Republic of the Philippines
Western Mindanao State University
College of Computing Studies
DEPARTMENT OF COMPUTER SCIENCE
Zamboanga City

Predicting Western Mindanao State University College Entrance Test Scores Based on Student Profile and Senior High School Grades Using Data Mining Techniques

In partial fulfillment of the requirements for the degree of
Bachelor of Science in Computer Science

Presented to the Faculty of
Department of Computer Science
College of Computing Studies

Theo Jay M'lleno G. Sanson
Jane Stephanie J. Domingo
Researchers

Mr. Jaydee C. Ballaho
Adviser

March 31, 2022

Western Mindanao State University
College of Computing Studies
DEPARTMENT OF COMPUTER SCIENCE
Zamboanga City

Approval Sheet

The Thesis attached hereto, entitled **"Predicting Western Mindanao State University College Entrance Test Scores Based on Student Profile and Senior High School Grades Using Data Mining Techniques"**, prepared and submitted by Theo Jay M'Lleno G. Sanson and Jane Stephanie J. Domingo in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science, is hereby **recommended for Oral Examination**.

MR. JAYDEE C. BALLAHO
Adviser

APPROVED by the Oral Examination Committee on _____ with a rating of **PASSED**.

MS. LUCY F. SADIWA, MSCS
Chairperson

MR. SALIMAR B. TAHIL, MEnggEd
Member

MS. MARJORIE A. ROJAS
Member

ACCEPTED in partial fulfillment of the requirements for the degree of **Bachelor of Science in Computer Science**

ODON A. MARAVILLAS JR, MSCS
Head, Department of Computer Science

RODERICK P. GO, Ph.D.
Dean, College of Computing Studies

Acknowledgment

To Roberto H. Domingo and Jennifer J. Domingo and to Eduardo F. Sanson and Lilian G. Sanson for providing love, patience, financial aid, and especially supporting us when it seemed that there was no other way for us to accomplish the study;

To our thesis adviser and mentor, Sir Jaydee C. Ballaho, for the gracious guidance during our research and the fatherly support;

To Ma'am Lucy F. Sadiwa, for her brilliant insights that inspired us to do our best for the research;

To Engr. Marvic Lines, for guiding us during the beginning of this research;

To Ma'am Marjorie Roxas and Sir Salimar Tahil for guiding us after our preliminary defense;

To Sir Eric H. Alfaro of WMSU University Registrar, for being most accommodating while we were collecting our research data;

To Sir Bong Salcedo of WMSU College of Law for his indispensable support;

To college best friend, Ate Emy Joy A. Omictin, for sharing in the same struggles in life, for when things got tough, being the only people who knew how to comfort each other, and for all the advice and laughs;

To our friends and classmates Ronald Dale, Mark Anthony, Jayson, and Migfren for all the laughs that they shared with us;

To Krizti Jessu V. Villocillo, for showing up when I needed someone the most, for being someone that pushed me to do better in this field I chose, and for giving me hope and showing boundless support;

To Shamy Rasma C. Jailani, for her invaluable encouragement during the bleakest moments of this endeavor;

To our good friends Jheff Nico, Mehrajz, Al-Kneedzfarl, Mylevorne, Khalil, and Aaron for the companionship and wild shenanigans;

To Kenneth Ray G. Rule, for telling us to keep on keeping on;

And finally, above all, to our Almighty God, whose love, patience, and compassion have provided us with the faith, courage, and strength to face adversity;

We express our deepest and sincerest gratitude and dedicate the accomplishment of this research to all of you. Much of the work put into this thesis would not have been possible without the love and help of our dear family and friends.

Abstract

College entrance test results are a significant factor in an applicant's admission into the college of their choice. Predicting their performance on the test would be an effective means for promising students to improve their chances of being admitted into their chosen course. However, in the Philippines, the feasibility of using senior high school academic performance as a predictor for the college entrance test has not been previously studied. This thesis investigated the viability of using past academic performance to predict student performance on the WMSU College Entrance Test (CET). This study used the K-12 Curriculum Grading System to predict the students' CET performance results.

The study uses machine learning to test the hypothesis that previous academic performance predicts college entrance test performance and determine the accuracy of college entrance test results. This was developed in parallel to the development of the online CET registration system made exclusively for the use of the WMSU TEC. The researchers then chose the Support Vector Machine Model due to its best performance in generalizing its output and data and providing a better accuracy in prediction.

The study results show that a higher past academic performance accurately predicted a higher CET score. A lower academic performance predicted a lower CET score. The final iteration of the ternary classification model accurately predicted 50% of the unseen data points. It suggests that the student's past academic performance, expressed in the form of their senior high school grade, is a predictor of CET performance. The research thus concludes that there exists a correlation between the K–12 curriculum grading system and the standardized examination of the WMSU CET.

Keywords: Machine Learning, Classification Algorithms, Standardized Exam Results, Academic Performance, Prediction Algorithms, Support Vector Machine Learning

Table of Contents

Approval Sheet i
Acknowledgment ii
Abstract iii
CHAPTER I INTRODUCTION 1
Background of the Study 1
Statement of the Problem 2
Objectives 2
Significance 3
Definition of Terms 6
CHAPTER II REVIEW OF RELATED LITERATURE 8
Related Studies 8
Prediction using Academic Performance 8
Academic Performance and Data Mining Techniques 8
Foreign Studies 9
Local Studies 11
Synthesis 12
Comparison Table of Related Systems 13
Conceptual Framework 14
CHAPTER III METHODOLOGY 16
Research Design 16
Data Gathering Techniques and Procedures 16
Statistical Tools 17
Linear Regression 18
Multiple Linear Regression 18
Lasso Regression 20
Ridge Regression 21
Tree Regression 22
Analytical Tool 23
Software Process 24
Planning 24
Analysis and Design 25
Development 29

Testing 32
System Architecture 34
Presentation Tier 34
Web Tier 34
Business Logic Tier 34
Data Tier 35
CHAPTER IV RESULTS AND DISCUSSION 36
Initial Strategy Regression 41
Regression Algorithms 43
Initial Strategy Results 43
Subsequent Strategy Classification 49
Classification Algorithms 52
Second Strategy Results 52
Numerical Phase Data Analysis 53
Categorical Phase Data Analysis 54
Classification Implementation and Model Evaluation 57
Front-end Integration 65
CHAPTER V CONCLUSION AND RECOMMENDATIONS 66
Conclusion 66
Recommendations 68
Bibliography 69
Appendix A Evaluation Tool 73
Appendix B System Development Test Cases 79
Appendix C Source Code 82
Appendix D Development Timeline 88
Appendix E Screenshot of the System 89
Appendix F Curriculum Vitae 96
Appendix G Certificate of Proofreading 98

List of Figures

Figure 1.Conceptual Framework....14
Figure 2. Linear Regressed Distribution [17]....18
Figure 3.Lasso Regression Equation....20
Figure 4 Linear vs. Ridge Regression [18]....21
Figure 5 Decision Tree Model generated through Regression....22
Figure 6.Software Process Model....24
Figure 7. Use Case Diagram....26
Figure 8.Entity Relationship Diagram....27
Figure 9.Activity Diagram....28
Figure 10 Network Architecture....31
Figure 11.System Architecture....34
Figure 12.Non-Null Matrix (Non-Nulls represented as black dashes)....37
Figure 13.Number of Non-Null Values....38
Figure 14.Pearson Correlation Heatmap....39
Figure 15 Spearman Correlation Heatmap....39
Figure 16 Distribution (Intro to Philosophy)....42
Figure 17.Scatter Plot (Intro to Philosophy vs English)....42
Figure 18. Transformed Intro to Philosophy Distribtution (Boxcox)....42
Figure 19 Random Forest Regression Residual Distribution Histogram....46
Figure 20 Residual vs Fit Scatterplot....47
Figure 21 Residual Q-Q Plot....48
Figure 22. Correlation Heatmap to sum of raw scores (Pearson)....53
Figure 23.Correlation Heatmap to sum of raw scores (Spearman)....53
Figure 24 SVM Confusion Matrix....58
Figure 25. SVM Normalized Confusion Matrix....59
Figure 26. Gradient Boosting Normalized Confusion Matrix....60
Figure 27. Logistic Regression Normalized Confusion Matrix....60
Figure 28 Random Forest Normalized Confusion Matrix....60
Figure 29. Category/Class Distribution....60
Figure 30 SVR ROC Curve....61
Figure 31.Logistic Regression ROC Curve....61
Figure 32. SVM Confusion Matrix (One-Hot-Encoded)....63

Figure 33.Random Forest Confusion Matrix (One-Hot-Encoded) Binary Classification..63
Figure 34 Neural Network Confusion Matrix 64
Figure 35 Track Type, Number of Science Subjects, Raw Score Category Scatter Plot 67
Figure 36.Logistic Heatmap (Normalized).....73
Figure 37.Logistic Heatmap.....73
Figure 39. OHE Logistic Heatmap (Normalized) 73
Figure 38.Logistic Curve.....73
Figure 40. OHE Logistic Heat map 73
Figure 41 OHE Logistic Curve 73
Figure 43.Gradient Boosting Heatmap.....74
Figure 42.Gradient Boosting Heatmap (Normalized) 74
Figure 44.Gradient Boosting Curve 74
Figure 45.OHE Gradient Boosting Heatmap(Normalized).....74
Figure 46.OHE Gradient Boosting Heatmap.....74
Figure 47.OHE Gradient Boosting Curve.....74
Figure 49. SVM Heatmap 75
Figure 48.SVM Heatmap (Normalized).....75
Figure 51.OHE SVM Heatmap (Normalized) 75
Figure 50.SVM Curve.....75
Figure 52.OHE SVM Heatmap.....75
Figure 53.OHE SVM Curve 75
Figure 55.Random Forest Heatmap 76
Figure 54.Random Forest (Normalized).....76
Figure 57.OHE Random Forest Heatmap(Normalized).....76
Figure 56.Random Forest Curve 76
Figure 59.OHE Random Forest Curve.....76
Figure 58.OHE Random Forest Heatmap.....76
Figure 60. Neural Network Heatmap (Normalized) 77
Figure 61.Neural Network Heatmap 77
Figure 63.OHE Neural Network Heatmap (Normalized).....77
Figure 62.Neural Network Curve 77
Figure 64.OHE Neural Network Heatmap.....77
Figure 65.OHE Neural Network Curve.....77
Figure 67.Naive Bayes Heatmap.....78

Figure 66.Naive Bayes Heatmap (Normalized) 78
Figure 68.Naive Bayes Curve 78
Figure 69. OHE Naïve Bayes Heatmap (Normalized) 78
Figure 70.OHE Naive Bayes Heatmap 78
Figure 71.OHE Naive Bayes Curve 78
Figure 72 Dashboard 89
Figure 73 List of Examinations 89
Figure 74 Adding New Examination 90
Figure 75 Examination Details 90
Figure 76 Edit Examination 91
Figure 77 Application Form (Upper) 91
Figure 78 Application Form (Lower) 92
Figure 79 Student Data View (Upper) 92
Figure 80 Student Data View (Lower) 93
Figure 81 Student Report Generation Output 93
Figure 82 Exam Report Generation Output 94
Figure 83 Student Status tracking form 94
Figure 84 Student Electronic Slip View 95

List of Tables

Table 1.Definition of Terms.....7
Table 2.Comparison Table of Related System.....13
Table 3.Model Scores.....46
Table 4 Numerical Testing Phase Model Scores 55
Table 5 Categorical Testing Phase Model Scores 55
Table 6 Model Parameters 56
Table 7 Implemented Classification Model Scores.....58
Table 8 One Hot Encoding Results.....62
Table 9 Binary Classification Model Results 64

CHAPTER I
INTRODUCTION

Background of the Study

The testing and evaluation center compiles and stores a large volume of data, such as the demographics found when students register for the College Entrance Test and their respective CET results. However, the large volume of the data prevented the WMSU TEC from garnering valuable insights, as manually sifting through this data and applying statistical tools was exorbitantly time-consuming. The researchers could use it to create an application that is both a prediction system and data visualization software that can yield helpful information for the WMSU TEC.

The primary technological gap lies in WMSU TEC storing its data in physical form and conducting much of its examination processes manually. Creating an online web application allows much of the automation of the examination scheduling processes.

The illustration of data in an easy-to-understand format that communicates its information visually provides an essential tool for decision-makers to base their decisions. Providing them with a visual representation of the data will give them meaningful insights into large volumes of data in a relatively short time. In recent days, and as seen in all sectors, the importance of data visualization has only grown.

A study has developed a prediction system for future applicants that would predict the range of their WMSU CET score and recommend areas for improvement to improve their chances of admission to their desired college. Simultaneously, the researchers also created a web application for use by the WMSU TEC that automates various examination scheduling processes such as room assignment, student application approval, and exam detail specifications.

Statement of the Problem

The primary problem that this study focuses on is that there is no previous study regarding the effectiveness of a student's senior high school grade and personal demographics as a factor in determining their performance on any college standardized tests.

Secondarily, both the WMSU's Testing and Evaluation Center and its Admissions Center have accumulated a large volume of data that have the potential to be used to assist them in making decisions that would be beneficial to both the school and any future college applicants. However, due to its large volume, they did not have the tools necessary to garner valuable information from the data. Thus, the potential insights that may be useful to the WMSU TEC were inaccessible by conventional means.

Apart from this, Corona Virus Disease Pandemic (COVID-19) Restrictions placed a strain on universities with processes requiring face-to-face steps, such as registration of college applicants for the College Entrance Test, which could be accomplished online more efficiently for both the applicants and the university.

Objectives

The researchers aimed to achieve the following objectives throughout the study:

Main Objectives

- To test whether student demographic data is a significant factor in determining the WMSU CET Score.
- To test whether student senior high school academic performance data is a significant factor in determining the WMSU CET Score.

Specific Objectives

- To create a web application that the WMSU TEC can utilize to digitalize the process of registration and examination management for the College Entrance Test.
- To create a platform for future college applicants to register for the WMSU CET online.
- To create a prediction system that predicts the range of the student's college entrance test score based on their inputted data and recommends areas of improvement to increase their chances of attaining a higher grade

Significance

Completing this study and its objectives would allow WMSU to recommend areas of improvement for schools around the country that will allow their students to achieve higher passing rates at the CET, thus increasing their chances of success, which will be beneficial to the institution and the students. Additionally, it uses any information contained within the raw input data. The completion of this thesis was unique in that it was the first of its kind to be conducted on the proposed data set, which consists of Western Mindanao State Senior High Grades and their corresponding College Entrance Test Results, to determine the significance of the Commission of Higher Education-approved K-12 Grading System as a factor in predicting the outcome of a standardized college entrance test.

The primary user and beneficiary of this system will be the WMSU TEC. A higher CET score will increase an applicant's chances of entering their desired college as it is a significant factor in their selection process. Therefore, secondary beneficiaries would be Western Mindanao State University and its students and college applicants, depending on the decisions made or recommended by the Western Mindanao State University TEC using the data gleaned from this study, as well as the various schools that could encourage their students to improve in a subject area so that they would have increased performance in the College Entrance Test.

Scope and Limitations

Scope

The system is available for use at the Western Mindanao State University. The data set was limited to the university's past college applicants, which served as the basis for developing the data mining system for future college applicants.

The system has two parts. The first is to incorporate future inputs such as student demographics and college entrance test scores into the prediction accuracy and data visualization produced by the classification algorithm created solely for data mining. The demographic data that the study used included the students' parents' combined income and the high school from which they graduated.

The second use is to provide a platform for the WMSU CET to allow WMSU college applicants to register for online entrance examinations to minimize the need for face-to-face interactions.

Limitation

The research is only intended for the Western Mindanao State University, its prospective college applicants, and the part of the school administration that handles the examination, which is the WMSU TEC.

The Western Mindanao State University Testing Evaluation Center and Admissions Center needed to grant permission for this study to use the students' senior high school grades, demographic data, and college entrance test results. The population is limited to Zamboanga City and the surrounding area's senior high school graduates intending to enroll in the WMSU Main Campus. The information present in these data sets will exclusively belong to WMSU college applicants for use in the system after data preparation.

The primary issue raised by this study arose from privacy concerns of the involved parties, mainly the past and future college applicants of WMSU. Data

anonymization was a high priority during the data preparation phase of the study to remedy this.

The study included only those students who passed the WMSU CET and enrolled in the university as only those students were required to submit a transcript of records.

Features

The system allows users who would be WMSU College applicants intending to register for the WMSU Main Campus and any external studies units to register for the WMSU CET. The system also allows the WMSU TEC Administrators to Schedule Examinations and Specify the Rooms for the Examinations.

The system will not allow the administration of WMSU CET online, as it will only allow the WMSU TEC Administration to schedule examinations and applicants to register for the examinations.

The target users of the system are the WMSU college applicants in the context of using the system to register online for the WMSU CET and of the WMSU TEC in the context of gaining insights from the data mining functionality of the system as well as managing college applicants for the entrance examination.

Definition of Terms

Term	Definition
1. Western Mindanao State University (WMSU)	The university where the study will take place, located in Zamboanga City, Philippines
2. Testing and Evaluation Center (CET)	The part of WMSU administration that handles the administration of the College Entrance Test, as well as other evaluation Processes
3. Machine Learning	Algorithms that improve automatically through experience and by the use of data.
4. Data Mining	The process of extracting and discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems.
5. Data Set	A set of similar Data to be used as inputs in algorithms
6. Data Point	An element in the data set that consists of attributes.
7. Attribute	The variable value in each column of all data points.
8. Linear Regression	A predictive machine learning model that tries to find the linear relationship between an independent variable and a dependent variable.
9. Classification	A Machine Learning Process that attempts to predictively categorize a data point based on previously trained attributes and correlations.
10. Feature	A predictor variable. This is the variable that is used to predict the label and is the main variable used to train and test the data. Multiple features can be used in prediction models.

11. Label	The variable being predicted, this is the target variable in which the model is trained to predict using feature variables.

Table 1.Definition of Terms

CHAPTER II
REVIEW OF RELATED LITERATURE

Related Studies

Prediction using Academic Performance

The novelty of this study comes from using high school GPA performance as a determinant for predicting standardized college entrance test scores. Previous studies merely explored the idea of the CET scores correlating with college academic performance. Many of these studies found a positive correlation between standardized test scores and college academic performance [1].

Allensworth and Clark found a consistently stronger relationship between HSGPAs and college performance than ACT scores. However, the strength of the relationship varies by high school [2]. The compiled literature showed that researchers agreed that HSGPAs are more robust predictors for college outcomes than standardized tests such as the ACT or SAT scores. They further solidify these claims by using statistical analysis with data from a 4-year college with a population of 55,084.

