A logo for college computing

Description automatically generated

**Assessment Cover Page**

|  |  |
| --- | --- |
| *Student Full Name* | Erika Vieira Alves |
| *Student Number* | 2024298 |
| *Module Title* | Machine Learning |
| *Assessment Title* | CA1 |
| *Assessment Due Date* | 17/11/2024 |
| *Date of Submission* | 17/11/2024 |

**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

[Introduction 2](#_Toc182758645)

[Data Characterization and Pre-Processing 2](#_Toc182758646)

[Exploratory Data Analysis (EDA): 3](#_Toc182758647)

[Cross-Validation with Varying Training Split 3](#_Toc182758648)

[Machine Learning Models 4](#_Toc182758649)

[SMOTE to balance the training dataset 4](#_Toc182758650)

[Tuning de Hiperparâmetros 5](#_Toc182758651)

[Conclusion 5](#_Toc182758652)

# Introduction

Academic performance is influenced by numerous factors, many of which are rooted in lifestyle choices such as study habits, sleep patterns, socialization, and extracurricular activities. This project aims to explore these correlations and predict students’ performance, represented by GPA, based on these lifestyle factors. Using machine learning techniques, this analysis uncovers patterns and highlights the critical elements that positively or negatively impact academic success.

The primary goal of this project is to apply classification and prediction algorithms to gain insights into the dataset and evaluate the effectiveness of machine learning models.

We will focus on:

* Determine the predominant lifestyle factors affecting academic achievement (GPA).
* Create prediction methods that estimate GPA utilising lifestyle data.
* Provide practical insights on how students can balance their habits to achieve better academic results.

I employed machine learning methods to predict GPA: Random Forest for complicated interactions, Logistic Regression for simplicity and efficacy, and SVM for delineating clear decision limits.

# Data Characterization and Pre-Processing

The dataset, sourced from Kaggle, is titled "Student Lifestyle Dataset." It comprises 2,000 student records, with the following 8 key variables:

* **Study\_Hours\_Per\_Day**: Daily hours dedicated to studying.
* **Extracurricular\_Hours\_Per\_Day**: Time spent on extracurricular activities.
* **Sleep\_Hours\_Per\_Day**: Daily sleep duration.
* **Social\_Hours\_Per\_Day**: Hours spent socializing.
* **Physical\_Activity\_Hours\_Per\_Day**: Time allocated to physical activities.
* **GPA**: Academic performance score.
* **Stress\_Level**: Stress level derived from variables such as study and sleep hours.

This dataset provides a comprehensive view of students' lifestyles, enabling an in-depth exploration of how choices affect academic outcomes.

**Project Steps:**

**Data Prepation and Import:** Importing libraries is necessary to use advanced and optimised features, which makes it easier to work with large amounts of data and run complex analyses. It also makes it easier to reuse code and organise projects, which encourages teamwork and community support.

**Data Cleaning and Preprocessing**: In this step, the **Student\_ID** column, deemed irrelevant for the analysis, was removed. I created GPA\_Class categories (Low, Moderate, Hgh) to prepare for machine learning models.

# Exploratory Data Analysis (EDA):

Exploratory analysis included identifying correlations between key variables, such as the impact of sleep and study hours on GPA. Visualizations, including scatterplots and histograms, were created to understand data distributions and detect potential outliers.

One critical observation was the inverse relationship between stress levels and GPA. Students reporting higher stress levels—often caused by an imbalance between study and sleep—tended to have lower academic performance. In contrast, activities like regular physical exercise showed a positive correlation with better performance.

# Cross-Validation with Varying Training Split

Cross-validation was utilised to assess the model's efficacy with three separate training divisions: 10%, 15%, and 25% of the dataset allocated for training. This method guaranteed a thorough comprehension of the effects of different training data proportions on model accuracy. The findings demonstrated uniform performance for the 10% and 15% splits, each attaining an accuracy of 0.88, signifying that the model effectively generalised despite constrained training data. However, with 25% of the training data, the accuracy decreased somewhat to 0.85, indicating possible variability or overfitting attributable to the enlarged training size in relation to the test set. These findings emphasise the necessity of evaluating various splits in cross-validation to determine the appropriate data distribution for accurate predictions.

