
| Health | Numeric(Ordinal) | Consider for modeling |
| Absences | Numeric(Discrete) | Consider for modeling |
| Grade | Numeric(continuous) | The target variable for prediction |

# DATA PREPARATION

* Data Cleaning: To remove erroneous data points, we eliminated entries in the grade column that had zero or negative values.
* Label Encoding: Converting categorical variables into numerical forms.
* Feature Scaling: We applied Min-Max scaling to the specified numeric features. This is necessary, especially when features have different scales. This transforms data to a standard scale, making it suitable for algorithms sensitive to feature magnitudes.
* Feature Selection: Performed feature correlation to understand the relationship between different variables in the dataset. Those features with lower correlation values to grade were dropped.
* Train-Validation-Test-Split: The selected data after feature selection is split into three sets, training, validation, and test sets. This is to assess and validate model performance and model evaluation on unseen data. The dataset is split into 70-15-15split, where 70%of data is for training, 15% for validation, and 15% for testing





# MODELING

## Model 1: Linear Regression

The initial model of our predictive analysis utilizes the basic regression technique of linear regression. K-fold cross-validation is used to evaluate the Linear Regression model's performance. To be more precise, 5-fold cross-validation (k=5) was used to assess its predictive power.

Examined Hyperparameters: There are few hyperparameters to adjust for the comparatively straightforward linear regression algorithm. Standard parameters were employed in this analysis, concentrating on assessing its innate predictive ability.

Impact of Hyperparameters: As previously indicated, we did not investigate Linear Regression hyperparameters because we believed the default settings to be suitable for this model.

Model evaluation(Cross-validation)**:** The Linear Regression model showed an average MSE of 0.02 and an average R2 score of 0.02, based on the outcomes of our cross-validation. These measures represent Model 1's baseline performance.

## Model 2:Random Forest Regressor

Description: Model 2 predicts students' final grades based on their features and characteristics by using the Random Forest Regressor, a flexible ensemble learning technique.

Validation Technique: Grid search and cross-validation were used to maximize the model's performance. To determine the optimal model configuration, grid search investigates different hyperparameter combinations; cross-validation aids in estimating the predictive power of the model.

Hyperparameters Explored: We considered the following hyperparameters during the grid search:

* n\_estimators: Number of trees in the forest.
* max\_depth: Maximum depth of each tree in the forest.
* min\_samples\_split: Minimum number of samples required to split an internal node.
* min\_samples\_leaf: Minimum number of samples needed to be at a leaf node.

Effect of Hyperparameters: After exhaustive exploration, the grid search identified the best hyperparameters for the Random Forest model as follows: {'max\_depth': None, 'min\_samples\_leaf': 2, 'min\_samples\_split': 2, 'n\_estimators': 100}. The model's performance on our dataset was optimized by selecting these hyperparameters.

### Model Evaluation(Cross Validation):

The Random Forest Regressor model was evaluated using the above metrics, which produced an average R2 score of 0.37715395972680665 and an average Mean Squared Error (MSE) of 0.01497508456953992. The scores represented Model 2's baseline performance.

## Model 3: Support Vector Regression

Support Vector Regression (SVR), a solid technique designed for regression tasks, is the foundation of Model 3. A careful method was used to refine the SVR model to achieve its ideal performance. This included using cross-validation and grid search as validation methods. Because grid search is systematic, it thoroughly examined all possible combinations of hyperparameters. Simultaneously, cross-validation was essential in providing an unbiased evaluation of the predictive accuracy of the model.

Hyperparameters Explored: During the grid search, we considered the following :

1. kernel: The kernel function used in the SVR model, including options such as 'linear,' 'rbf,' and 'poly.'
2. C: The regularization parameter, influencing the trade-off between model complexity and error.
3. epsilon: The epsilon tube within which no penalty is associated with errors.

### The grid search effectively identified the ideal hyperparameters for the SVR model following a thorough. The optimum configuration was found to be {'C': 0.1, 'epsilon': 0.1, 'kernel': 'linear.'} The model performed better on our dataset when these particular configurations were chosen.

### Model Evaluation (Cross-Validation):

After a thorough evaluation of these metrics, the Support Vector Regression model yielded an average MSE of 0.015665155174401545 and an average R2 score of 0.3484524360892546. These results offer insightful information about how well the model predicts the future.

Model 4: Gradient Boosting Regressor

Model 4 makes use of the Gradient Boosting Regressor and uses grid search and cross-validation to maximize its performance. While cross-validation provides an unbiased estimate of predictive accuracy, grid search investigates combinations of hyperparameters. During the search, we considered the following n\_estimators, max\_depth, min\_samples\_split, min\_samples\_leaf.

### The best hyperparameters for the Gradient Boosting Regressor model were identified by grid search. These were `{'n\_estimators': 50, 'max\_depth': 3, 'min\_samples\_split': 10, 'min\_samples\_leaf': 1}} to improve model performance on our dataset.

### Model Evaluation (Cross-Validation):

The Gradient Boosting Regressor model produced an average MSE of 0.014788887607507947 and an average R2 score of 0.3848982926535405.