Using variance analysis, Tatar and Düstegör drew a similar conclusion that using course grades achieves higher accuracy in predicting college performance [3]. Furthermore, they demonstrated that individual course grades are better for short-term prediction, and GPAs are better for long-term predictions of academic performance.

Academic Performance and Data Mining Techniques

The prediction of student performance using Multiple Linear Regression (MLR) techniques is not novel, with one example being a model developed by Yang et al., which was able to get the optimal scores by combining Principal Component Analysis with traditional MLR [4]

As proven by Huang and Fang, various academic fields apply this technique. They were able to develop a set of multivariate regression models for predicting student performance in a particular course, specifically Engineering Dynamics, by using the cumulative GPA earned in four prerequisite courses, including Engineering Statics, Calculus I and II, and Physics, as well as scores from three previous exams, as independent variables. [5] The circumstances of their study are somewhat parallel with ours, with a dataset consisting of the GPAs of various subjects and a single dependent variable that this study is trying to achieve.

These studies only focus on the correlations between CET Scores, college grades, and high school GPAs with college performance. This study is different in that it focuses on the relationship between the Highschool GPA and the College CET scores and uses it as a basis for a prediction system between those two variables.

Foreign Studies

Applying both classification or association and clustering data mining techniques is not a novel idea. Meng, H. et al. used both techniques to analyze geoscientific research [6]. They used both concepts independently, with the clustering algorithm handling clusters with identifiable characteristics such as shapes, sizes, and densities. In contrast, they used association analysis to mine the continuous attributes to glean potentially helpful and insightful information in geoscientific applications. They utilized quantitative association rules to deal with the relationships in their data.

Ramasamy and Nirmala developed a similar system that predicted disease with a similar approach, using clustering and association to mine data on a dataset with multiple data types [7]. They used the Association Rule mining technique to find relationships between the patient's data and the hospital's database. In contrast, they used keyword-based clustering algorithms to find the disease with the highest probability of affecting the patient. They have stated that using both algorithms allows the system to have better efficiency and quicker processing.

Haraty R. et al. developed a clustering algorithm that outperformed the K-means algorithm called the G-means algorithm [8]. They instead used the HITS algorithm to

correlate both a user network and a course network to a server as the basis for their Dynamic Online Course Recommendation. They considered the users' knowledge level instead of other studies that focus primarily on accuracy. Their experiments with real e-learning datasets showed that their model did well when recommending online courses, which means that taking the user's knowledge level into account when making course recommendations is a good idea.

A qualitative review by Khalid et al. showed that at least half of the literature regarding implementing recommendation systems for Massive Open Online Courses (MOOCs) is for learning and course recommendation systems. [9] They instead used the HITS algorithm to correlate both a user network and a course network to server as a basis for their Dynamic Online Course Recommendation. They took into account the users' knowledge level as opposed to other studies that focus primarily on accuracy. The results of their experiments in real e-learning datasets showed that their model performed well in online course recommendation with an increase in learning result quality, which indicates that there is a benefit when taking into account the user's knowledge level in course recommendation.

A qualitative review by Khalid et al. showed that at least half of the literature regarding the implementation of recommendation systems for Massive Open Online Courses (MOOCs) are for learning and course recommendation systems. [10] Content-based filtering was the most common implementation method for both types of systems. However, in conjunction with content-based filtering, applied using hybrid algorithms. [10], At the same time, course recommendation systems also used collaborative-based filtering methods [10]. Both types of systems in recent years have seen an upsurge with the use of neural networks, pattern mining, and deep learning for the preprocessing of data.

A study by Goga et al. found that the Random Tree Performance Algorithm is the optimal algorithm for predicting academic performance from factors such as family background and previous academic achievements [11]. he Goga et al. study's conclusion is consistent with the findings of other studies cited by Goga et al. Similar studies discovered that decision tree classes are the most accurate algorithms for predicting students' academic performance. [12] Grewal and Kaur became capable of creating an intelligent course recommendation system. Then, they used the association rule to

determine which students' and individual characteristics were in line. Additionally, they can also utilize the Fuzzy Set and Rough Set theories. In contrast, they deduced the classification rules for the prediction model from relevant data using selective sampling processes. The accuracy is first validated using a third-party testing data set. Grewal and Kaur recommended that the Student Recommender System is essential in helping students select courses of their choice. They also conclude that it provides practical advice and counseling for 10+2 students.

These studies focused on applying combined data mining techniques for educational recommender systems. They showed that there could be correlations between secondary-level academic performance and tertiary-level academic performance. The outcomes of the studies by Yang and Jiang, Khalid et al., and Goga et al. all point to a feasible recommendation system based on data mining techniques. However, they used data of a different nature to the one used in this study. It mainly consisted of categorical data such as past achievements, knowledge levels, and family background, among other things. The variables in this study are primarily continuous. The study attempts to identify a continuous variable, namely the range of the CET score that the student will receive based on the input variables. It contrasts with the categorical values, namely the predicted course recommendation systems that the studies mentioned above.

Local Studies

These studies focused on applying combined data mining techniques for educational recommender systems. They showed that there could be correlations between secondary-level academic performance and tertiary-level academic performance. The outcomes of the studies by Yang and Jiang, Khalid et al., and Goga et al. all point to a feasible recommendation system based on data mining techniques. However, they used data of a different nature to the one used in this study. It mainly consisted of categorical data such as past achievements, knowledge levels, and family background, among other things. The variables in this study are primarily continuous. The study attempts to identify a continuous variable, namely the range of the CET score that the student will receive based on the input variables. It contrasts with the categorical

values, namely the predicted course recommendation systems that the studies mentioned above.

Synthesis

The development of the online registration application as part of this study was not the first time that the WMSU TEC had implemented online registration for students. Previously, they accepted applications from the students online through the Google Forms application. The limitation of this application meant that this could only be implemented as the first step of the process, still requiring the student to come to school to complete the payment process and receive their examination slips, which contain details about the examination the student was assigned to and must be presented to the proctor when taking the examination. Developing the system meant that the electronic slip could instead be claimed online and printed and presented to the proctor on the day of the examination. However, it did not consider the process of payment as part of the development of the system.

Various registration software programs also exist. Tierny states that some of the top application/class management software of 2021 include Jotform, Regpack, Enrollware, and Student Management by ACEware. This software allows the user to create form questionnaires that students can use to register for a class or course. It is contrary to the requirement specified by the WMSU TEC that students register for an examination and allows TEC administrators to schedule examinations and assign students to them. Some of the software, such as Student Management by ACEware, generates statistical reports based on the data from the form. The study's software also generates reports based on demographic data from the students. However, the researchers also developed a visualization system to demonstrate statistical insights better using bar graphs, heatmaps, and plot charts.

An examination scheduling application is part of the INFOSILEM | Berger-Levrault educational management suite. It allows the user to schedule examinations in multiple rooms and assigns students to those examinations in a single step. This application focuses on schedule quality management in that it takes into account student information. It schedules the examinations according to the student's needs and the

instructor's preferences. This type of generic software is limited in that the requirements outlined by the WMSU TEC require a more bespoke approach to software implementation. INFOSILEM does not accept and store student demographic data, nor does it store the CET scores of the students in the system. It cannot create forms required for student registration. It is primarily intended for educational institutions to administer exams to regularly enrolled students.

Comparison Table of Related Systems

Attribute	**Google Forms**	**Form Management Software**	**INFOSILEM**	**This study's System**
Able to create and distribute Forms and store data into a database	✔	✔	X	✔
Able to generate statistical reports on existing data	✔	✔	X	✔
Able to Visualize Statistical Reports with Plot charts, Bar Graphs, and Heatmaps	✔	✔	X	✔
Able to Schedule Examinations	**X**	**X**	✔	✔
Able to Assign Students to Examinations	X	X	✔	✔
Able to Store Examination Scores	X	X	X	✔
Able to Generate Examination Slips for admission	X	X	X	✔
Able to Show students their application Status	X	X	X	✔

Table 2.Comparison Table of Related System

Conceptual Framework

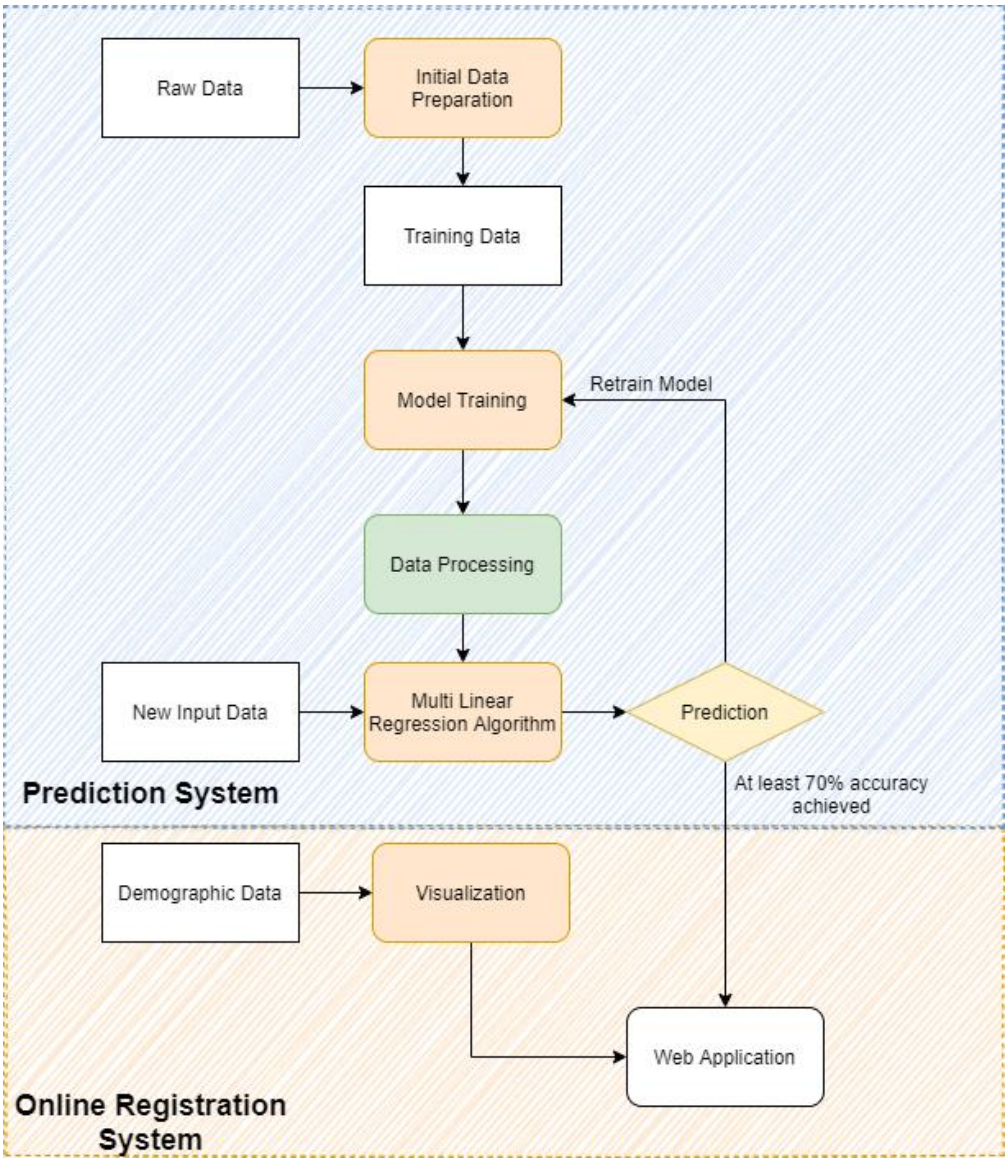


Figure 1.Conceptual Framework

The system's primary goal is to use data from the Western Mindanao State University Testing and Evaluation Center and the Admissions Center to predict a student's college entrance test score outcome using regression or classification.

Before being inserted into the database, the historical samples in the dataset are most often physical or document-based. The data mining algorithm could look for patterns and try to recognize them. The standardized raw data is inserted into a

readable CSV format before being fed to the regression algorithms. It is used to train the regression machine learning algorithm and determine whether the algorithm accurately predicted the data point's value. The baseline of 70% accuracy was determined to be satisfactory for the regression problem, with a 10% margin of prediction error. Future applicants to the school will directly input their data into the database through the online registration functionality developed alongside the data mining system. Reports for the WMSU TEC assist in decision-making through the data mining system generated using data visualization.

Following the successful development of the machine learning algorithm, the researchers moved on to data visualization, visualizing the patterns, insights, and knowledge gleaned from demographic data in a readable and easy-to-understand format.

CHAPTER III
METHODOLOGY

Research Design

This study aims to predict a student's WMSU CET performance using their past academic GPAs as a basis for prediction by utilizing regression or classification algorithms as the primary techniques utilizing a machine learning algorithm.

This research employs an applied research design, primarily since these researchers experimented with various regression and classification techniques to predict students' college entrance test performance using historical data from their high school grades as a predictor of high school performance, which could be influencing the student's college entrance test performance, thereby potentially developing the basis for a basic prediction system for student college entrance test performance.

Data Gathering Techniques and Procedures

This research aims to observe patterns using Data Mining techniques focusing on regression or classification. The data set consists of each student's GPAs in their respective high schools courses. The study then used the prepared data to train the data mining algorithms.

The data set was historical data, mainly the past three years of the college application and CET results acquired from WMSU TEC. This data exists in a physical copy of their transcript of records, which the researchers requested access to from the WMSU Admissions Center through the WMSU TEC. To ramp up the digitization process, the researchers scanned the documents first, then encoded the data into an excel spreadsheet with various columns for each subject.

The data used for alteration and testing came from WMSU historical documents and records, primarily from the Registrar's Office.

In particular, the system used three years of past data which was requested and encoded into a format readable by the system, using a program which accepts data inputs and adds them to a database, to increase the efficiency and reduce the time it takes for all the data to be encoded.

The development of the online registration system and machine learning concepts allows the data mining functionality of the system to accept new inputs, further increasing its reliability and longevity. First, the data set is cleansed and validated using data cleansing and validation methodologies.

Statistical Tools

In this research, the researchers utilized a variety of data mining techniques in order to find the technique with the best accuracy when predicting the range of the student's CET Scores. These techniques were used to determine whether the data acquired, which in this case was the students' demographic data as well as their past academic performance taken from the GPA records of their High School Grades, affected the CET results of the student. Due to the nature of the data, the study has multiple independent variables that are continuous for SHS GPA and some non-continuous values that the dependent variable could be based on. The researchers focused on testing different regression analysis techniques that accept multiple independent variables and assigns corresponding weight to them accordingly to produce the expected dependent variable. The model best suited for the dataset was chosen, taking into account its size, the type of data present, degree of linearity, collinearity, variance, presence of outliers, bias, etc. *See Algorithms Section.*

Linear Regression

The baseline Regression algorithm is the Linear Regression Machine Learning Algorithm. Linear Regression Algorithms attempt to predict the dependent variable (in the diagram represented *y*) by trying to identify the best linearly fit line between *y* and the dependent variable, its coefficient or slope (represented as b_1), and the intercept (represented as b_0) [13].

$$y = b_o + b_1 x$$

In simpler terms, a simple linear regression tries to draw a line across a plane between two variables, the figure below will represent the dependent and independent.

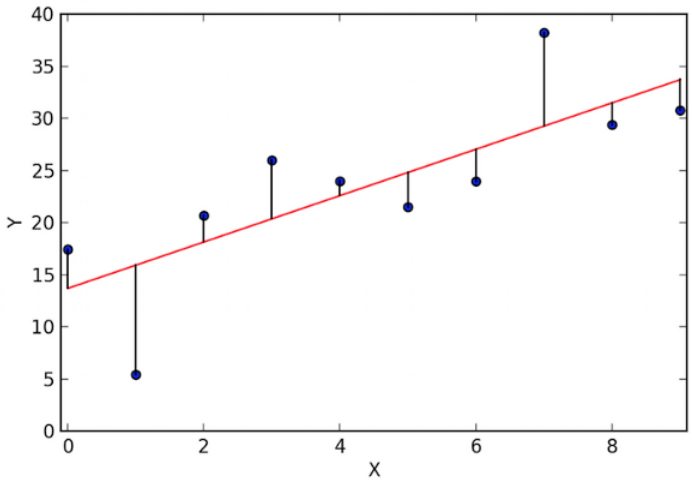


Figure 2. Linear Regressed Distribution [14]

It does this by utilizing the ordinary least squares method that tries to ascertain the coefficient of the slope that would best fit the data values of each element by using the minimum sum of squares to find the best line through the data elements. With this, the algorithm can create a rough estimate of where the dependent variable would be if given the value of an independent variable, with a reasonable margin of error.

Multiple Linear Regression

Multiple Linear Regression (MLR) is similar to Linear Regression. It uses the same logic as the minimum sum of squares to determine the best fit line through the

data, with which it predicts the position of the dependent variable, taking independent variables as inputs. While Simple Linear Regression only takes one independent variable to predict the value of the dependent variable, MLR can take multiple independent variables, assigning corresponding weights to their values in the calculation. It means that one independent variable could be more critical in determining the value of the dependent variable than another independent variable [13].

However, MLR requires various assumptions to be true:

- That there is a linearity in the relationship between the dependent variable and the independent variables.
    - This will be done by selecting 5 random independent variables as a sample and testing each of them individually with the dependent variable on a scatterplot diagram to check if they have a linearity relationship. Linearity would be proven if at least 4 out of the 5 chosen independent variables show a linear relationship with the dependent variable (80% success rate).

- That there is no multicollinearity between the independent variables.
    - This is apparent and immediately assumed to be true, as in any educational context, the GPA of a single subject does not directly affect the GPA of another subject.

- That the observations for the dependent variable are selected independently and at random.
    - This is immediately assumed to be true as there was no further criterium used for choosing samples for the data set apart from what was described in the population section (3.1.1) of this document.

- That the Regression Residuals or the error differences between the predicted and actual outputs are normally distributed.
    - This will be done using a Histogram and Q-Q-Plot.

Lasso Regression

Most minor absolute shrinkage and selection operator, or LASSO Regression, is an algorithm based on the Linear Regression model that utilizes shrinkage. The data values are shrunk towards a center point, such as the mean. This procedure performs better in models with fewer parameters and high multicollinearity between the predictor values. The equation for this is represented similarly to the linear regression equation [15].

$$\sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Figure 3.Lasso Regression Equation

The main addition is the presence of lambda as a factor in determining the coefficients' sum of the absolute value. The lambda, in this case, represents shrinkage. With a lambda value of 0, all features are considered part of the equation. When it approaches infinity, it implies that no feature is selected. Thus, the higher the lambda value, the more minor features are considered for the model. A higher lambda value means a higher model bias, while a lower one means a higher variance.

