Applied Machine Learning: Report

Name: Daniel Cyril Obon
Student ID: C2650218

I. Introduction

Machine Learning has rapidly transformed industries by enabling systems to learn from data, adapt to changes and make predictions or decisions without explicit programming. Machine learning focuses on creating systems that can automatically improve their performance through experience. Recent advancements in machine learning have been fuelled by the emergence of innovative algorithms and theory, as well as accessible online data and affordable computational resources. Data-driven machine learning approaches are increasingly being adopted across various domains, including healthcare, manufacturing, education, financial modelling, law enforcement, and marketing, enabling more evidence-based and informed decision-making in these areas [1].

This report aims to assess the accuracy and effectiveness of various machine learning models in predicting a student's gender based on their test scores and participation in a test preparation course, by comparing the accuracy rating received from each group member's model. It explores how machine learning techniques can be utilized to uncover patterns in the data and enhance predictive capabilities in this context. Moreover, this report focuses on utilizing the Logistic Regression model to make predictions within the specified constraints. To achieve this, the training and analysis of the data will be conducted using RStudio, leveraging its extensive suite of relevant libraries.

II. Dataset

For this assignment, the dataset titled "Student Performance Prediction" from Kaggle has been selected. This dataset provides details on the academic performance of high school students in mathematics, including their grades and demographic characteristics. The information was gathered from three different high schools located in the United States [2]. It includes a file named exams.csv (Figure 1), which contains 8 variable columns and 1000 student data entry rows.

This dataset was chosen because it contains data on 1000 students, providing a sufficiently large sample size to enhance the accuracy and reliability of the analysis. Additionally, this dataset includes relevant variables such as math scores, reading scores, writing scores, and test preparation course completion, all of which are crucial for predicting a student's gender. However, the dataset also contains variables such as race/ethnicity, parental level of education, and a student's lunch price, which may be considered less relevant to the task of predicting a student's gender and are therefore excluded from the primary analysis when normalising the dataset.

	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
1	female	group D	some college	standard	completed	59	70	78
2	male	group D	associate's degree	standard	none	96	93	87
3	female	group D	some college	free/reduced	none	57	76	77
4	male	group B	some college	free/reduced	none	70	70	63
5	female	group D	associate's degree	standard	none	83	85	86
6	male	group C	some high school	standard	none	68	57	54
7	female	group E	associate's degree	standard	none	82	83	80
8	female	group B	some high school	standard	none	46	61	58
9	male	group C	some high school	standard	none	80	75	73
10	female	group C	bachelor's degree	standard	completed	57	69	77
11	male	group B	some high school	standard	none	74	69	69
12	male	group B	master's degree	standard	none	53	50	49
13	male	group B	bachelor's degree	free/reduced	none	76	74	76
14	male	group A	some college	standard	none	70	73	70
15	male	group C	master's degree	free/reduced	none	55	54	52
16	male	group E	master's degree	free/reduced	none	56	46	43
17	female	group C	some college	free/reduced	none	35	47	41
18	female	group C	high school	standard	none	87	92	81
19	female	group E	associate's degree	free/reduced	none	80	82	85
20	female	group D	associate's degree	standard	completed	65	71	74
21	male	group C	high school	free/reduced	none	66	66	62
22	female	group D	associate's degree	standard	completed	67	71	76
23	female	group B	some college	standard	none	70	71	71
24	male	group E	associate's degree	standard	none	89	88	86
25	male	group D	associate's degree	standard	completed	99	85	88
26	male	group B	some college	standard	none	74	83	72

Showing 1 to 27 of 1,000 entries, 8 total columns

Figure 1: exams.csv Dataset

III. Problem

The problem at hand involves predicting the gender of a student based on their performance in mathematics, reading, and writing, as well as their participation in a test preparation course. The dataset provides multiple features related to student performance, yet it is unclear which features such as test scores or test preparation are most indicative of gender, making the task challenging.

The problem is further compounded by the inclusion of less relevant variables such as race/ethnicity, parental education level, and lunch price, which could introduce bias into the predictions. Data bias arises when the datasets used to train machine learning models are incomplete or lack proper representation, resulting in skewed outputs. This can occur due to collecting data from biased sources, excluding essential information, or including errors in the dataset [3]. Deciding how to handle these variables depending on each applied machine learning model is crucial for ensuring the fairness and accuracy of the model.

