CITY UNIVERSITY

ASSIGNMENT COVER SHEET

Student Name & ID	1. CAROLLISA A/P WADIWILL - 202501010307 - BCSSE	
Subject Code	BIT 4333	
Subject Title	INTRODUCTION TO MACHINE LEARNING`	
Assignment Topic/Title	FINAL PROJECT	
Subject Lecturer	MR NAZMIRUL IZZAD NASSIR	**Date Received**
Subject Tutor	-	
Due Date	4TH DECEMBER 2025	
Word Count	-	

FORMATTING and REFERENCING

Your assignment must meet the formatting and referencing requirements noted by the lecturer. By signing below, you are confirming that you have met those requirements.

DECLARATION

This assignment is my own original work. No part of this work has been copied from any other source or person except where due acknowledgement is made, and no part of been previously submitted for assessment at this or any other institution. For the purposes of assessment and standards, I give the University permission to retain this assignment; provide a copy to other assessors; and evaluate its academic integrity through the use of a plagiarism checking service.

CITY UNIVERSITY

PROJECT TITLE: PREDICTING STUDENT PERFORMANCE USING MACHINE LEARNING

Executive Summary

This project explores machine learning regression variables that affect the performance of students. The purpose is to estimate the score of students in their exams basing on the demographic factors, behavioral factors and academic factors. It uses a cleansed data of student variables including hours studied, attendance, parental involvement, access to resources, sleep hours, past scores, and level of motivation.

Some of the regression models that were trained are Linear Regression, SVR, XGBoost Regressor, and CatBoost Regressor. The Linear Regression model was the most suitable performing model with MSE = 3.256 and R2 = 0.770. The analysis of the features importance revealed that the best predictors are attendance, hours studied, and previous scores.

To illustrate interactive predictions of student exam scores, a Streamlit application was created, which allows using varied values of each feature to see what prediction will occur.

Problem Statement

To make appropriate education planning and interventions, it is important to know the key factors that determine the performance of students.
Conventional statistical techniques can be ineffective to reflect complicated associations among the variables. This project aims to:

- Determine the student factors that have the most significant impact on exam performance.
- Machine learning regression predicts exam scores correctly.
- Demonstrate an interactive application that can be used to simulate the performance prediction of the students under different features.

CITY UNIVERSITY

Dataset Source

The data has been an open educational content that has student demographic, behavioral and academic data. Key features include:

- Hours_Studied per week
- Attendance rate
- Previous_Scores
- Motivation_Level
- Parental_Involvement
- Sleep_Hours per day

Optional variables: Access_to_Resources, Extracurricular_Activities, Tutoring_Sessions, Internet_Access, Family_Income, Teacher_Quality, School_Type, Peer_Influence, Physical_Activity

The data was processed to deal with missing data and discrepancy. Model compatibility was done by encoding categorical variables, and numeric features were normalized.

Link: https://www.kaggle.com/datasets/lainguyn123/student-performance-factors

CITY UNIVERSITY

Methodology

1. Data Preprocessing

- Numeric values that were missing were replaced by the median.
- Missing categorical value was replaced by mode.
- Outliers and relationships were detected using boxplots and the correlation analysis.
- One-hot encoded categorical features and numeric features were standardized with the help of StandardScaler.

2. Feature Selection

Regression predictors that were identified as the key ones were the following:

- Hours_Studied
- Attendance
- Previous_Scores
- Motivation_Level
- Parental_Involvement
- Sleep_Hours

The Streamlit app had optional features that could be explored further but never in model training.

3. Model Development

The regression models trained and compared below were:

- Linear Regression - chosen because of simplicity and ease of interpretation.
- Support Vector Regressor (SVR)
- XGBoost Regressor

- CatBoost Regressor

Mean Squared error (MSE) and the R 2 metrics were used to evaluate the models. Linear Regression model was selected because it was the best model in terms of the lowest MSE and maximum R2.

Results:

Model	MSE	R2
Linear Regression	3.256	0.770
SVR	3.380	0.761
XGBoost	4.614	0.674
CatBoost	3.734	0.736

Feature Importance (Linear Regression Coefficients):

Feature	Importance
Attendance	2.290
Hours_Studied	1.757
Previous_Scores	0.706
Tutoring_Sessions	0.626
Peer_Influence_Positive	0.517
Distance_From_Home_Near	0.419
Extracurricular_Activities_Yes	0.286
Motivation_Level (encoded)	- 0.277
Parental_Involvement (encoded)	- 0.535
Sleep_Hours	- 0.018

Top contributing factors: Attendance, hours studied and previous scores

CITY UNIVERSITY

4. Deployment

- Scalerization of objects (scaler) and trained model in .pkl files.
- Interactive prediction is possible with Streamlit application:
- The values of both main and optional features are entered by the users.
- Attributes are coded and normalised to training data.
- The Exam_Score is projected as real-time.

Example input for prediction:

```
example = {
    "Hours_Studied": 20,
    "Attendance": 85,
    "Previous_Scores": 70,
    "Motivation_Level": "Medium",
    "Parental_Involvement": "High",
    "Sleep_Hours": 7
}
```

5. Repository Structure

```
StudentPerformanceML/
│
├── data/
│   └── StudentPerformanceFactors_Cleaned.csv
│
├── docs/
│   └── BIT4333_FinalProject_Report.pdf
│
```

```
├── models/
│   ├── 01_DATAPREPROCESSING.ipynb
│   ├── 02_MODELDEVELOPMENT.ipynb
│   └── 03_MODELTESTING.ipynb
│
├── slides/
│   └── BIT4333_FinalProject_Slides.pptx
│
├── streamlit_app/
│   └── app.py
│
├── .gitignore
├── README.md
└── requirements.txt
```

All processing, models and code are designed so as to facilitate reproducibility and facilitated deployment.

Conclusion

Linear Regression model has a high accuracy in predicting the student exam score with high accuracy (R2 = 0.770) and low mean squared error. The most influential factors were defined as attendance, hours studied and previous scores.

The adoption of a Streamlit application makes it possible to analyze the performance of students in real-time on the basis of hypothetical conditions so that teachers and learners could plan interventions or study approaches efficiently.

Acknowledgements

I would be glad to judgefully acknowledge that Mr. Nazmirul Izzad Bin Nassir guided and supported me throughout this project. His guidance proved necessary in making me complete this project successfully and on my own..