Ridge Regression

Ridge Regression is another type of regression algorithm that is best used when there are more coefficients or factors in the prediction model than observations. In contrast to Multiple Linear Regression, this algorithm is better suited for multicollinearity between variables, which might have been a possibility in the senior high school GPAs of the students. It considers multicollinearity, which could be interpreted as a relationship between independent variables and extreme outliers that could affect linear regression models. It incorporates them as bias, known as the "Ridge Regression Penalty." The introduction of bias into the model will generate a significant, though not entire, drop in invariance [15].

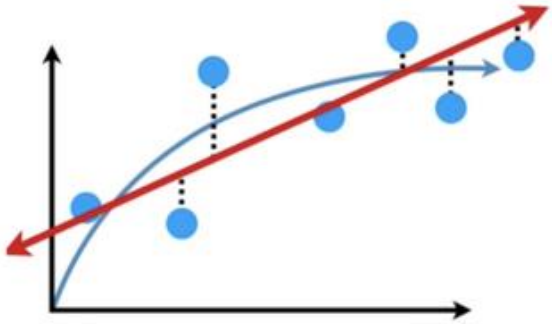


Figure 4 Linear vs. Ridge Regression [15]

The figure demonstrates that introducing the Ridge Regression Penalty to the model (represented by the blue line) allows it to better fit the data set for longer term predictions as opposed to the linear model (represented by the red line) which has a high residual output.

As opposed to using the Ordinary Least Squares method of Linear Regression, it instead uses Variable Standardization, which uses the subtraction of the averages between the dependent and independent variables and a division divides the result with the standard deviations. The Ridge Regression follows a formula that can be represented as:

$$Y = XB + e$$

In this equation, the dependent variable is represented by Y, while the sum of the squares of coefficients for independent variables are represented by X, and the regression coefficients used to evaluate the weight of the independent variables is B, while *e* represents the sum of errors as residuals.

Tree Regression

Tree Regression uses Binary Recursive Partitioning. It iteratively bifurcates the data into branches or partitions. It continues for each partition into smaller groups as the method moves up each branch. It does this after splitting the data into the training and validation sets. After doing so, the model must be 'pruned,' in which the model's branches are removed to reduce overfitting by using the validation set of data.

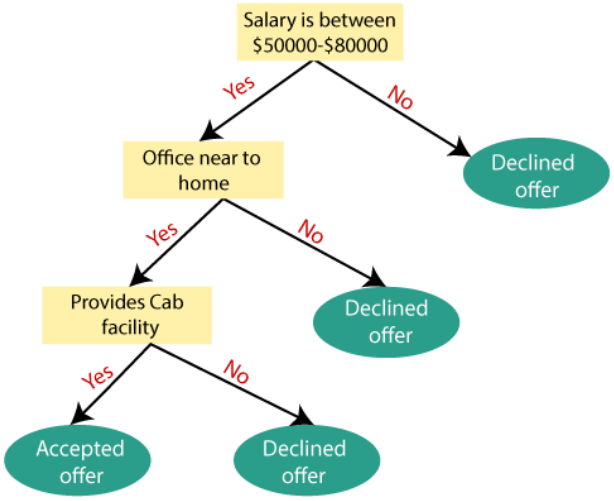


Figure 5 Decision Tree Model generated through Regression

The figure represents a decision tree model. Decision trees are advantageous because they require less data cleaning, the model's performance is not heavily reliant on the linearity of the data set, and the hyper-parameters to be tuned are almost null. It is, however, prone to over-fitting, which could be better handled by using the Random Forest algorithm instead. However, Random Forest relies on supervised training data and, in that instance, would be inappropriate for the unlabeled data set of this study [15].

For all algorithms, the researchers used a 40:60 split of the total data to train the model, where 60% of the data will be used to train the model, 20% will be used to test it, and 20% will be used to validate it. The split was conducted randomly to minimize bias.

Analytical Tool

Python was used as the primary analytical tool, using python libraries such as pandas, matplotlib, and numpy for the initial exploratory data analysis and the results analysis

Software Process

Planning

This research requires the development of a web-based application that provides a platform for college applicants to register for the WMSU College Entrance Test online, which will streamline the process of registration and admission while reducing the risk of contracting COVID-19 for both the students and the administrative officials of the university, thus increasing the security of both parties. It also integrates the data mining algorithms into the application during development.

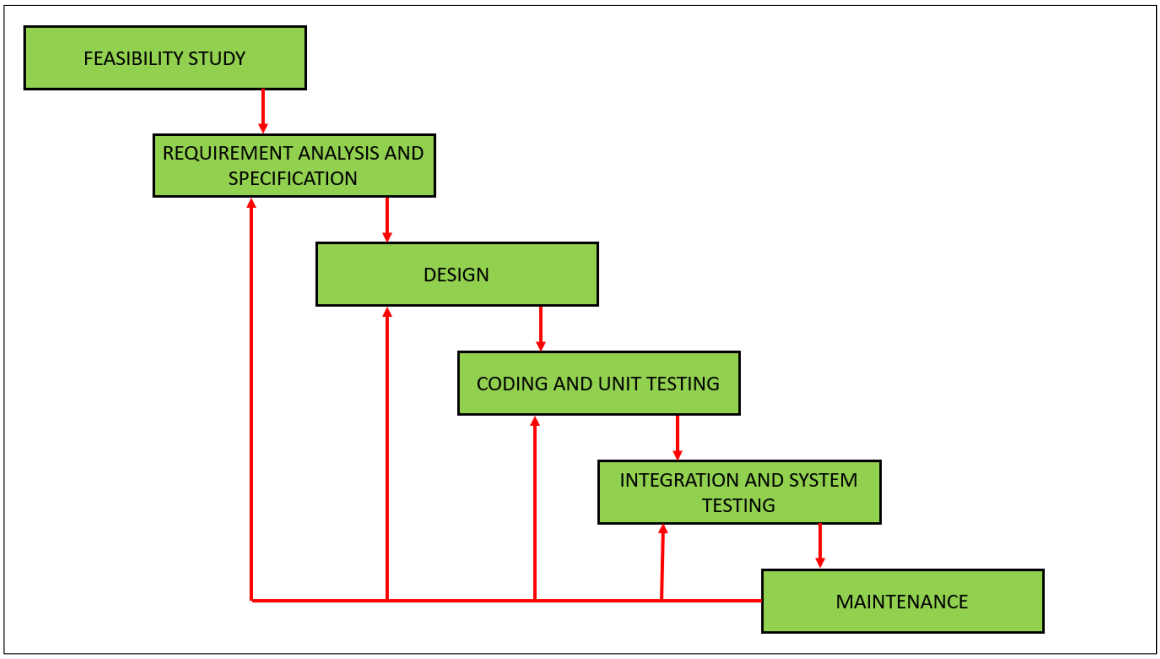


Figure 6.Software Process Model

This study's model for development was the Hybrid Waterfall and Iterative Model. This model was chosen because some of the specifications this model recommends were met. It includes how the requirements of the project are well defined and understood, the new technology, which in this context is the use of data mining and machine learning algorithms, was being learned and used by the researchers, some working functionality can be developed early for feedback, and implementing inevitable changes to the system has a low cost.

The new technology involved creating a web-based database with data mining functionality for future inputs. As this type of web application was one of the very first that the researchers created, an iterative model was the most effective method for the design and implementation as the researchers learned, giving opportunities to request feedback early on in the design phase, reducing the risk for the system as much possible, while still adhering to the requirements given.

The iterative model also allowed for the implementation of new features, as this part of the study the researchers considered optional, allowing the design and development of the system in parallel with the main focus of the study, separately in a modular fashion along with the primary requirement.

The web-application platform will allow future researchers to acquire future data to train the algorithm further. At the start of each academic year, there will be a new batch of college applicants that will register their information as part of the requirements for admission. The development will consist of four phases: planning, development, testing, and implementation.

Analysis and Design

Based on the request of the WMSU TEC, outlined are the first iteration of the requirements from the system:

- The system should be able to allow college applicants to submit CET registration forms.
- The system must allow the WMSU TEC to view these forms.
- The system must allow the WMSU TEC to approve or deny registration.
    - The system must allow the WMSU TEC to specify the reason for denying a registration
- The system must notify the student (via email or text) whether their registration was approved or denied.
- The system must allow the WMSU TEC to assign the student the Room and Schedule in which they will take the exam.

- The system must notify the student (via email or text) their assigned room and schedule for the exam.
- The system should be able to generate data visualization based on the data mining algorithms that will be used to develop it.
- The system should be able to re-visualize new data based on the new inputs it will accept in the future.

Use Case Diagram

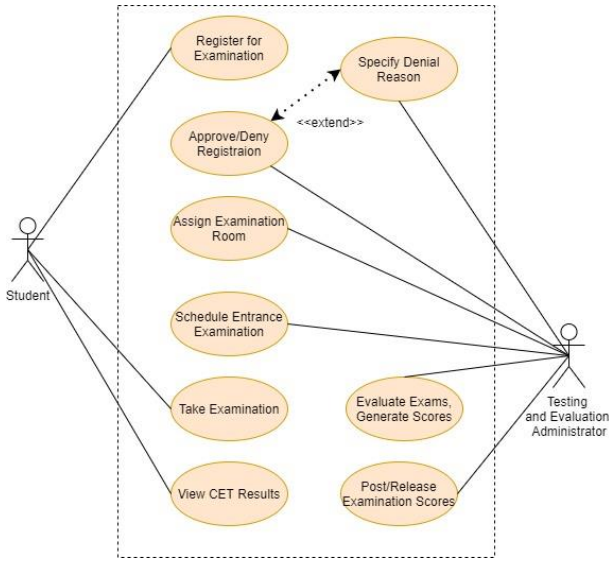


Figure 7. Use Case Diagram

The primary transaction of the system used by the client can be described as follows:

- The student registers for the examination by submitting the requirements, which include the registration form, which they have to acquire from the TEC before filling it out, among other things.
- The TEC approves the registration.
- The TEC assigns the student a room and a schedule for when they will take it on.
- The student takes the examination.
- The TEC evaluates the examinations and posts/releases the scores.

- The student can view the result and claim a document from the TEC that certifies that they achieved that particular score.

Much of these steps can be digitalized for a more streamlined process, especially the registration phase for the student. If it is done online, the bulk of the workload usually meant for distributing and accepting registration forms can be reduced, increasing the efficiency of the transaction.

Data Model

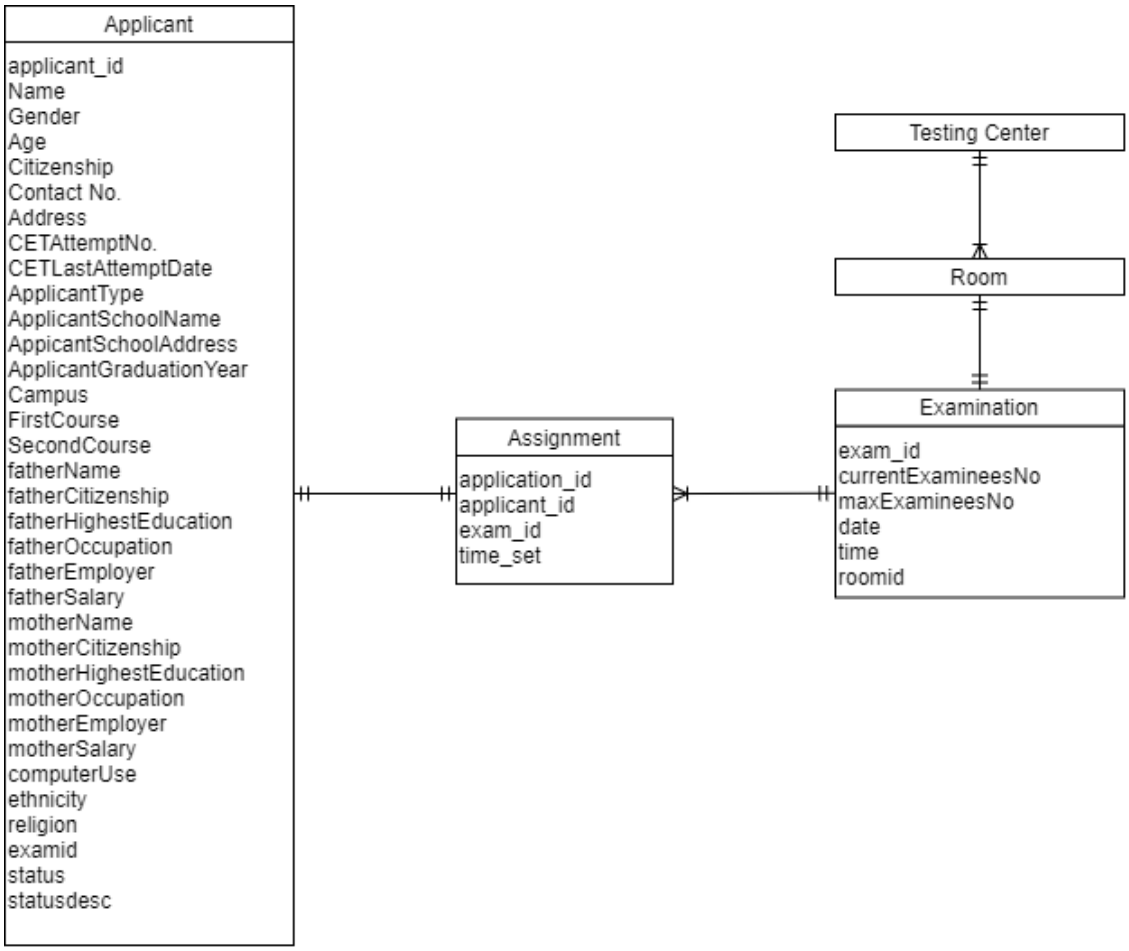


Figure 8.Entity Relationship Diagram

For the online registration system, the system allows the users to fill out a data form, which will then have to be approved by the WMSU TEC administrator. Once approved, the student's data will be inserted into the system and the TEC administrator

can then assign the student the date of the exam and the room number. Once this was done, the system assigns an exam ID code to the student.

Activity Diagram

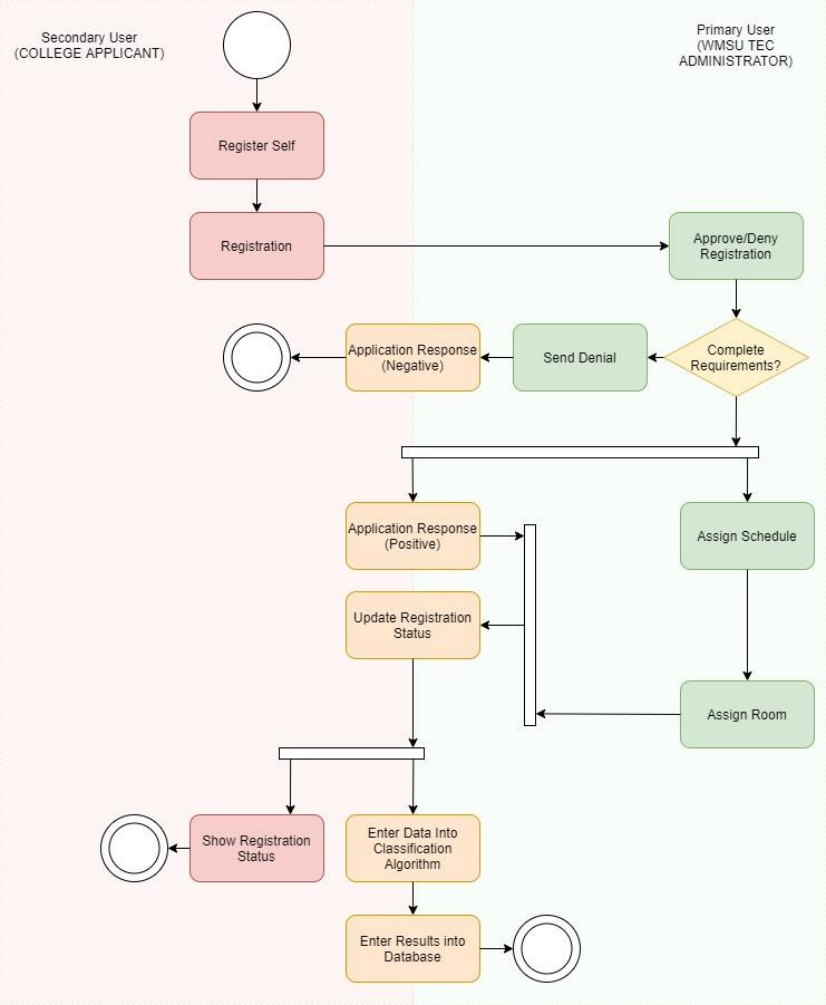


Figure 9.Activity Diagram

The activity diagram describes the general flow of the system. It starts with the student's registration, which is sent to the primary user for approval. Here, the primary user can decide if the registration meets the requirements outlined by the TEC and, once approved, will assign a schedule and a room to the college applicant. One of the main requirements that the system cannot handle is payment. One way to circumvent this is to create a 2-step approval system. The first step is to approve the main requirements without the payment. After the main requirements are approved, the college application

will be notified that they may pay the examination fee at the designated administrative building in WMSU. The system can confirm the payment. The method by which the system will approve the payment will vary according to the policies outlined by the WMSU TEC. One possibility is submitting the receipt number, which will be cross-checked with the receipt database in the administration's finance wing. Other methods were outlined as the study underwent the software development phases. Once fully approved, the system will then run the data through the data mining algorithms before entering the results into the database.

Development

The software was developed using the Visual Studio Code IDE in a virtual environment installed with the plugins specified below. The researchers then deployed the web application to the Heroku hosting service with the domain name "wmsu-tec.herokuapp.com." At the same time, the system was tested on a local server.

Since the study involved the development of a web-based product, the deployment was done through an online domain-hosting service, which can later be migrated to a physical server in the Western Mindanao State University IT Building. The system's database was also deployed through an online database management system.

Software Requirements

OS: Windows 7 or later, Mac OS 10.12 or newer, OS X 10.9 or newer, Android 5.0 or newer, iOS 11.4 or later.