Thus, the problem requires careful preprocessing to remove irrelevant or potentially biased variables while retaining those most predictive of the target outcome. Additionally, a detailed analysis of each variable will be conducted after preprocessing, including examining the correlation of each variable with a student's gender. Besides that, the performance of

different machine learning models will also be evaluated and compared, as implemented by each group member, to determine which approach yields the highest accuracy.

IV. Data Preparation and Exploration

Data preprocessing is a critical step in enhancing the efficiency of machine learning models by improving the quality of the input features. For instance, research using the widely recognized Framingham Heart Study dataset demonstrated that effective preprocessing techniques significantly boosted the predictive accuracy of otherwise underperforming classifiers. The findings highlight how preprocessing can play a vital role in improving model performance, particularly in tasks such as assessing the risk of coronary heart disease [4].

	gender	test preparation course	math score	reading score	writing score
1	female	1	59	70	78
2	male	2	96	93	87
3	female	2	57	76	77
4	male	2	70	70	63
5	female	2	83	85	86
6	male	2	68	57	54
7	female	2	82	83	80
8	female	2	46	61	58
9	male	2	80	75	73
10	female	1	57	69	77
11	male	2	74	69	69
12	male	2	53	50	49
13	male	2	76	74	76
14	male	2	70	73	70

Showing 1 to 15 of 1,000 entries, 5 total columns

Figure 2: Dataset After Preprocessing (Normalisation)

To prepare the dataset for analysis, several preprocessing steps were carried out to ensure data quality and suitability for machine learning models. The raw dataset, exams.csv, was first imported into RStudio and its structure was inspected to understand the variables Firstly, less relevant variables such as race/ethnicity, parental education level, and lunch price were excluded from the analysis to focus on variables most relevant to the prediction. The variable “preparation course completion” was encoded into numerical values (1 for completed and 2 for none) to facilitate processing by machine learning models. Similarly, gender was converted to a factor, which is a categorical variable rather than a numeric or character variable that is to be served as the target variable.

The dataset was then cleaned by removing rows containing missing values using the na.omit() function. Thus, to verify the data integrity of the dataset, it was checked for any remaining null values to ensure it was free of incomplete entries. Following preprocessing, the data was split into training and testing subsets to enable model training and evaluation. By employing the createDataPartition function, it ensures an even distribution of the target variable (gender) across both testing and training datasets. This preprocessing workflow enhances the dataset's quality and structure, contributing to more reliable and accurate predictions during analysis.

V. Selected Algorithm (Logistic Regression)

Logistic regression is used to examine the relationship between predictor variables and categorical outcomes. Typically, the outcome is binary, such as the presence or absence of a condition (e.g., non-Hodgkin lymphoma), in which case the model is called a binary logistic model. When the model includes only one predictor variable, it is known as a simple logistic regression. If the model incorporates multiple predictors, such as risk factors and treatments, which may include both categorical and continuous variables, it is referred to as a multiple or multivariable logistic regression [5].

Thus, for this group assignment I have selected the Logistic Regression model to predict the gender of a student. The reason this algorithm is suitable for this case study is due its effectiveness in modelling binary outcomes, such as predicting a student's gender based on various predictor variables. The model is capable of handling both categorical and continuous variables, such as the scores in mathematics, reading, writing, and participation in a test preparation course, making it versatile for this analysis. Additionally, logistic regression provides a clear understanding of the relationship between the predictors and the outcome, offering probabilities that can be interpreted as the likelihood of a given outcome (in this case, gender). Furthermore, logistic regression is relatively simple, computationally efficient, and widely used for binary classification problems, which makes it an ideal choice for this assignment.