# Machine Learning Models

GPA was predicted using many machine learning techniques:

* **Random Forest:** Selected for its ability to clarify complex component interactions.
* **Logistic Regression:** For classification problems, a simple yet effective model
* **Support Vector Machine (SVM):** Best for spotting different decision lines.

Cross-validation and hyperparameter optimisation via Grid Search were employed to train and evaluate each model. Performance indicators, such as confusion matrices, classification reports, and accuracy scores, were employed to evaluate the models.  
Conclusions: The Random Forest algorithm exhibited the highest accuracy in predicting GPA among all examined models. It also emphasised the most significant variables, including:

* **Study Hours Per Day:** The primary determinant of GPA.
* **Sleep Hours Per Day:** Essential for sustaining low stress and elevated productivity levels.
* **Physical Activity Hours:** Moderately influential, however advantageous overall.

While, Logistic Regression yielded acceptable outcomes, indicating certain linear correlations across variables, the Random Forest model identified more intricate patterns.

# SMOTE to balance the training dataset

# The class imbalance in the GPA dataset was fixed with SMOTE, which made the model work better by creating fake cases for classes that weren't well-represented. This made learning more fair and improved generalisations.

# The results show: Cross-validation accuracy tests show that SVM does better than the others, with an 87.95% score. Logistic Regression comes in second with an 87.40% score, and Random Forest comes in third with an 86.75% score. There was good performance from both models, but the SVM was more accurate because it could set clear decision limits. There aren't many differences, which means that all of the models are good at this job.

# Tuning de Hiperparâmetros

GridSearchCV was employed to perform an exhaustive search over hyperparameter combinations, using cross-validation to evaluate each scenario. This ensured the selection of parameters that optimized metrics like accuracy or F1-Score.

**Impact of Tuning:**

* **Random Forest:** Improved accuracy and stability by reducing overfitting through depth limitation and optimized splits.
* **SVM:** Achieved optimal performance with C=1 and kernel=rbf, providing greater precision and clearer class separation.
* **Logistic Regression:** No hyperparameters were tuned due to its deterministic nature, but cross-validation ensured appropriate data selection.

Results: Tuning significantly improved Logistic Regression's mean CV accuracy from 87.40% to 88.05%, while Random Forest saw a smaller improvement (86.75% to 87.40%). SVM performance remained stable (87.95% to 87.90%), indicating that tuning was most impactful for Logistic Regression.

# Conclusion

The analysis provided significant results for academic and educational organisations. It emphasised the importance of balancing study and sleep to achieve optimal performance. Furthermore, extracurricular and physical activities indirectly contributed to improved outcomes by reducing stress and improving concentration.  
These data can guide sensible treatments include time management courses or techniques to enhance sleep quality, therefore helping students to achieve improved academic performance.  
The results of this study proved that academic performance is highly influenced by lifestyle choices. Apart from verifying these results, machine learning gave GPA a consistent predictive model.

Highlight the need of a healthy lifestyle and provide the path for next research to investigate certain interventions for best performance of students.  
This study can be expanded to incorporate supplementary variables, such the influence of social media engagement or meal quality on academic achievement. Moreover, advanced algorithms such as neural networks may be analysed to contrast outcomes and potentially improve accuracy.

# References

A case study prepared for HDip in DAB/ AI Applications by Dr. Muhammad Iqbal

Iqbal, Muhammad . (2024). CASE STUDY - Hyperparameter Tunning and GridSearchCV. Class materials, CCT College Dublin.

Iqbal, Muhammad . (2024). Comparison\_Classification\_Models. Class materials, CCT College Dublin.

Student Lifestyle Dataset (2024) Kaggle. Available at: https://www.kaggle.com/datasets/steve1215rogg/student-lifestyle-dataset (Accessed: 11 November 2024).