The packages used for development and deployment of the website are described in the *requirements.txt* in the workspace. The content for the file is as follows:

- asgiref==3.4.1
- boto3==1.18.53
- botocore==1.21.53
- certifi==2021.5.30
- chardet==4.0.0

- charset-normalizer==2.0.6
- click==8.0.1
- colorama==0.4.4
- Django==3.2.7
- django-crispy-forms==1.13.0
- et-xmlfile==1.1.0
- greenlet==1.1.2
- idna==3.2
- ijson==3.1.4
- jmespath==0.10.0
- jsonlines==2.0.0
- linear-tsv==1.1.0
- mysqlclient==2.0.3
- openpyxl==3.0.9
- python-dateutil==2.8.2
- pytz==2021.1
- requests==2.26.0
- s3transfer==0.5.0
- six==1.16.0
- SQLAlchemy==1.4.25
- sqlparse==0.4.2
- tabulator==1.53.5
- unicodecsv==0.14.1
- urllib3==1.26.7
- xlrd==2.0.1

Hardware Requirements

The system requirements for the user will depend entirely on the internet browser that they will be using to access the application, as it is web based. These are the usual minimum requirements for common internet browsers:

- CPU: Pentium 4 or newer with SSE2
- RAM: 512MB

- Storage Space: 200MB

The system requirements used for development are as follows:

- CPU: Intel Core i3 or Newer
- RAM: 8GB
- Storage Space: 128GB

Network Requirements

- Network Min. Bandwidth: 500kbps

The Network Requirements follow a basic Network Architecture typical of Web Hosted Applications. The parts of the Network are as follows:

1. 1 Website Server
2. 1 Client
3. Dedicated Internet Line
4. Public IP Address
5. Domain Name System (DNS)

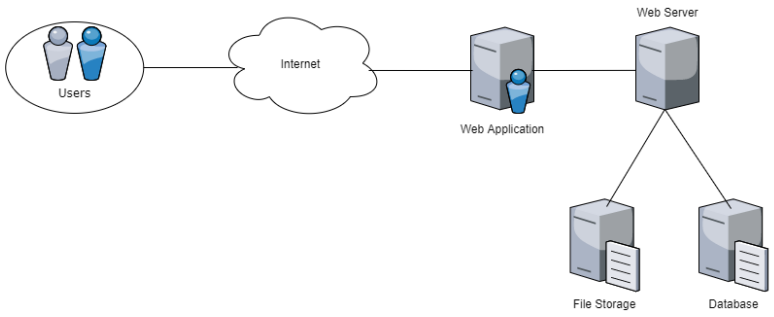


Figure 10 Network Architecture

Development Tools

Since the software will be using Python for both the data mining functionality and the backend framework for the web application, the researchers used Visual Studio

Code to develop the system. However, the development also used various frameworks to reduce the time needed for development.

`Visual Studio Code: Visual Studio Code is a multilingual programming IDE developed by Microsoft that supports a variety of plugins, packages, and languages that will fit the developer's needs. HTML, CSS, PHP, and JavaScript support make this a robust IDE, especially for web developers.

Back End Frameworks

- Django – A Python-based framework that allows the client browser and the database to communicate with each other.
- Sci-kit Learn – A Python-based software machine-learning library that has various machine learning models that can be implemented with a given data set.
- NumPy – A Python-based mathematics library containing various functions for statistical analysis.
- Matplotlib – A Python-based statistical library that generates graphical outputs such as charts, graphs, and other figures based on numerical values as inputs.

Front End Frameworks

- Bootstrap – A front-end HTML/CSS framework for designing the User Interface (UI) of the website.

Testing

The researchers emphasized the testing of the machine learning prediction model, which means that the testing of the system to be deployed was secondary. Nevertheless, all the functional requirements were still tested to ensure that the system was usable for future use. Testing the machine learning prediction model is discussed in a later section of this document.

The researchers conducted an Integrated type of testing, wherein separated features were combined and tested as a group, grouped by the respective modules that the features were a part of. *See Appendix B*

The scope of testing mainly included the three primary modules necessary for the system's functioning and constituent features. These three modules were the Application of the user, Login Validation of the user, and Exam Scheduling Functionality.

To evaluate the system, future studies can implement a user evaluation test to ensure that the system is working as intended. The user evaluation will consist of survey-type questions for the two user types, the primary and the secondary. These will ask the user to rate the system's UI, UX, and ease of use compared to the original face-to-face method of the system's digitalized transactions. The feedback form will be based on the 1004INT – Information Systems UID Feedback Framework. *See Appendix A.*

Researchers will then apply statistical methods to the results, after which they can consider the web application implementation if the results report at least an 80% satisfaction rate. Future researchers can also conduct another user evaluation test for the primary users that asks them about the efficacy of the data visualization project developed according to the data they provided. These evaluation surveys will ask the user if the data visualization was deemed proper and if there are any improvements that future researchers could make. As there are no previous data analysis methods, a success rate of 80% with a 10% margin of error will be considered a success.

System Architecture

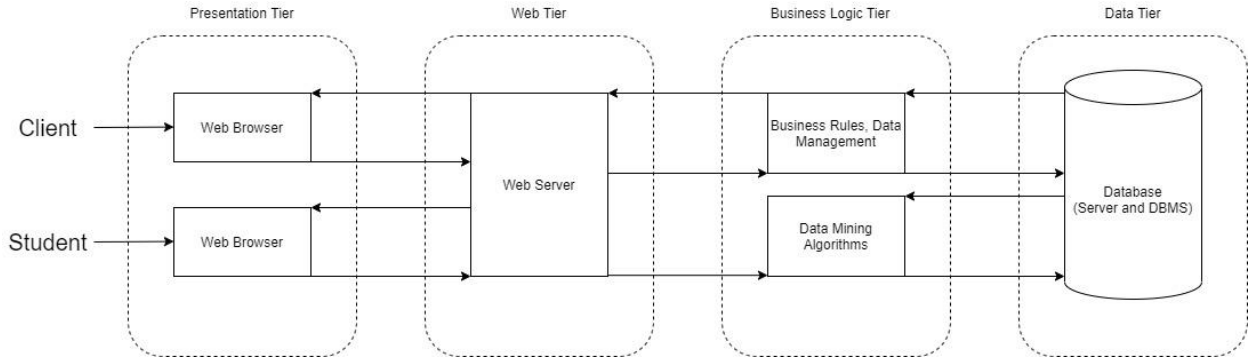


Figure 11.System Architecture

Presentation Tier

For the system architecture, the researchers opted for a 4-tier system architecture as the application is web-based. In the presentation tier, the primary user will be the client, which, in this context, is the WMSU TEC. They will be able to view and manipulate the data, such as the schedules and room availability. The secondary users will be the students. Since they are only registering for the CET, they will most likely only use the application once. They can only submit their registration form and can view their data and their registration status. Both of the users will communicate with the webserver using web browsers.

Web Tier

In the Web Tier, it is here that the server will allow communication between the users' web browser and the Application proper that is within the Business Logic Tier.

Business Logic Tier

The business logic tier comprises of two main components. The first is the application that will handle the commands, logical decisions, and the data management, by the client and the database, The second is the application that will handle student

data and put it through the data mining algorithms set in place in order to generate new visualizations for use by the WMSU TEC.

Data Tier

All relevant data will be stored in the Data Tier, as it will hold the database server that will communicate with the application.

CHAPTER IV
RESULTS AND DISCUSSION

Upon receiving and then encoding the data, the researchers conducted an exploratory data analysis on the dataset. It was during this time that the researchers discovered the first limitation, in that the data only included the academic background of the student, as the WMSU Admissions Department does not require the student to submit their CET forms upon enrolment; therefore, the data that the study had access to was the students' 2nd-year senior high school grades, schools, age, academic track, and their respective CET scores which included the raw scores for all five modules that make up a CET score, and their percentile rankings.

The dataset consists of 650 instances and 132 features. The researchers found 650 samples satisfactory. It was determined using Slovin's formula to calculate the required amount given the desired confidence level of 95%. According to the WMSU TEC, the target population is comprised of the successful applicants who enrolled. This amounts to 3,456 for 2018, 3,707 for 2019, and 6,188 for 2020, for a total of 13, 351. Using Slovin's formula, it was calculated that 388 samples were needed for the study. Of the 132 features, 22 are not related to student grades. Of the 22, 16 are related to the student's CET results, including their raw scores and percentile ranks for the entrance exam modules. Some features are not necessary for the study, such as Student Names, and thus they will be dropped later on.

During the exploratory data analysis, the researchers discovered multiple exciting findings regarding the data that would influence the strategy. The first finding was that the data is highly fragmented and varied, with 77.0% of features missing. It is because students enroll from many schools and strands, and each school has a different curriculum lined up for each strand. Each strand has a standard curriculum that the Philippine Department of Education (DepEd) has to follow. Despite this, we have found that each school implements the curriculum differently. Often, this can be in the form of shifting the schedules or adding subjects unique to the school or very uncommon in other schools, such as Zamboanga Chong Hua High School's "Chinese Language"

subject or Pilar College's "Church and Missionary Discipleship" subject. It presents a problem, as the researchers only have access to the second-semester grades of the students' second senior high school year. The fragmentation can be visualized in the non-null matrix below.

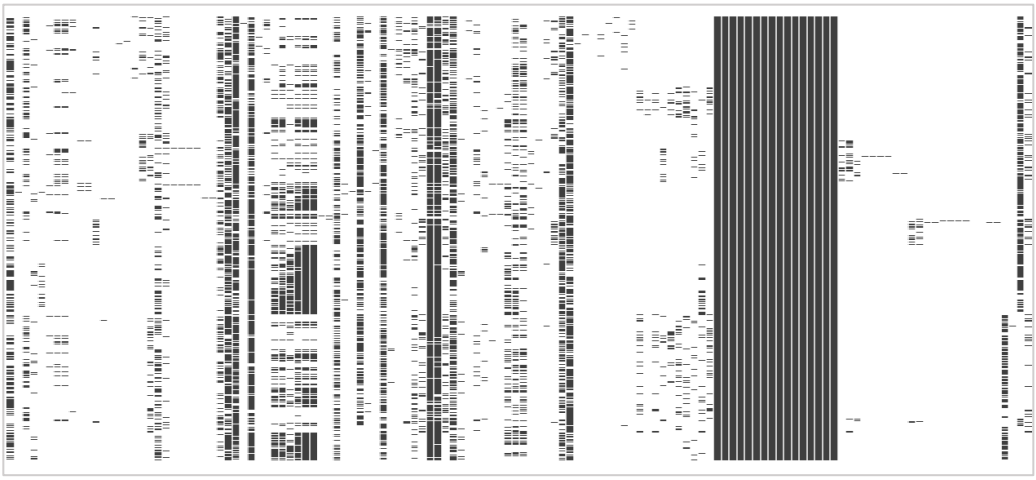


Figure 12.Non-Null Matrix (Non-Nulls represented as black dashes)

It is also apparent that some subjects are less fragmented than others and in order to visualize, the researchers took the number of non-null values per subject. The following figure displays only subjects with the number of non-null values greater than 200.

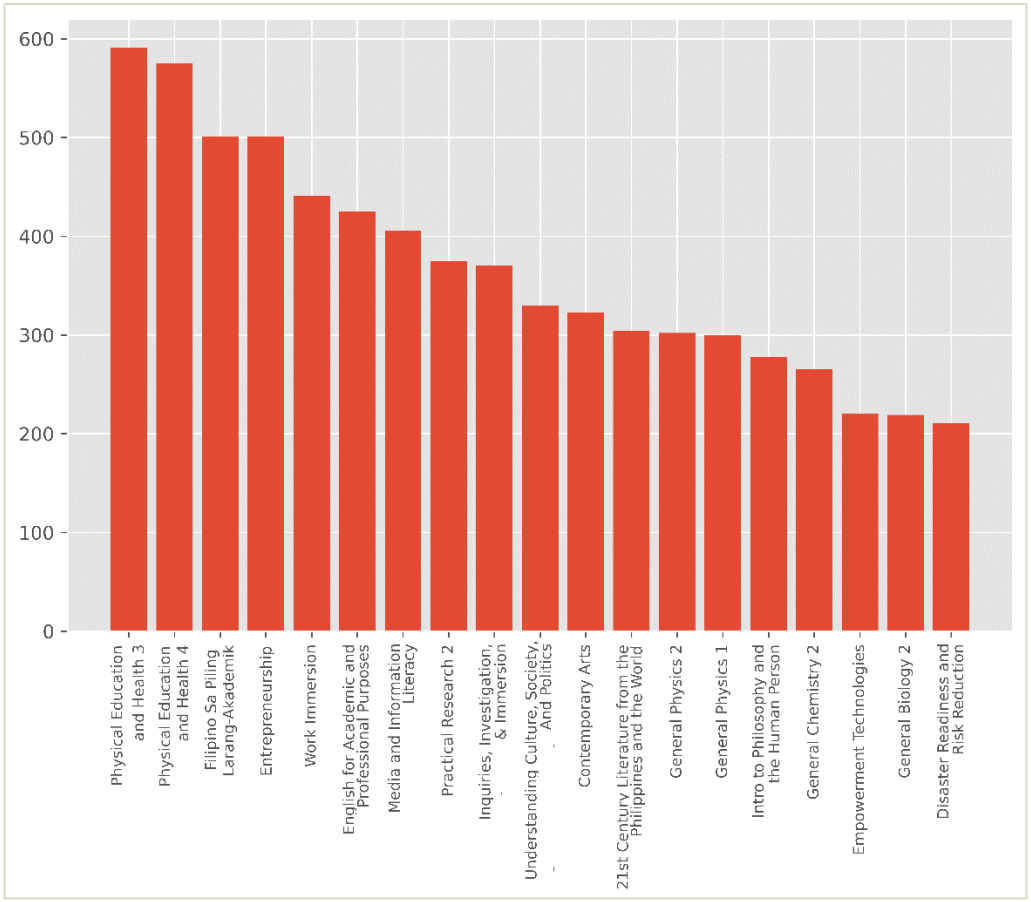


Figure 13.Number of Non-Null Values

The next step was to calculate the coefficients for all subjects per module, as well as for the total score of all the modules. It was calculated using both Pearson and Spearman coefficient techniques so that both linear and non-linear relationships can be accurately measured. This can be visualized into a heatmap.

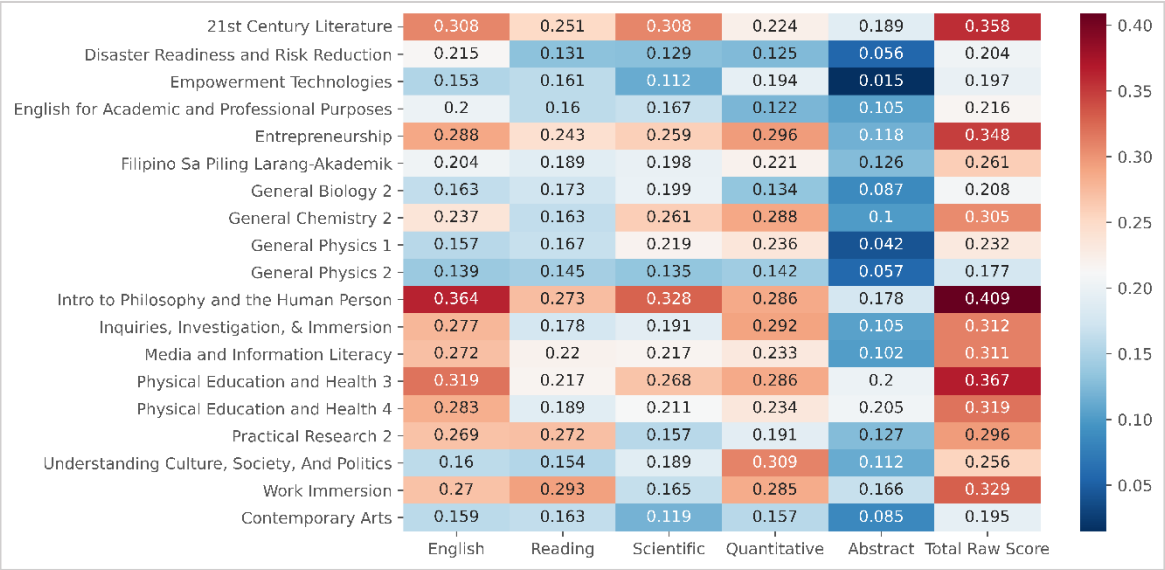


Figure 14.Pearson Correlation Heatmap

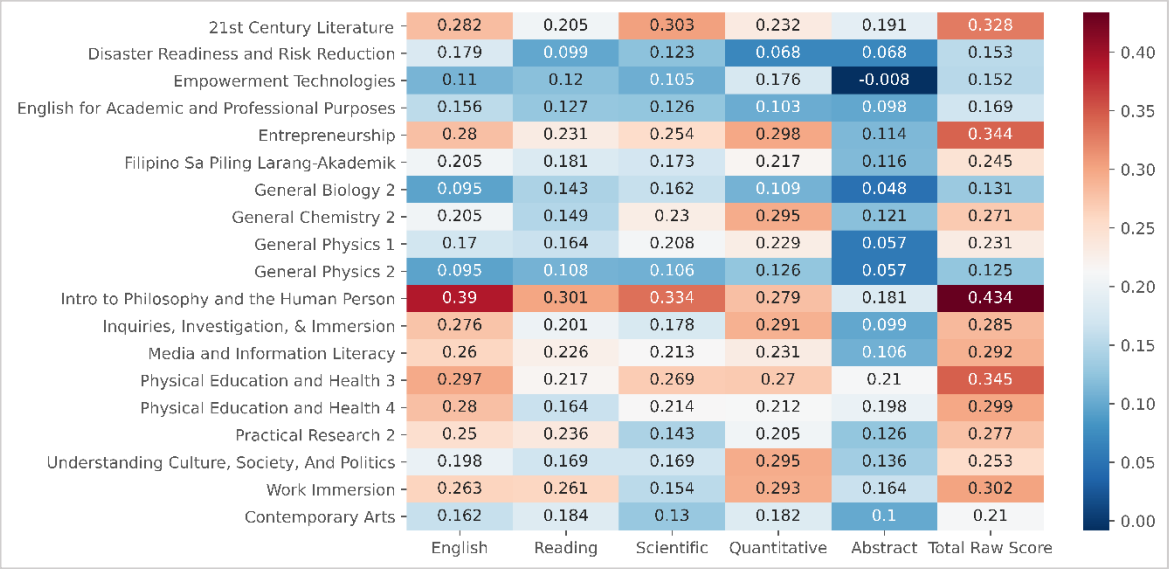


Figure 15 Spearman Correlation Heatmap

As is apparent in the heatmap, different subjects have different levels of correlation with each score. The researchers focused on the subjects with correlations between ± 0.30 and ± 0.49 using either Pearson's or Spearman's coefficients as variables with correlations in this range are considered to have a moderate relationship with the target variable. [16] However, the researchers decided to reduce the lower threshold to ± 0.29, as this would give more features with a minor 0.01 difference in the coefficient. Despite this, the "Abstract" module of the exam still does not have any strong

correlation with any of the listed subjects, thus would not be part of the first strategy for the model creation process. The subjects correlated with each score are listed below.