VI. Critical Evaluation

a) Prediction Accuracy

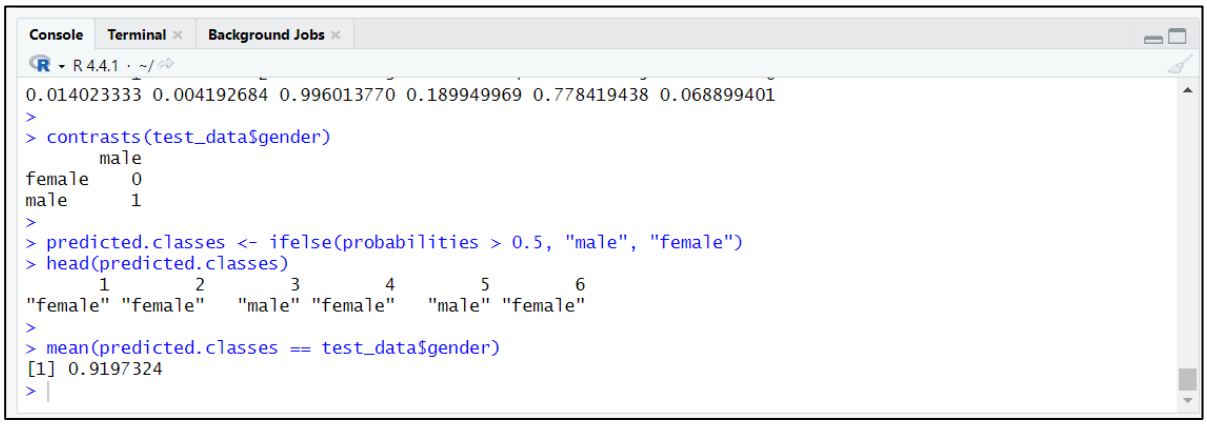

Figure 3: Logistic Regression Accuracy Using Test Data for Gender Prediction (91.07%)

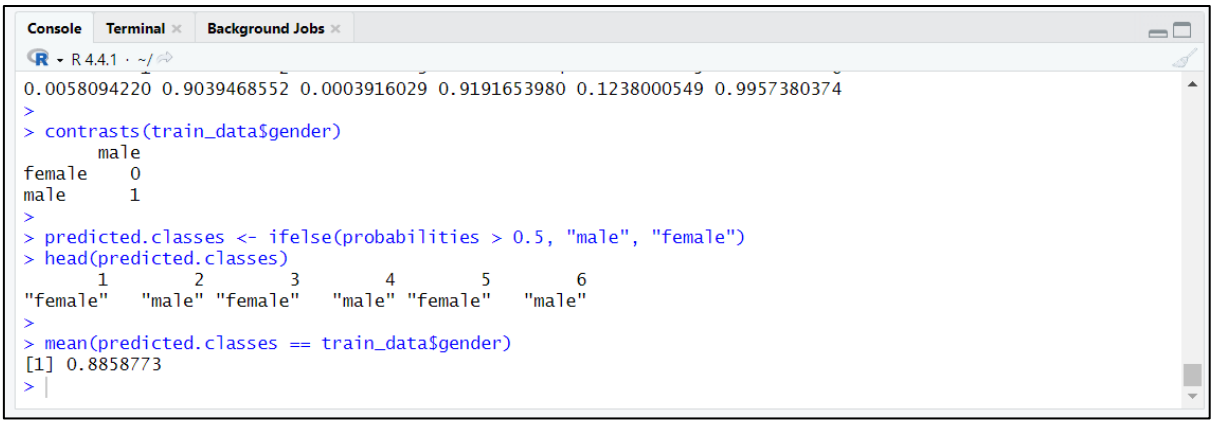

Figure 4: Logistic Regression Accuracy Using Training Data for Gender Prediction (88.59%)

From Figure 3 and 4, it could be noticed that the prediction using testing data has a higher accuracy (91.07%) compared to using training data (88.59%). This difference in accuracy could suggest that, since the test and training data are sampled from the same overall dataset but are different subsets, there might be a coincidence where the specific test data contains more representative or simpler cases for the model to classify, leading to a higher accuracy on the testing set. In contrast, the training set might have more complex, noisy, or varied examples, leading to lower accuracy on that data.

To mitigate this discrepancy in accuracy, the k-fold cross-validation method could be implemented instead of relying on a single splitting of data. K-fold cross-validation is a technique where the dataset is randomly divided into k equal-sized subsets or "folds." The value of k represents the number of partitions the data is split into. For instance, with a k-value of 10, the dataset would be divided into 10 separate groups. Nine of these parts are used for training the model, while the remaining one is used for testing. This process is repeated ten times, with each fold serving as the test set once, and the

remaining nine folds used for training. The final performance is then averaged across all iterations. [6].

Accuracy Using Testing Data for Gender Prediction	
Applied Algorithm	Accuracy Percentage
K-Nearest Neighbours (KNN)	0.8566667 (85.67%)
Decision Tree	80.60%
Naïve Bayes	61.62%
Logistic Regression	0.9197324 (91.97%)

Table 1: Accuracy Using Testing Data for Gender Prediction with Difference Algorithms

Accuracy Using Training Data for Gender Prediction	
Applied Algorithm	Accuracy Percentage
K-Nearest Neighbours (KNN)	0.9171429 (91.71%)
Decision Tree	85.16%
Naïve Bayes	68.19%
Logistic Regression	0.8858773 (88.59%)

Table 2: Accuracy Using Training Data for Gender Prediction with Difference Algorithms

The tables above display the accuracy rating obtained by other group members, using testing and training data for gender prediction. When comparing the accuracy of different machine learning models for gender prediction, it is evident that Logistic Regression consistently performs well, particularly in the testing dataset (91.97%) and the training dataset (88.59%). This suggests that the model is not only able to generalize effectively to unseen/new data, but it also performs reliably on the training set in predicting a student's gender.