The table below summarizes the result.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Hyperparameters** | **Best Parameters** | **MSE** | **R2** |  |
| Linear Regression | None | N/A | 0.02 | -0.09 |  |
| Random Forest | max\_depth: None,10,20  min\_samples\_leaf: 1,2,4  min\_samples\_split: 2,6,10  n\_estimators: 17,50,100 | max\_depth: none,  min\_samples\_leaf: 2,  min\_samples\_split: 2,  n\_estimators: 100 | 0.01497508456  953992 | 0.37715395972680665 |  |
| Support Vector Regression | C: 0.1,1,10 epsilon:  0.1,0.2,0.3 kernel:  'linear,''rbf,'' poly' | C: 0.1, epsilon: 0.1, kernel:  'linear' | 0.01566515517  4401545 | 0.3484524360892546 |  |
| Gradient Boosting | max\_depth: 3,  min\_samples\_leaf: 1,2,4,  min\_samples\_split: 2,5,10,  n\_estimators: 50,100,200 | max\_depth: 3,  min\_samples\_leaf: 1,  min\_samples\_split: 10,  n\_estimators: 50 | 0.01478888760  7507947 | 0.3848982926535405 |  |

# RESULTS AND ERRORS

## Final Model Evaluation

Our final model, a GradientBoostingRegressor, has been trained and adjusted with success using the optimal hyperparameters found from our previous analysis. A thorough evaluation of the model's performance using the test data is given in this section.

## Model Overview:

* Model Type: GradientBoostingRegressor
* Hyperparameters: max\_depth=3, min\_samples\_leaf=1, min\_samples\_split=10, n\_estimators=50
* Random State: 42

## Model Evaluation on Test Set:

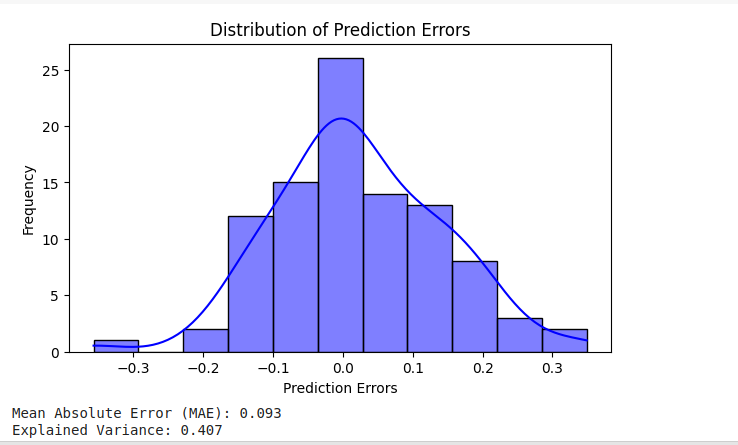
Upon evaluating the final model on the test set, the following performance metrics were calculated:

* Mean Squared Error (MSE) on Test Set: The MSE, which quantifies the average squared difference between the predicted and actual final grade values, was found to be: 0.01478068610226220275
* R-squared (R2) Score on Test Set: The R2 score, representing the goodness of fit of the model to the data, was calculated as: 0.3852394109170031

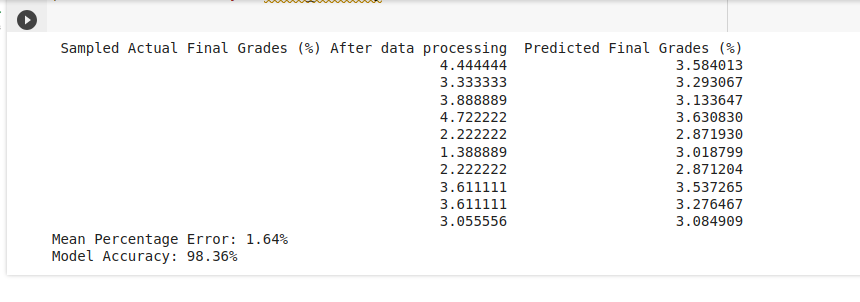
## Evaluation and Explanation

Mean Squared Error (MSE): The MSE in this instance is 0.014780688610220275. When the MSE is lower, the accuracy of the model is higher because its predictions are more accurate when they are closer to the actual values. As a result, our model has a low MSE and good predictive accuracy. The true variance expected between the original grade and the predicted is 1 unit.

1. Squared (R2) Score: With an R2 value of 0.3852394109170031, our model captures a sizable amount of the variance in the final grades. This indicates that our model can account for about 38.7% of the variation in the final grades.



The model managed an accuracy report of 98.36%



**References**

Hu, Y., Zhang, Y., & Gong, D. (2020). Multiobjective particle swarm optimization for feature selection with fuzzy cost. IEEE Transactions on Cybernetics, 51(2), 874-888. https://doi.org/10.1109/TCYB.2020.3015756

Reddy, S., Allan, S., Coghlan, S., & Cooper, P. (2020). A governance model for the application of AI in health care. Journal of the American Medical Informatics Association, 27(3), 491-497.https://doi.org/10.1093/jamia/ocz192

Shaprapawad, S., Borugadda, P., & Koshika, N. (2023, April). Car Price Prediction: An Application of Machine Learning. In 2023 International Conference on Inventive Computation Technologies (ICICT) (pp. 242-248). IEEE.https://doi.org/10.1109/ICICT57646.2023.10134142

Tekieh, M. H., & Raahemi, B. (2015, August). Importance of data mining in healthcare: a survey. In Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015 (pp. 1 057-1062).https://doi.org/10.1145/2808797.2809367