- English
    - 21st Century Literature
    - Intro to Philosophy and the Human Person
    - Physical Education and Health 3
- Reading
    - Intro to Philosophy and the Human Person
    - Work Immersion
- Scientific
    - 21st Century Literature
    - Intro to Philosophy and the Human Person
- Quantitative
    - Entrepreneurship
    - Inquiries, Investigation, & Immersion
    - Understanding Culture, Society, and Politics
    - General Chemistry 2
    - Work Immersion
- Total Raw Scores
    - 21st Century Literature
    - Entrepreneurship
    - General Chemistry 2
    - Intro to Philosophy and the Human Person
    - Inquiries, Investigation & Immersion
    - Media and Information Literacy
    - Physical Education and Health 3
    - Physical Education and Health 4
    - Practical Research 2
    - Work Immersion

The training and testing samples used in the study are the result of train and test splitting method, with a 70:30 ratio split for training and testing samples respectively.

This means that 70% of the data will be used for training the model while 30% will be reserved to serve as unseen samples for prediction.

Initial Strategy Regression

The initial strategy involved the use of regression algorithms to predict the score of the student based on their subject grades. It quickly became apparent that the highly fragmented nature of the data would make it incredibly difficult for the creation of a model relying on multiple subjects as features. As an example, when attempting to develop a model for the English module of the CET, the first two subjects in the list which are "21st Century Literature" and "Intro to Philosophy to the Human Person" would have 304 and 278 entries respectively when they are separated. When grouped together, however, the number would drop down to 101, which represents the number of students that have both subjects. This presents a problem since the models the researchers have lined up cannot take blank values as features. The problem is exacerbated the more features are included. The large discrepancy of null values between the subjects also makes it difficult to use imputation as a statistical technique to fill in these blanks. Using the earlier example, imputing around 285 entries for either "21st Century Literature" or "Intro to Philosophy of the Human Person" will severely affect the quality of the model no matter which method is used. If the mean, median, or mode methods are used, it would shift the distribution of the scores, as more than 66% of the entries were imputed. At the same time, this would also border on doctoring data, as there would be very little chance that a student of the Humanities and Social Sciences (HUMSS) academic track would have subjects such as "General Chemistry 2" or "Physics 2," and in the same vein, Science, Technology, Engineering and Mathematics (STEM) students would not have a grade in "Disciplines in Social Sciences".

The remaining feasible option for this strategy was to create a model using only a single feature, and a single target label. Using the correlation coefficients as part of the initial step of feature selection, the researchers were left with a reduced amount of features to test the model with. The testing began with the feature with the highest correlation coefficient with English, the first of the exam modules. Intro to Philosophy and the Human Person has a 0.39 Pearson coefficient, and there was an attempt to visualize its relationship with English Proficiency.

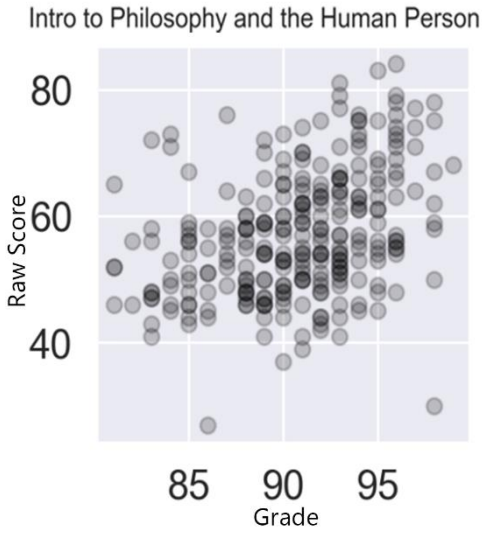


Figure 17.Scatter Plot (Intro to Philosophy vs English)

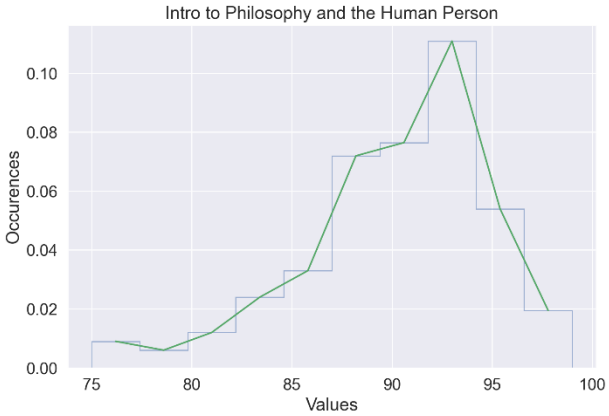


Figure 16 Distribution (Intro to Philosophy)

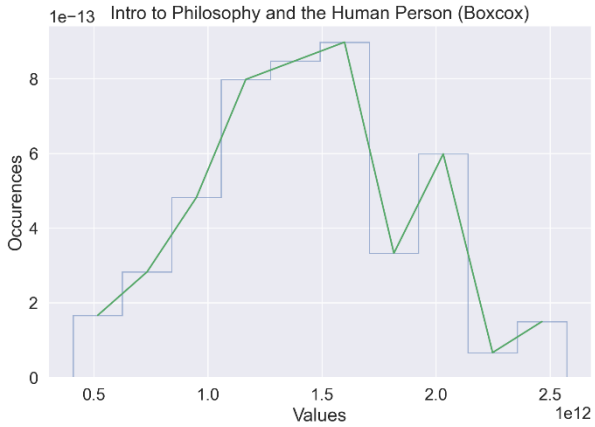


Figure 18. Transformed Intro to Philosophy Distribtution (Boxcox)

The second complicating issue with the initial strategy encountered is that the variance of the data is extreme, given the sample size that the study has, with this example having a variance of 16.04. Removing outliers has had little impact, and would only be effective should a significant amount of the data set be filtered out. Despite this, as seen in *Figure 14*, there is a clear upward trend between the score and the grade, with the possibility of a line being reasonably drawn from the lower left towards the up-right corner diagonally. This signifies that a relationship exists between the score and the grade. It was found that the distribution for the grades is moderately negatively skewed. This distribution is the trend for all the subject-related features. This will present a problem when using algorithms that assume normal distributions. That is why when creating the models, it was also tested with the same dataset that was transformed using Boxcox transformation. (*See figure 16*) The researchers then proceeded to test the features individually towards predicting the score label with the following algorithms:

Regression Algorithms

- Linear Algorithms
  - Linear Regression
  - Lasso Regression
  - Ridge Regression

- Non-Linear Algorithms
  - Decision Tree Regressor
  - Random Forest Regressor
  - Support Vector Regression (SVM)
    - RBF
    - Linear

Initial Strategy Results

The results of the initial strategy were inconclusive, at the least. The high variance of the data takes its toll on the methods of evaluation that were used to determine the quality of the data. The means of evaluation the study uses are the standard methods used by statistical studies in regression.

For a regression model to be considered reasonable:

- That the coefficient of determination or R-Square (R2) score of the model is close to 1.0, with models that have a score of 1.0 being able to predict 100% perfectly.

- The Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) are low compared to the maximum values of the data set. The MSE measures the deviation of the predictions from the actual inputs. At the same time, RMSE is the square root of MSE. It is a more reliable metric when comparing the performance of multiple models to each other.

- The Mean Absolute Error (MAE) is a low value, the lower, the better. The researchers used K-fold cross-validation to determine the MAE, with the constant value k = 10.

- The residuals of the prediction are uniformly distributed, with little to no increase or decrease invariance. It can be visually confirmed using a 'prediction residual vs. predictor variable' scatter plot.

- The distribution of the residuals is normalized and can be visually confirmed using a histogram and Q-Q plots of the distribution of the residuals.

Following these metrics, the quality of the developed models is poor, despite extensive testing and data preprocessing.

The researchers tested each feature individually to a specific label multiple times, using a different preprocessing technique each time and using an ensemble of techniques during certain testing phases. The first technique is to filter out outliers. Outliers were tested using box-plot graphs, and these outliers were removed using the IQR. The second technique was the normalization of features and labels to a '0 to -1 range'. The third technique is using the Box-Cox transformation for features to attempt to create a more normalized distribution of the feature values. The impact on the model's performance varies per testing phase, with varying, albeit minuscule, degrees of improvement or deterioration.

The single feature was exponentially increased from a range of 2 to a variable n, and new columns for each exponent were added to the dataset. All of the tests that were conducted using n = 16 create a less linear model that accounts for variation based on the alpha value of either the Ridge or Lasso algorithm.

The results for all the subjects in each module follow a specific trend. The quality of the models is difficult to ascertain due to the low number of entries per feature and high variance (See Appendix E). After extensive testing, the scores represented below are the highest scores taken from the highest scoring model.

Module 1 - English Proficiency					
Subject	**Model with best fit**	**R^2 Score**	**MSE**	**RMSE**	**MAE (Standard Deviation)**
21st Century Literature	Ridge Regression (a=1e-08)	0.055	120.048	10.956	8.338 (1.564)
Intro to Philosophy and the Human Person	Decision Tree Regressor	0.21	84.841	9.21	7.758 (0.721)
Physical Education and Health 3	Ridge Regression (a=1e-15)	0.158	83.183	9.12	Accuracy: 7.740 (0.562)
Module 2 – Reading Comprehension					
Intro to Philosophy and the Human Person	Linear Model	0.048	13.980	3.739	3.009 (0.449)
Work Immersion	Random Forest Regressor	0.167	14.052	3.748	2.930 (0.404)
Module 3 – Scientific Skills					
21st Century Literature	Ridge Regression (a=20)	0.068	18.636	4.317	4.009 (0.643)
Intro to Philosophy and the Human Person	Ridge Regression(a=0.0001)	0.116	19.081	4.368	3.316 (0.399)
Module 4 – Quantitative Skills					
Entrepreneurship	Lasso Regression (a=0.0001)	0.095	26.872	5.183	3.767 (0.483)
Inquiries, Investigation, & Immersion	Random Forest Regression	0.044	31.702	5.630	3.955 (0.532)
Understanding Culture, Society, and Politics	Linear Regression	0.096	25.317	5.031	3.869 (0.439)
General Chemistry 2	Linear Regression	0.101	29.994	5.476	4.067 (0.618)
Work Immersion	Ridge Regression (a=0.0001)	0.121	24.678	4.967	3.658 (0.371)
Total Module Score					
21st Century Literature	Decision Tree Regression	0.129	339.273	18.419	15.997 (2.590)
Entrepreneurship	Linear Regression	0.115	327.338	18.092	14.626 (1.979)
Intro to Philosophy and the Human Person	Random Forest Regressor	0.209	362.571	19.041	14.532 (1.965)

Physical Education and Health 3	Ridge Regression (a=1e-15)	0.19	322.268	17.951	15.205 (1.103)

Table 3.Model Scores

After visualizing the errors of the highest-scoring models, the researchers reviewed the results. They decided to suspend further development with the initial strategy. Because, as is apparent in the table, despite moderate correlations between the features and their respective scores, the correlation coefficients or R2 scores are consistently low, with the highest score for the entire table being 0.21 for the feature 'Intro to Philosophy and the Human Person' for both the English module and for the Total Module Score indicating a 21% prediction accuracy. The low R2 score does not necessarily mean that the model is not accurately predicting the model. It can explain using other means of evaluation, where it would see that the model follows the data trend. For this example, the error residual distribution for the Random Forest Regression model is represented below, with Intro to Philosophy and the Human Person as the feature and Total Raw Score as the label.

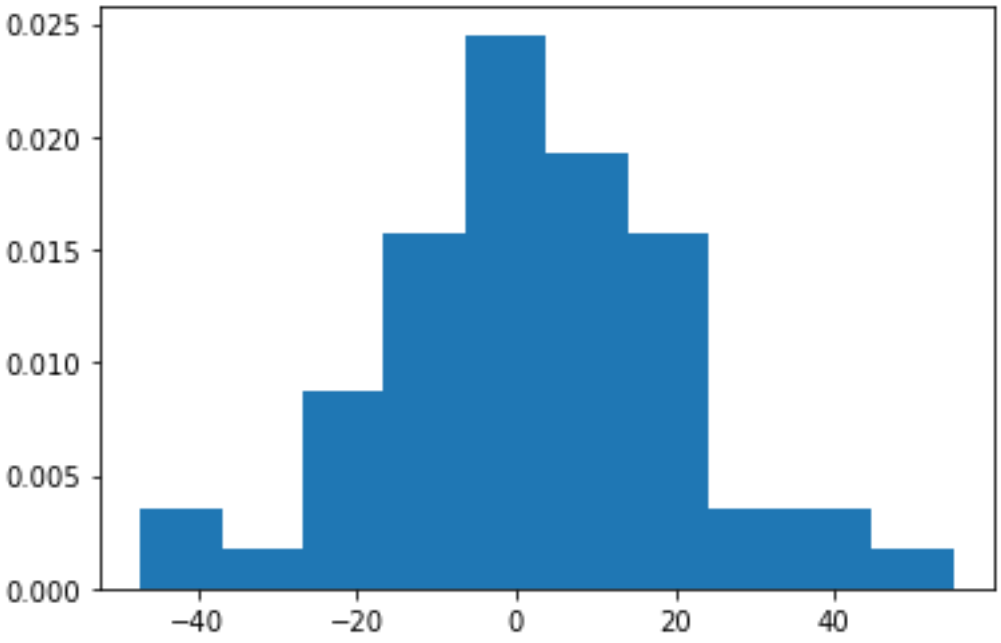


Figure 19 Random Forest Regression Residual Distribution Histogram

The error residual of the predictions follows a normal gaussian distribution. This signifies that th05e bias of the model is minimal, and that it follows the actual trend of the data.

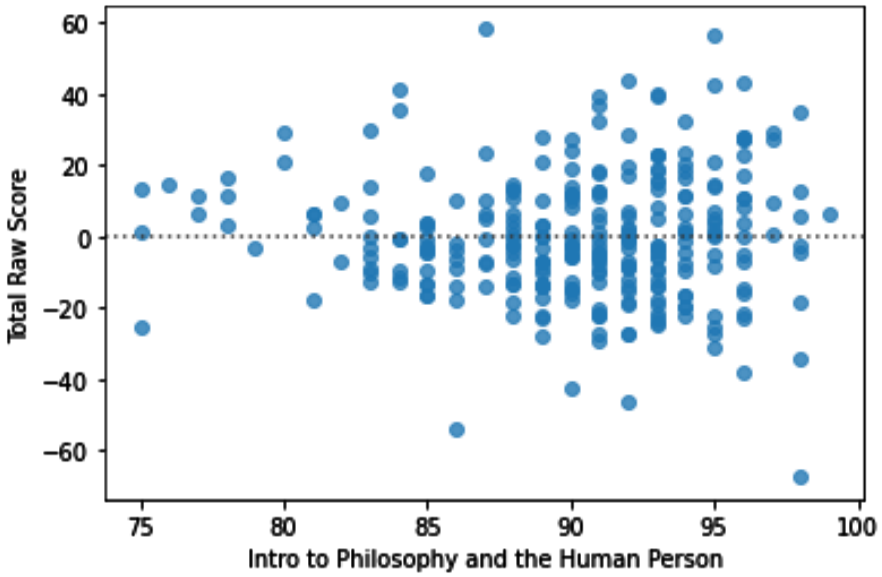


Figure 20 Residual vs Fit Scatterplot

The residual scatterplot show that the errors follow a uniform distribution around the 0 value of the z-axis, and does not form specific clusters, therefore it reinforces the hypothesis that the model follows the actual trend of the data, though the variance of the residuals does not make this apparent at first glance. This fact is more pronounced in the following figure.

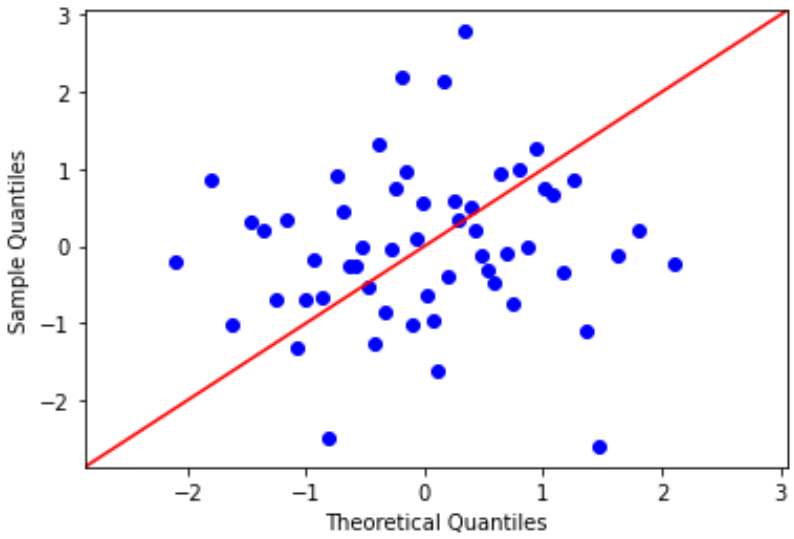


Figure 21 Residual Q-Q Plot

The Q-Q plot exhibits more clearly the variance for the dataset when using single features to a single label. In order for a model to be considered good, the blue plots should follow the red diagonal theoretical/sample quantile line without skewing upwards or downwards. The residuals meet this condition, as it clearly follows a uniform distribution along the red line. The problem comes from the variance of the residuals, due to the dataset being trained on a highly varied data set with a low amount of entries.

Due to the high variance, the researchers decided that there was a need for another course of action in order to create a model to determine a student's performance for the WMSU CET.