The higher accuracy observed for Logistic Regression could be attributed to its simplicity and effectiveness for binary classification, such as gender prediction. It models the linear relationship between predictors and the target variable, resulting in strong performance without overfitting. In contrast, K-Nearest Neighbours (KNN) shows a significant accuracy drop from training (91.71%) to testing (85.67%), indicating potential overfitting. Naïve Bayes and Decision Trees also exhibit lower accuracy (61.62% and 80.60%, respectively), likely due to Naïve Bayes' assumption of feature independence, which may not be the case in this context, and Decision Trees' difficulty balancing complexity, leading to overfitting or underfitting.

b) Graph Analysis

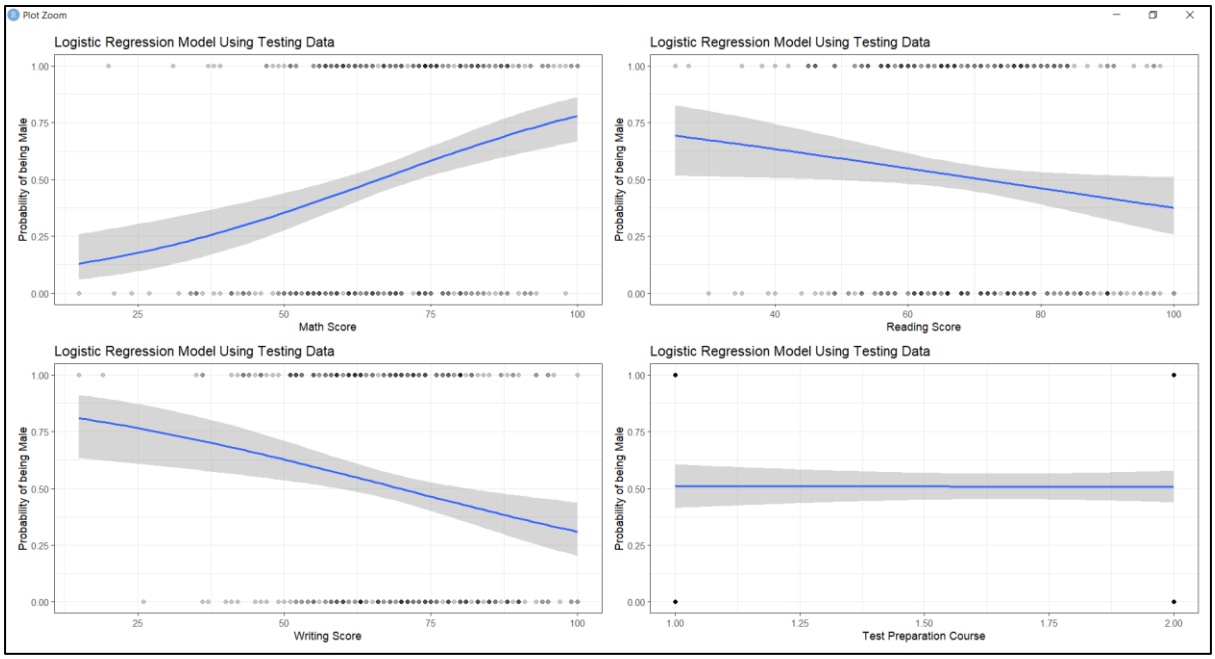

Figure 5: Logistic Regression Using Testing Data Gender Probability Graph

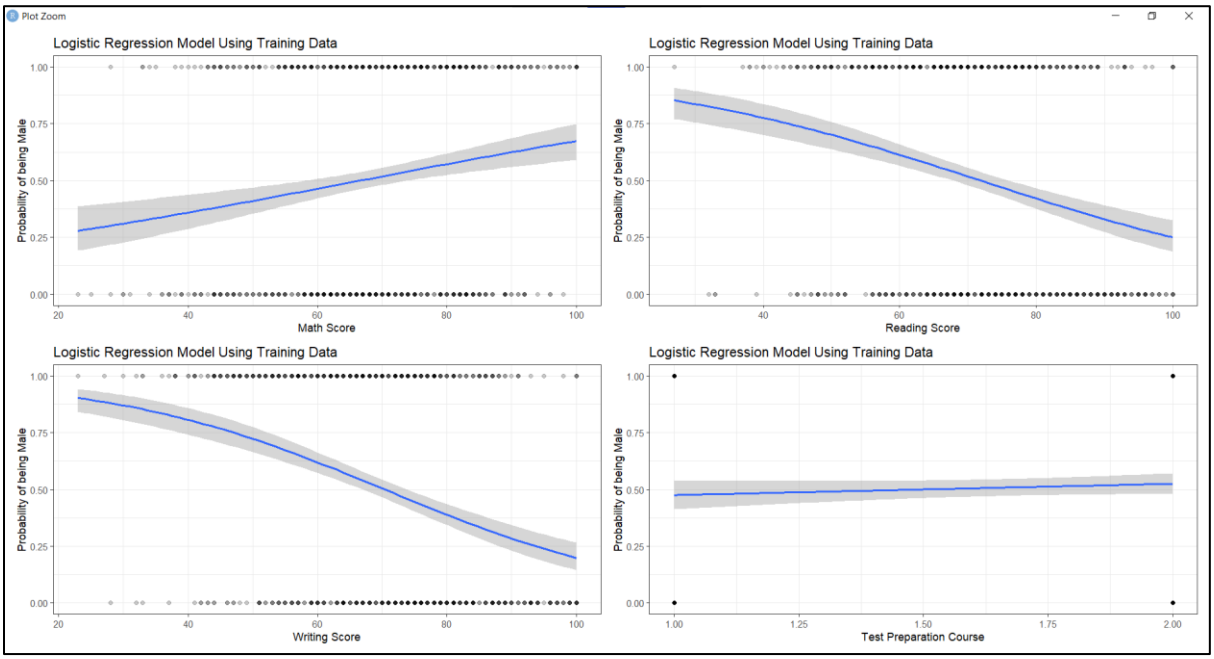

Figure 6: Logistic Regression Using Training Data Gender Probability Graph

Analysing the graphs in Figure 5 and 6, they suggest that there is a noticeable gender-based trend in academic performance, with males generally achieving higher scores in mathematics, while females tend to perform better in reading and writing. This could imply that there are potential differences in how male and female students approach or excel in these subjects. Additionally, the flat trend observed for the test preparation

course suggests that the completion of the course does not significantly influence the likelihood of a student being male or female, indicating a balanced distribution of gender across those who participated in the course. These observations may point to broader educational or societal patterns in performance and behaviour, though further analysis would be required to understand the underlying factors.

VII. Conclusion