Subsequent Strategy Classification

With the counsel of this study's adviser, the researchers instead opted to shift to another strategy. This strategy involves using feature construction as a form of feature engineering. The features constructed were based on the academic data of the student. Instead of using each subject as an individual feature, they were grouped into three main subgroups, English, Math, and Science. These subgroupings were used because the WMSU CET focuses mainly on these three subjects. Modules 1 and 2, or English proficiency and reading comprehension, focus on English. Module 3, or Scientific Skills, focuses on science, while Module 4, or Quantitative Skills, deals with math subjects. The information about the subjects deemed related to the subgroup was taken directly from the DepEd website. [17] Non-related subjects that do not belong to any subgroups, such as "Filipino sa Piling Larang," were dropped from the study. Listed are the following subgroups and their respective subjects used in the study:

ENGLISH

- Community Engagement, Solidarity, and Citizenship
- Creative Non-Fiction
- Creative Writing
- English for Academic and Professional Purposes
- Empowerment Technologies
- Organization Management
- Intro to Philosophy and the Human Person
- Trends, Networks and Critical Thinking in the 21st
- Oral Communication
- Understanding Culture, Society, And Politics
- Introduction to World Religions and Belief System
- Trends, Networks, and Critical Thinking in the 21st Century
- Media and Information Literacy
- Reading and Writing
- Personal Development
- Philippine Politics and Governance
- 21st Century Literature from the Philippines and the World

MATH

- Advanced Statistics
- Fundamentals of Accounting / Business Management 2
- Applied Economics
- Entrepreneurship
- Business Enterprise Simulation
- Business Finance
- Pre-Calculus
- Business Mathematics
- Statistics and Probability
- Basic Calculus

SCIENCE

- Disaster Readiness and Risk Reduction
- General Biology 1
- Earth & Life Science
- General Biology 2
- General Chemistry 1
- General Physics 1
- General Physics 2
- Physical Education and Health 4
- Physical Education and Health 3
- Physical Science
- General Chemistry 2

The grades for each subject in a subgroup were added together and divided by the number of subjects the student has, effectively getting the average of all subjects that apply to that particular student. The academic features include the average grades and number of subjects for each subgrouping, the sum of the number of subjects for each subgrouping, and the mean of all three subgroups' average grades. For clarity, average English grade, English No. of Subjects, Math Average Grade, Math No. of Subjects, Science Average Grade, Science No. of Subjects, EMS Average Grade, and EMS No. of Subjects.

In addition, using the student's information regarding which school they graduated senior high school from, whether the school was public or private, a high school, university, or college, and whether the school was located in either district 1 or 2 of Zamboanga City. The researchers then encoded all categorical information into a numeric coded format. The categorical data includes the school, category of school, type of school, district of school, senior high school strand of the student, and the track of the student.

The researchers also created a new type of predicted label called "Sum of Raw Scores – Categorical." This label is based on the total score of the student for the CET and generalized into three main categories: Below Average, Average, and Above Average. The researchers calculated the average total score for all three years. It yielded an overall average of 135. It should be noted that each year has different averages when calculated on its own. The total raw scores for the school year 2018-2019 yield an average of 131, 2019-2020 yield 136, and 2020-2021 yield 138. The researchers chose to use the average for all three years to create a more robust and singular model instead of having to create a separate model for each year for the study and each year after the implementation of the system. The Sum of Raw Scores – Categorical was done by setting a range for the average. It was decided to use 7 points above or below the average; thus, the "average" score spanned 128-to 142. Analysis showed that the distribution of samples with this range of scores adds up to approximately 20.46% of the population. Ranges of ±10, ±5, and no range/average were also tested.

The researchers opted to test multiple algorithms with the constructed and preprocessed features and chose the model with the most consistent model performance. The algorithm's input will only include the constructed and preprocessed features, not the individual subjects. The target variable is the total raw score. The testing will check whether the raw numerical score or its categorical form will perform better for the model. The following algorithms were used in the testing:

Classification Algorithms

- Linear Algorithms
  - Logistic Regression (For Classification)

- Non-Linear Algorithms
  - Neural Network Regression/Classification
  - Gradient Boosting Regression/Classification
  - Random Forest Regression/Classification
  - Support Vector Regression (SVM Linear)
  - Naives Bayes
    - Gaussian
    - Multinomial
    - Complement
  - K Nearest Neighbors Regressor/Classifier

Second Strategy Results

The results of the second strategy yielded better results than the initial strategy. The researchers first implemented feature selection by conducting testing for all the models for both types of labels and using backward feature selection. It was determined that the models work best when the 'school' feature is dropped. Trying to drop any other feature lowers the model scores. This could be due to the high variance of the 'school' feature combined with the small dataset. For the first phase of testing, the researchers tested for which label type, categorical or numerical, is better predicted for the model. Predictions for categorical labels were developed with classification models, while for regression models were used for numerical labels. It is apparent however, that both tests will not be applicable to all algorithms as some algorithms can only predict categorical values, and some algorithms can only predict numerical values. To specify, Linear Regression will be unique to the numerical phase of the test, and Naïve Bayes and Logistic Regression will be unique to categorical phase of the test. It was found that there is a large improvement to model performance when trying to predict the categorical labels instead.

Numerical Phase Data Analysis

The numerical phase analysis involved trying to ascertain the Pearson and Spearman correlations between the features and the target variables.

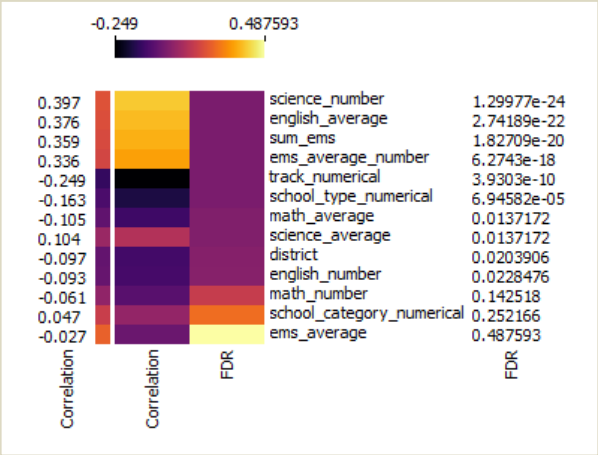


Figure 22. Correlation Heatmap to sum of raw scores (Pearson)

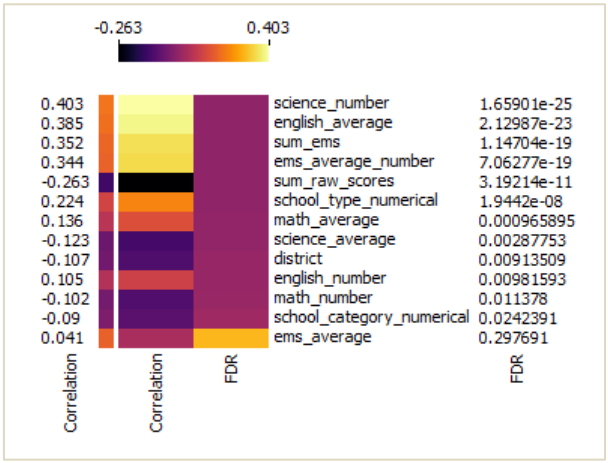


Figure 23.Correlation Heatmap to sum of raw scores (Spearman)

The treatment of outliers was tested with three techniques, which are including the outliers, a single box-plot technique, and the double box-plot technique. The single box-plot method removes the outliers of the feature with the highest correlation, and leaves 426 out of 645 samples. The double box-plot filtered the outliers of the first two features with the highest correlations scores, and left the study with 243 samples. The inclusion of all outliers produced the models with the highest model score, which is

Gradient Boosting Regression with a score of 0.372. This could be attributed to the increased number of samples being used, using the entire dataset without filtration.

Categorical Phase Data Analysis

For categorical data, the same testing methods were used. Including all the data without filtering outliers yielded an average model score of 0.523 for all models. Box plots were used to remove the outliers of the most relevant features to the label. The relevance of the features was determined using x2 or ANOVA analysis of the features over to the subgroups of the labels. According to this metric, the most relevant features were the number of science subjects and the number of all three subject groups combined (EMS No. of Subjects). Using the single box-plot technique leaves 449 samples. It raises the model score slightly, with the highest scoring model being Gradient Boosting with 0.709. The most effective technique was to use the double box-plot filtering, leaving the model with 307 samples but raising the model score to a range between 0.836 to 0.854 for different models. It means that the system can accurately predict with at least 83% confidence whether the student will receive an Average, Above average, or Below average score for their CET. During this phase, the different ranges for the “Sum of Raw Scores – Categorical” were tested to optimize the model. It should be noted that a range of zero, where the three categories are reduced to two, to just above average and Below Average, yields the highest result, with the Random Forest model producing Accuracy and Recall scores of 0.95 and an AUC of 0.991. (See Appendix E) It means that the classification models can predict whether the student will be above average or below Average based on their data with an accuracy of 95%. Testing for the other ranges resulted in the Average category ranging ±7 to perform the best by returning the highest precision scores.

After receiving the results with the optimized models, the results were documented in the following table, which represents the model scores for both categorical and numerical label testing phases.

Numerical Phase				
Model	**R^2 Score**	**MSE**	**RMSE**	**MAE**
Lasso Regression	0.256	295.951	17.203	13.427
Gradient Boosting	0.320	270.433	16.445	12.637
Random Forest	0.317	271.799	16.486	12.523
SVM	0.2	318.326	17.842	13.691
kNN	0.152	337.132	18.361	14.02
Neural Network	-3.940	1964.805	44.326	37.566

Table 4 Numerical Testing Phase Model Scores

Categorical Phase				
Model	**Precision**	**Classification Accuracy / Recall**	**Area Under Cover**	**F1 Score**
Gradient Boosting	0.787	0.803	0.925	0.794
kNN	0.845	0.769	0.845	0.742
Logistic Regression	0.805	0.828	0.930	0.813
Naïve Bayes	0.799	0.756	0.923	0.773
Neural Network	0.775	0.807	0.921	0.786
Random Forest	0.790	0.836	0.927	0.799
SVM	0.742	0.803	0.913	0.761

Table 5 Categorical Testing Phase Model Scores

The R^2 Model scores indicate that the classification models performed significantly better than the regression models. Therefore, the models can predict the category of performance for the student, whether they would score average, above

average, or below average, with an 80% degree of accuracy. The researchers decided to use the categorical labels in the prediction of student performance in further implementations and development of the model. For future reference, the following table lists the parameters that were determined to achieve the highest model score for each algorithm.

Model	**Parameters**
Lasso Regression	Alpha = 0.5 Fit Intercept = True
Gradient Boosting	No. of Trees = 100 Learning Rate = 0.100 Tree Depth Limit = 3 Minimum Subset Size = 2
Random Forest	No. of Trees = 100 Minimum Subset Size = 5
Neural Network	Neurons in Hidden Layers = 100, Activation = Adam Solver = ReLu Alpha = 0.0001 Max No. of Iterations = 200
Logistic Regression	Type = Ridge Strength (C) = 1000
SVM	Type = SVM Kernel = Linear Numerical Tolerance = 0.0010
kNN	No. of Neighbors = 5 Metric = Euclidean Weight = Uniform

Table 6 Model Parameters

Classification Implementation and Model Evaluation

All the classification algorithms were implemented in a Python environment in order to conduct further testing and model evaluation. The evaluation metrics consist of the accuracy score, recall score, confusion matrix heatmap, and the ROC Curve chart. The accuracy score is taken using k-fold cross validation, with k = 10. The recall score uses scikit-learn's *recall_score()* function with the *average* set to *'macro'* due to the model being a multiclass problem. Only the training set of the split data was used in determining the scores.

The researchers chose to evaluate the models with consistent accuracy scores of above 0.80. The models that consistently return the scores, using the metrics above, that indicate its high quality is the SVM model with the highest scores, followed by Logistic Regression, followed by Random Forest, followed by Gradient Boosting. During testing using the scikit-learn's built-in testing function, the Random Forest model often has varying accuracy scores, but averages with an accuracy score of around 0.85. Under certain conditions, it returns a 0.90 model score. For this reason, K-Fold cross validation was used for the determination of the accuracy score, recall score, precision score, and the ROC/AUC Score. The values are the mean of the scores through the K-number of iterations of cross validation, while the values in parentheses are the standard deviation between the results.

Model	**Accuracy Score**	**Recall**	**Precision**	**ROC/AUC Score**	**F1 Score**
SVM	0.877 (0.05)	0.747 (0.1)	0.774 (0.167)	0.943 (0.04)	0.742 (0.124)
Logistic Regression	0.829 (0.043)	0.715 (0.064)	0.730 (0.104)	0.922 (0.027)	0.708 (0.073)
Gradient Boosting	0.841 (0.056)	0.715 (0.089)	0.735 (0.129)	0.893 (0.056)	0.714 (0.1)
Random Forest	0.853 (0.069)	0.727 (0.115)	0.739 (0.159)	0.914 (0.058)	0.723 (0.130)
Neural Network	0.8 (0.033)	0.631 (0.038)	0.552 (0.075)	0.874 (0.046)	0.585 (0.046)

Naïve Bayes (Complement)	0.825 (0.026)	0.642 (0.023)	0.552 (0.017)	0.89 (0.066)	0.591 (0.02)

Table 7 Implemented Classification Model Scores

Due to the problem of uneven class distribution, the researchers focused on the more robust F1 score, which is the weighted mean between recall and precision. The mean of false positive and false negative metrics will better quantify the quality of the model.

The SVM model consistently outputs the highest accuracy score and recall score, thus it was used as a basis for the evaluation phase of testing. It is known that the model can predict the category of the student with an 87% accuracy rate though it has an F1 score of 0.742. This means that while 87% of the total dataset was correctly predicted, the model incorrectly predicts or assigns students of a certain class to a certain target class or category due to the low frequency of their true category, a result of class imbalance. In order to confirm this, it can be visualized into a heatmap confusion matrix to see which categories were correctly predicted, and which categories were incorrect.

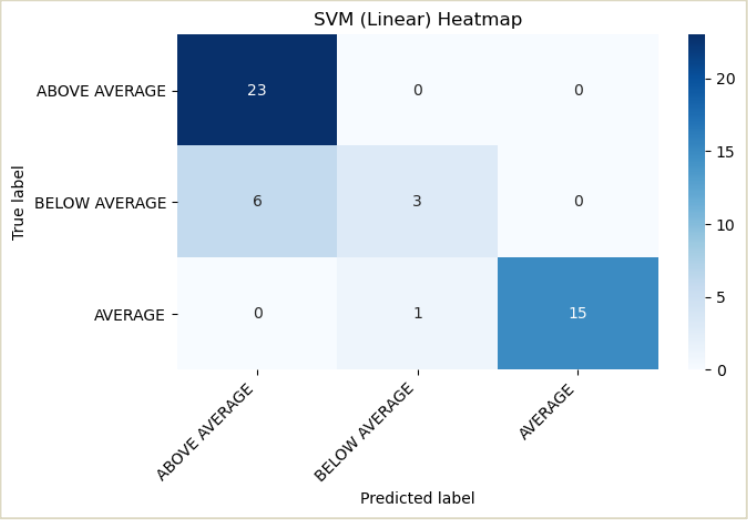


Figure 24 SVM Confusion Matrix

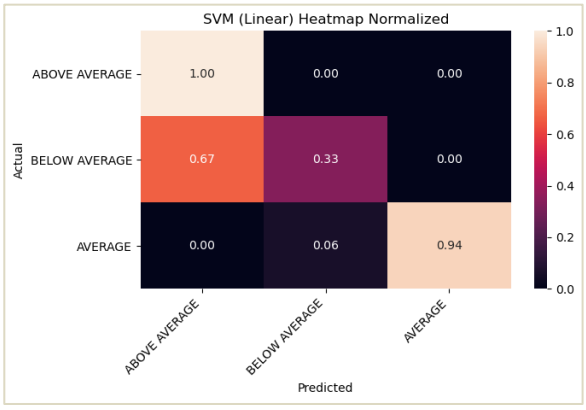


Figure 25. SVM Normalized Confusion Matrix

The heatmap (*figure 22*) shows the number of correctly predicted variables diagonally from the top left corner to the lower right. The normalized heatmap (*figure 23*) shows the normalized values of the first heatmap, which are equivalent to a range between 0.00 to 1.00. Values with 1.00 indicate that 100% of the actual values were predicted by the model into that category. It should be noted that all models have a tendency to produce a low accuracy percentage for the 'AVERAGE' category (*See figures 24, 25, 26*). This is attributed to the class imbalance, as the number of instances for 'AVERAGE' are slightly lower than the Above or Below Average categories (*See figure 27*).

The performance of the model can vary when evaluating using the confusion matrix, which is why the researchers base the findings on the model F1 Score, as it outputs the average weighted mean of the recall and precision scores after a repeating train/test splitting process using the K-fold cross validation.

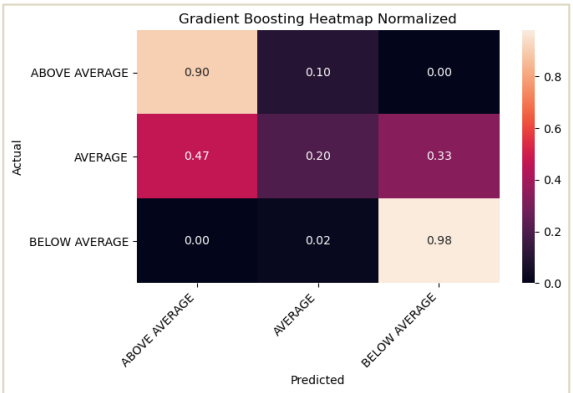


Figure 26. Gradient Boosting Normalized Confusion Matrix

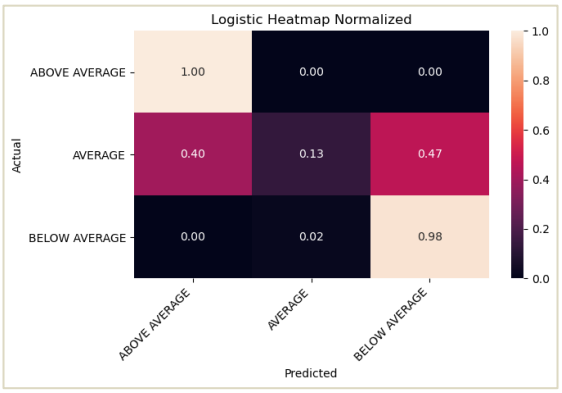


Figure 27. Logistic Regression Normalized Confusion Matrix

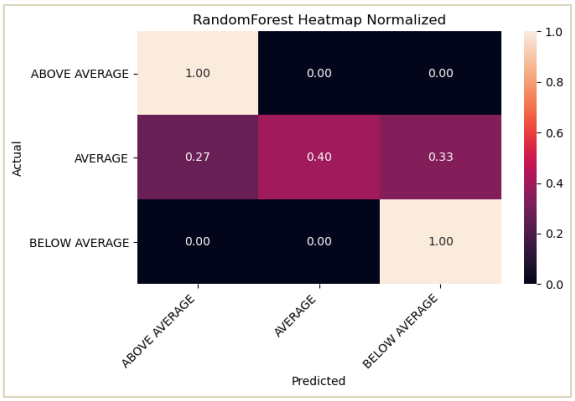


Figure 28 Random Forest Normalized Confusion Matrix

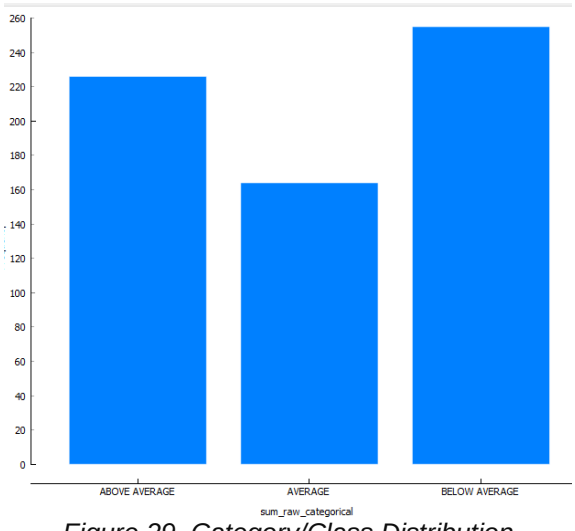


Figure 29. Category/Class Distribution

In order to further determine the actual quality of the models, the researchers visually evaluated each model using the AUC/ROC. For the ROC, the faster a line approaches 1, the higher quality the model is considered to be. The AUC or Area Under Curve score quantifies this metric, visible in the charts as the values next to the category in the chart's legend. First, the two highest scoring models, SVM and Logistic Regression, were compared.