In conclusion, this case study highlights the challenges and benefits that could be gained from using machine learning models like Logistic Regression. The analysis revealed that Logistic Regression performed best in terms of accuracy, outperforming other models in this context. However, potential limitations in the model's consistency were identified, highlighting the need for more rigorous evaluation techniques. In the future, addressing this issue through the application of advanced methods, such as k-fold cross-validation, could provide a more comprehensive assessment of model performance, ensuring fairness and better prediction accuracy.

Additionally, this case study has helped improve my understanding of applied machine learning techniques and data preprocessing. The experience has enabled me to develop critical skills in data analysis, model evaluation, and identifying potential biases, which are valuable in fields like data science, education technology, and AI development. Furthermore, working collaboratively within a group has significantly enhanced my communication and teamwork skills, strengthening my ability to effectively interact in a professional environment.

VIII. Peer Points

Name	Peer Points (Total 13)
Danile Cyril Obon	4
Raymond Yau	3
Samual Winter	3
Harvey Peacock	3

References:

[1] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255-260, 2015.

[2] R. Kiattisak, "Student performance prediction," *Kaggle*, 2023. [Online]. Available: https://www.kaggle.com/datasets/rkiattisak/student-performance-in-mathematics?resource=download. [Accessed: Dec. 28, 2024].

[3] E. Ferrara, "Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies," *Sci*, vol. 6, no. 1, p. 3, 2023.

[4] O. Sami, Y. Elsheikh, and F. Almasalha, "The role of data pre-processing techniques in improving machine learning accuracy for predicting coronary heart disease," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 6, 2021.

[5] T.G. Nick and K.M. Campbell, "Logistic regression," *Topics in Biostatistics*, pp. 273-301, 2007.

[6] C. A. Ramezan, T. A. Warner, and A. E. Maxwell, "Evaluation of sampling and cross-validation tuning strategies for regional-scale machine learning classification," *Remote Sensing*, vol. 11, no. 2, p. 185, 2019.