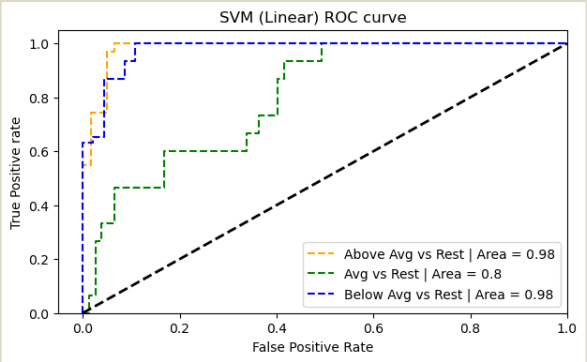


Figure 30 SVR ROC Curve

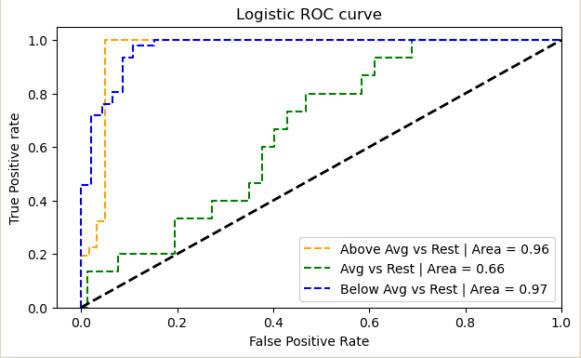


Figure 31.Logistic Regression ROC Curve

As mentioned above, the faster a line approaches 1, the better the model is considered to be. Not only can this be visually confirmed, it can also be quantified by finding the model with the higher AUC. In the figures above, the SVM model should perform more robustly when compared to Logistic Regression, despite having a very similar accuracy score of 0.86. The researchers determined the average AUC of all three categories per model to be 0.92 for SVC and 0.86 for the Logistic Curve.

The researchers also compared the SVM ROC Curve with the other models. All models performed worse when compared to the SVM model using this metric, with Gradient Boosting returning an average AUC of 0.90, Random Forest with 0.91, Neural Network with 0.8, and Naïve bayes with 0.83.

One-Hot-Encoding Approach

After trying to implement a one-hot-encoding approach for the categorical data, namely *school_category_numerical*, *school_type_numerical*, *district*, and *track_numerical*, a more accurate model was achieved when evaluating by the model's confusion matrix. As mentioned in *section 4.1.6*, the models have difficulty classifying score into the average range, as the percentage of actual vs. predicted instances for the "Average" Category was very low. An effect of this technique is that the overall model accuracy scores are lower, with a drop of at least 3% to 6% for some models. However, this method consistently returns a higher F1 Score, with the F1 Score for the SVM model returning a value of 0.791. This is an increase to 0.038 for model performance, since the class imbalance renders the F1 Score a more useful metric rather than the accuracy score.

Model	**Accuracy Score**	**Recall**	**Precision**	**ROC/AUC Score**	**F1 Score**
SVM	0.819 (0.043)	0.771 (0.060)	0.782 (0.072)	0.937 (0.019)	0.763 (0.067)
Logistic Regression	0.832 (0.035)	0.766 (0.040)	0.788 (0.060)	0.938 (0.031)	0.763 (0.046)
Gradient Boosting	0.816 (0.071)	0.764 (0.094)	0.785 (0.113)	0.927 (0.052)	0.759 (0.098)
Random Forest	0.819 (0.073)	0.751 (0.102)	0.764 (0.157)	0.932 (0.044)	0.740 (0.123)
Neural Network	0.775 (0.094)	0.682 (0.106)	0.68 (0.142)	0.896 (0.068)	0.655 (0.132)
Naïve Bayes (Complement)	0.774 (0.055)	0.712 (0.085)	0.706 (0.133)	0.881 (0.053)	0.692(0.105)

Table 8 One Hot Encoding Results

While the model score is lower, overall recall and precision scores are higher. The confusion matrix shows that the “Average” category has a higher tendency to be predicted correctly compared to non-one-hot-encoding techniques.

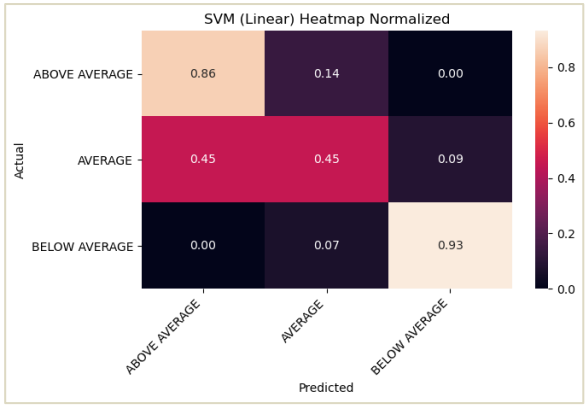


Figure 32. SVM Confusion Matrix (One-Hot-Encoded)

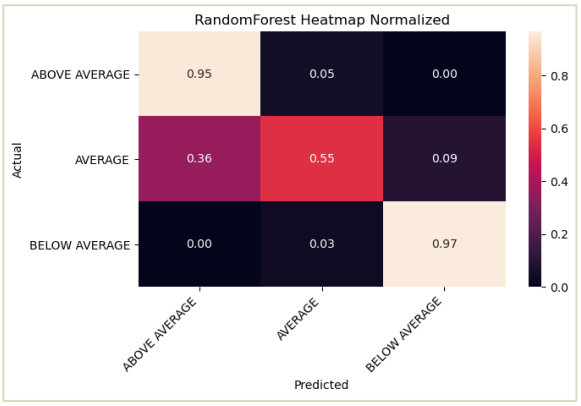


Figure 33.Random Forest Confusion Matrix (One-Hot-Encoded) Binary Classification

It was mentioned briefly before that turning the model into a binary classification problem vastly improves the model. The researchers measured the scores and found that the models are of a higher quality than the ternary model.

Model	**Accuracy Score**	**Recall**	**Precision**	**ROC/AUC Score**	**F1 Score**
SVM	0.915 (0.071)	0.905 (0.043)	0.916 (0.072)	0.956 (0.037)	0.914 (0.071)

Logistic Regression	0.906 (0.054)	0.903 (0.059)	0.91 (0.05)	0.958 (0.027)	0.903 (0.027)
Gradient Boosting	0.886 (0.053)	0.881 (0.051)	0.891 (0.056)	0.937 (0.041)	0.883 (0.053)
Random Forest	0.894 (0.065)	0.892 (0.063)	0.896 (0.068)	0.954 (0.037)	0.892 (0.065)
Neural Network	0.91 (0.04)	0.909 (0.045)	0.916 (0.031)	0.95 (0.035)	0.907 (0.043)
Naïve Bayes (Complement)	0.898 (0.045)	0.898 (0.047)	0.899 (0.046)	0.942 (0.037)	0.896 (0.046)

Table 9 Binary Classification Model Results

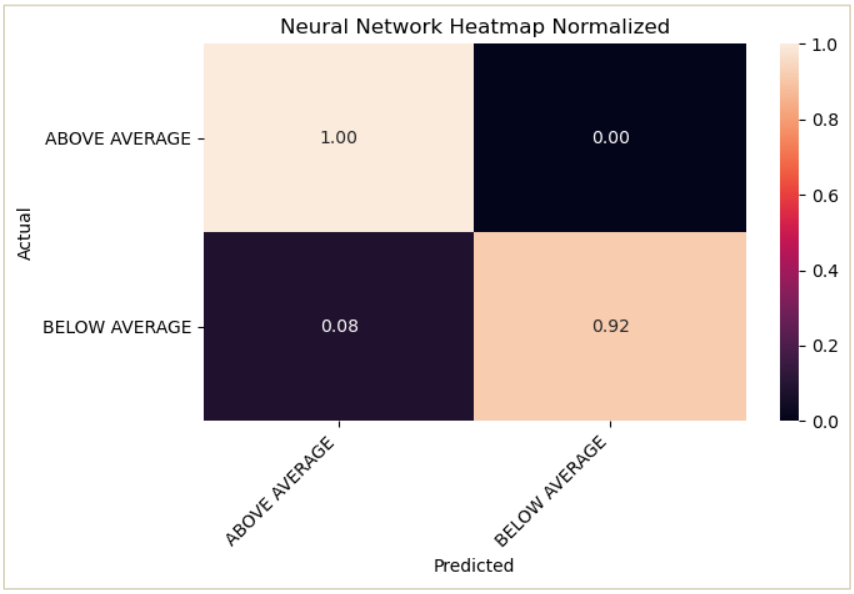


Figure 34 Neural Network Confusion Matrix

The confusion matrix also shows a higher level of performance compared to the ternary classification problem. The researchers attribute this to the fact that because of the binary classification, the data set is split into halves. Therefore, there are more instances for each class, giving the algorithms a greater opportunity for training. The ternary classification problem further splits the data set into three sets. Given a greater dataset, it is possible to develop models with the ternary classification with a similar quality.

The researchers opted against using the binary model for this study as its secondary purpose of is to create a prediction model for students to use in order to predict their performance with reasonably accurate results. The objective to determine whether or not SHS grades can determine the performance of the student can be done with a binary system, but the researchers opted for a multiclass system for the implementation of the system in order for students to get more information out of the prediction, as opposed to just telling them whether they would score higher than the average passing score or lower. The researchers decided to not exchange information density for accuracy.

Front-end Integration

The model was successfully deployed on to a front-end website, where in the student must input their grade data. The front-end accepts both demographic data and grade data in a user-friendly format. The SVM Classification model that produced a 90% accuracy score was used. The researchers tested the model on the web application with 14 unseen predictions, using data that was unused with both testing and training of the model. The model outputted satisfactory results, and the results generate the following confusion matrix.

		Predicted		
		Below Average	Average	Above Average
Actual	Below Average	4	1	1
	Average	2	2	0
	Above Average	2	1	1

Based on 14 students, the model can predict with a 50% true accuracy when introduced with the unseen predictions. It can be insinuated that the model tends to produce “Below Average” Predictions for a majority of the students, as 8 out of 14 students were classified into that category. The results of this testing are not indicative of the model’s true performance, as the sample size is extremely small when compared to the sample size used for training and testing the model. A more precise accuracy range and score can be achieved with the use and deployment of the website.

CHAPTER V
CONCLUSION AND RECOMMENDATIONS

Conclusion

The first conclusion of the study is that, despite the Confusion Matrix (*See figures 23 and 26*) indicating that Random Forest Model predicted the values more accurately, when based on the evaluation metrics, SVM is the superior model for implementation due to the following reasons:

- The SVM model produces the highest accuracy score when using 10-fold cross validation techniques.

- The SVM model has the second highest recall score. Random Forest Regression has the highest score by this metric, but does not meet the standards of the other metrics.

- The SVM Model has the highest AUC/ROC Score, implying that it is the most robust model among the rest.

This does not indicate however, that the Random Forest model is significantly inferior. The researchers consider the differences between Random Forest model and the SVM model to be negligible, and either could be used to develop the model for implementation with the prediction software.

The second conclusion is that students enrolled in the Academic Track perform better on the CET Examination. The results of analysis (*See figure 29)* showed that the number of science subjects is strongly correlated with the Total Raw Score in the CET.

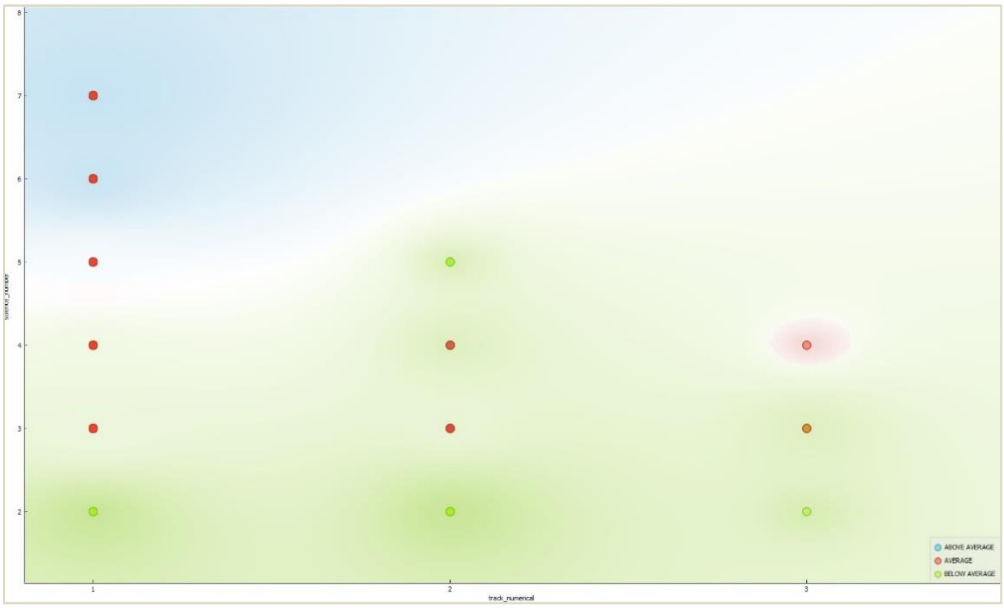


Figure 35 Track Type, Number of Science Subjects, Raw Score Category Scatter Plot

The figure shows that students who enrolled in the Academic Track (numerical value of 1) are more likely to achieve an “ABOVE AVERAGE” score compared to students of Technical Vocational track or other tracks (2 and 3 respectively) indicated by the color of the background.

The third conclusion is that there is a possibility that the demographic data of the student is a predictor for their performance on the CET. The researchers base this conclusion on the fact that included into consideration the type of their senior high school and whether it is part of a college, university, or high school, as part of the demographic data of the student, as private schools are more likely to admit students belonging to a family with a higher social standing than the families of students from public schools. The data under these features have a negligibly measurable positive correlation with the CET performance.

The fourth conclusion is that, with a satisfactory prediction accuracy rate of 86%, it is proven that a student’s past academic background and performance are strongly correlated with their CET scores, and are a strong predictor for CET performance.

Recommendations

The first recommendation of the study is for future researchers to include additional data, as this study's method of data collection was severely limited, impaired by the circumstances under which this study was taken. A more significant amount of data would amplify the accuracy of the statistical results given, with a strong possibility of improving the accuracy of the models and making it possible to develop the regression models for a more accurate presentation of the data. The acquisition of data can be both exhaustive and time-consuming. Thus, the first recommendation leads to the second;

The researchers recommend that the Online Student CET Registration System be used to test the model further and collect additional data for its improvement. This study's data collection method included asking for data from two separate organizations within the university, the WMSU Testing and Evaluation Center and the WMSU Admissions Department. Each organization collects its data independently. The union of data collection into a single platform that was developed would serve as a basis for the repository of data for future studies and improvements to the software.

Bibliography

[1] M. Alipio, "Predicting Academic Performance of College Freshmen in the Philippines using Psychological Variables and Expectancy-Value Beliefs to Outcomes-Based Education: A Path Analysis. Faculty of Radiologic Technology, Davao Doctors College," [Online]. Available: https://edarxiv.org/pra6z/. [Accessed 25 August 2021].

[2] A. E. &. Kent.C, "High School GPAs and ACT Scores as Predictors of College Completion : Examining Assumptions About Consistency Across High Schooks. Educational Reseracher. 49(3)," [Online]. Available: https://doi.org/10.3102/0013189X20902110. [Accessed 25 August 2021].

[3] A. E. &. D. D. Tatar, "Prediction of Academic Performance at Undergraduate Graduation: Course Grades or Grade Point Average? MDPI Applied Sciences, 10.," 2020. [Online]. Available: https://doi.org/10.3390/app10144967. [Accessed August 2021].

[4] S. J. H. L. O. H. T. H. A. Y. Q. H. J. C. H. O. H. &. L. A. J. Q. Yang, " Predicting Students' Academic Performance Using Multiple Linear Regression and Principal Component Analysis. Journal of Information Processing, 26, 170–176.," 2018. [Online]. [Accessed August 2021].

[5] S. &. F. N. Huang, "Regression Models For Predicting Student Academic Performance In An Engineering Dynamics Course. Utah State University," 2010. [Online].

[6] H. S. Y. S. F. &. S. H. Meng, " Application research of cluster analysis and association analysis. Institute of Electrical and Electronics Engineers. Published.," 2010. [Online]. [Accessed October 2021].

[7] S. &. N. K. Ramasamy, "Disease prediction in data mining using association rule mining and keyword based clustering algorithms. International Journal of Computers and Applications, 42(1), 1–8.," 2017. [Online]. Available: 89N3PDyZzakoH7W6n8ZrjGDDktjh8iWFG6eKRvi3kvpQ. [Accessed October 2021].

[8] "An Enhanced k-Means Clustering Algorithm for Pattern Discovery in Healthcare Data. International Journal of Distributed Sensor Networks, 11(6), 615740.," 2015. [Online]. Available: https://doi.org/10.1155/2015/615740. [Accessed September 2021].

[9] X. &. J. W. Yang, "Dynamic Online Course Recommendation Based on Course Network and User Network. Communications in Computer and Information Science, 180–196.," 2019. [Online]. Available: https://doi.org/10.1007/978-981-15-1301-5 15. [Accessed August 2021].

[10] A. L. K. &. Y. A. Khalid, "Recommender Systems for MOOCs: A Systematic Literature Survey (January 1, 2012—July 12, 2019). International Review of Research in Open and Distributed Learning, 21(4).," 2020. [Online]. [Accessed September 2021].

[11] "Machine Learning Decision Tree Classification Algorithm - Javatpoint. (2018). Www.Javatpoint.Com.," 2018. [Online]. Available: https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm. [Accessed October 2021].

[12] Q. A. A.-S. E. M. &. A.-N. M. I. Al-Radaideh, "Mining student data using decision trees. In the Proceedings of the 2006 International Arab Conference on Information Technology," 2006. [Online]. Available: https://www.acit2k.org. [Accessed 28 August 2021].

[13] R. Bevans, "An introduction to multiple linear regression [Online]," Scribbr, 26 October 2020. [Online]. Available: https://www.scribbr.com/statistics/multiple-linear-regression/#:~:text=Multiple%20linear%20regression%20is%20a,variables%20using%20a%20straight%20line. [Accessed August 2021].

[14] A. Bonner, "The complete beginner's guide to machine learning: simple linear regression in four lines of code! Medium.," 6 April 2019. [Online]. Available: https://towardsdatascience.com/simple-linear-regression-in-four-lines-of-code-d690fe4dba84. [Accessed 8 September 2021].

[15] C. Maklin, "Machine Learning Algorithms Part 11: Ridge Regression, Lasso Regression And Elastic-Net Regression. Medium.," 31 December 2018. [Online]. Available: https://medium.com/@corymaklin/machine-learning-algorithms-part-11-ridge-regression-7d5861c2bc76. [Accessed October 2021].

[16] Statistics Solutions, "Pearson's Correlation Coefficient.," 9 June 2021. [Online]. Available: https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/pearsons-correlation-coefficient/. [Accessed September 2021].

[17] Republic of the Philippines Department of Education., "Senior High School Core Curriculum Subjects.," [Online]. Available: https://www.deped.gov.ph/k-to-12/about/k-to-12-basic-education-curriculum/senior-high-school-core-curriculum-subje. [Accessed 13 December 2021].

[18] M. &. H. A. Amazona, "Modelling Student Performance Using Data Mining Techniques. Proceedings of the 2019 5th International Conference on Computing and Data Engineering - ICCDE' 19. Published.," 2019. [Online]. Available: https://doi.org/10.1145/3330530.3330544. [Accessed 5 September 2021].

[19] J. &. H. A. Arcos, "Analyzing Online Transaction Data using Association Rule Mining. Proceedings of the 2019 7th International Conference on Information Technology: IoT and Smart City. Published.," 2019. [Online]. Available: https://doi.org/10.1145/3377170.3377226. [Accessed 5 September 2021].

[20] C. S. M. I. A. F. E. S. A. A. M. J. &. S. J. Casuat, "A Development of Fuzzy Logic Expert-Based Recommender System for Improving Students'Employability. 2020 11th IEEE Control and System Graduate Research Colloquium (ICSGRC). Published.," 2020. [Online]. Available: https://doi.org/10.1109/icsgrc49013.2020.9232543. [Accessed 12 September 2021].

[21] D. McKay, "Knowing Your Aptitude Can Help You Choose a Career. The Balance Careers.," 21 August 2019. [Online]. Available: https://www.thebalancecareers.com/what-is-aptitude-526175. [Accessed September 2021].

[22] B. R. R. &. C.-B. A. Fabito, "Correlation between Student Entrance Exam Results and Academic Performance: Case of a College in a Philippine University," 2019. [Online]. Available: https://national-u.edu.ph/wp-content/uploads/2019/08/6-JSTAR5-Fabito.edited.edited1.pdf. [Accessed September 2021].

[23] S. &. S. M. Geiser, "Validity Of High-School Grades In Predicting Student Success Beyond The Freshman Year: High-School Record vs. Standardized Tests as Indicators of Four-Year College Outcomes. Research & Occasional Paper Series, 6(7).," 2007. [Online]. [Accessed September 2021].

[24] A. V. N. &. V. J. Likas, "The global k-means clustering algorithm. IAS Technical Report Series, IAS-UVA-01-02, pp.12.," 2001. [Online]. [Accessed August 2021].

[25] A. F. E. Y. P. &. B. M. M. Montalbo, "Admission Test as Predictor of Student Performance in Political Science and Psychology Students of Rizal Technological University. Asia Pacific Journal of Multidisciplinary Research, 6(3," August 2018. [Online]. [Accessed August 2021].

[26] T. N. J. P. a. H. P. Nguyen, "A comparative analysis of techniques for predicting academic performance 37th ASEE/IEEE Frontiers in Education," 2007. [Online]. [Accessed October 2021].

[27] P. N., " Factors Affecting High School Students' Career. Preference: A Basis for Career Planning Program. International Journal of Sciences: Basic and Applied Research," 2016. [Online]. Available: http://www.urs.edu.ph/wp-content/uploads/2016/06/2261-4881-1-PB.pdf. [Accessed August 2021].

[28] Q. M. N. a. K. N.V., "Drop Out Feature of Student Data for Academic Performance Using Decision Tree Techniques, Global Journal of Computer Science and Technology Vol. 10 Issue 2, pp2-5," 2010 . [Online]. [Accessed October 2021].

[29] D. B. G. G. L. G. U. T. &. Y. L. R., "Predictive Analysis Using Data Mining Techniques and SQL (HCT-I-003).," April 2014. [Online]. Available: https://www.dlsu.edu.ph/wp-content/uploads/pdf/conferences/research-congress-proceedings/2014/HCT/HCT-I-003-FT.pdf. [Accessed September 2021].

[30] J. Thompson, "Choosing the Right Clustering Algorithm for your Dataset. KDnuggets.," 2019. [Online]. Available: https://www.kdnuggets.com/2019/10/right-

clustering-algorithm.html. [Accessed September 2021].

[31] K. Tierney, "12 best class registration software solutions for 2021. The Jotform Blog.," 29 September 2021. [Online]. Available: https://www.jotform.com/blog/class-registration-software/#Enrollware. [Accessed October 2021].

[32] B. B. a. P. S. Yaday S.K., "Mining Education Data to Predict Student's Retention: A comparative Study, International Journal of Computer Science and Information Security (IJCSIS), Vol. 10, No. 2, 113-117," 2012. [Online]. [Accessed September 2021].

[33] G. L. Team, "A complete understanding of lasso regression," GreatLearning Blog: Free Resources what Matters to shape your Career, 22 March 2022. [Online]. Available: https://www.mygreatlearning.com/blog/understanding-of-lasso-regression/. [Accessed 30 March 2022].

Appendix A Evaluation Tool

Logistic Regression

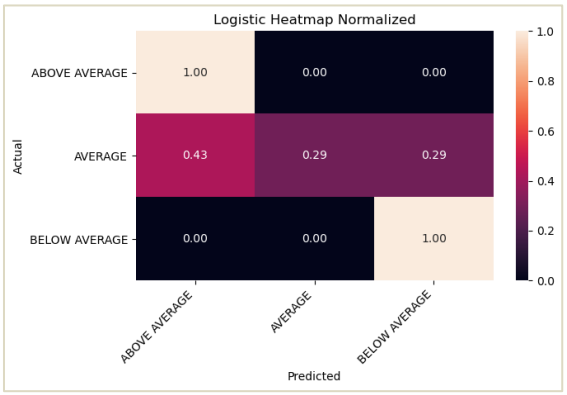


Figure 36.Logistic Heatmap (Normalized)

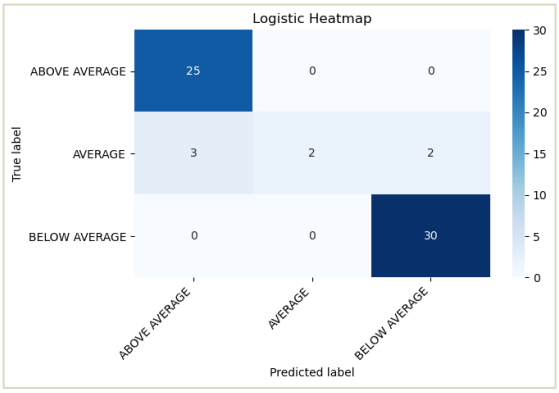


Figure 37.Logistic Heatmap

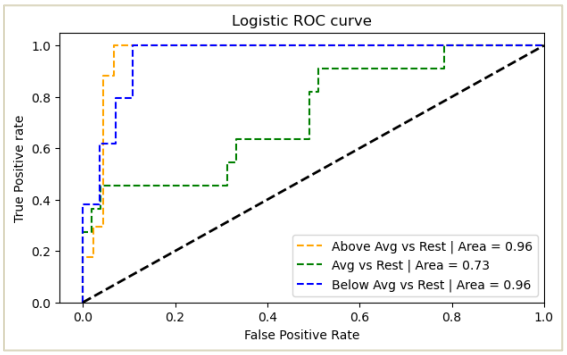


Figure 39.Logistic Curve

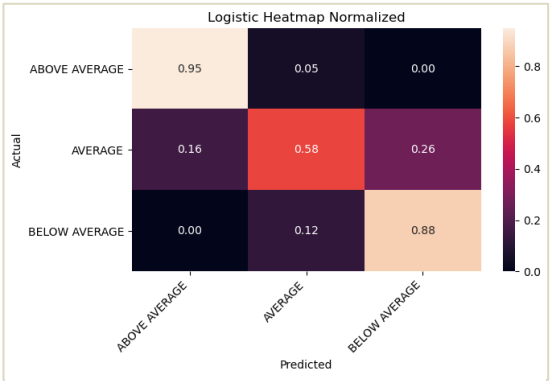


Figure 38. OHE Logistic Heatmap (Normalized)

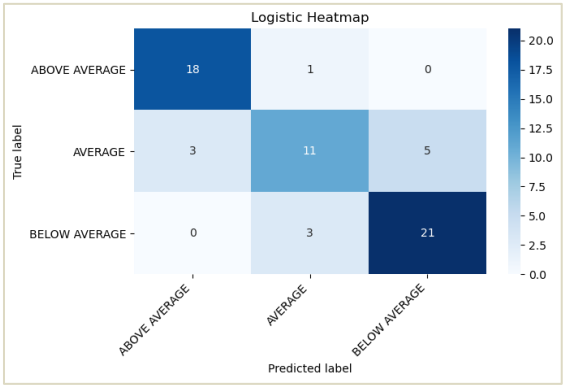


Figure 40. OHE Logistic Heat map

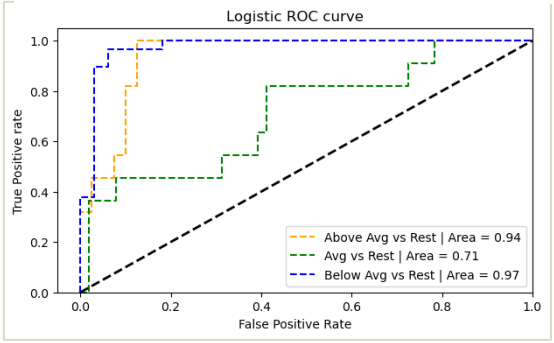


Figure 41 OHE Logistic Curve

Gradient Boosting

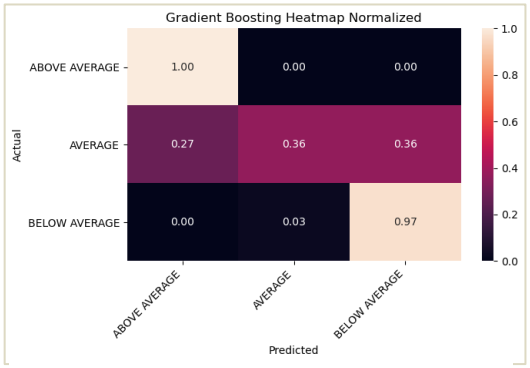


Figure 43.Gradient Boosting Heatmap (Normalized)

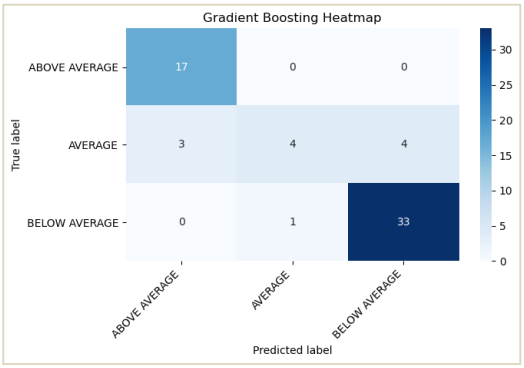


Figure 42.Gradient Boosting Heatmap

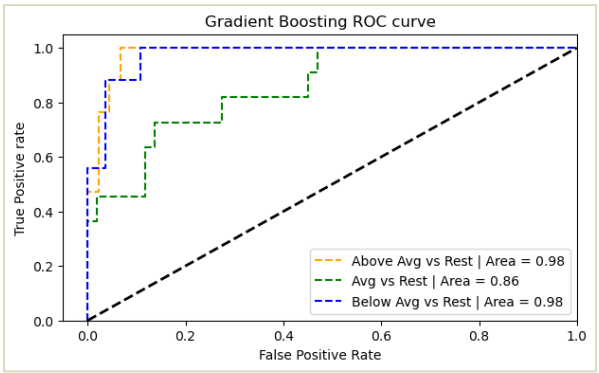


Figure 44.Gradient Boosting Curve

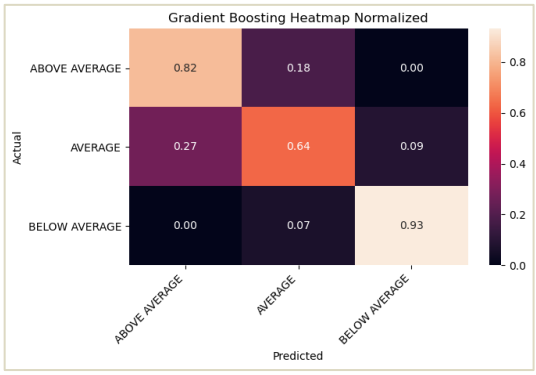


Figure 45.OHE Gradient Boosting Heatmap(Normalized)

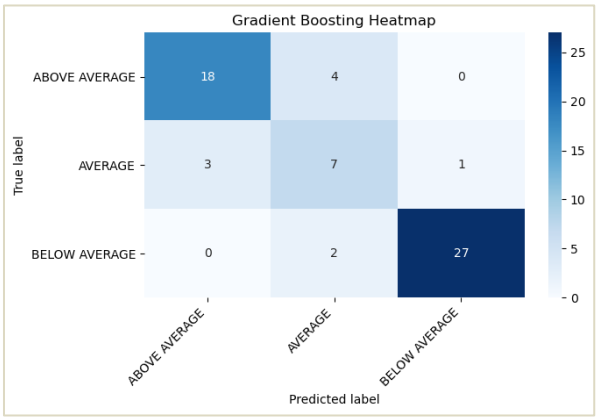


Figure 46.OHE Gradient Boosting Heatmap

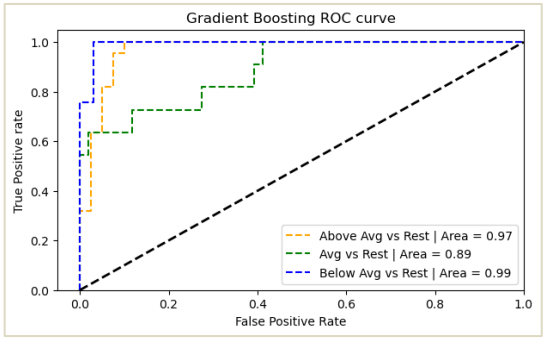


Figure 47.OHE Gradient Boosting Curve

SVM

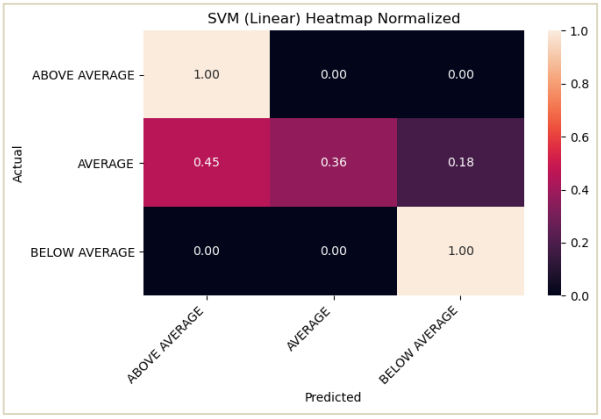


Figure 49.SVM Heatmap (Normalized)

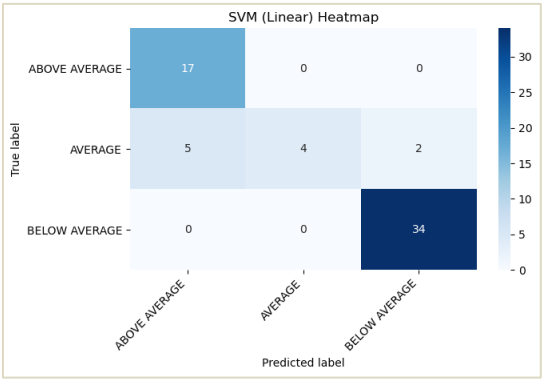


Figure 48. SVM Heatmap

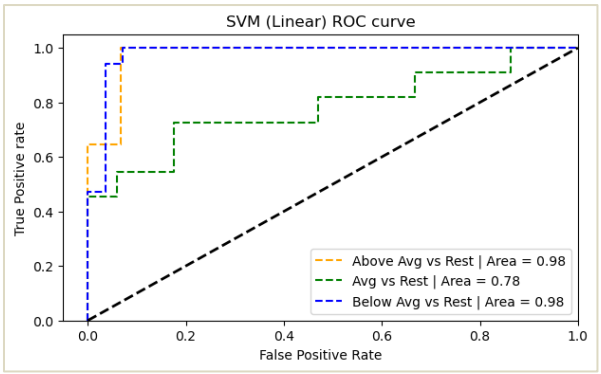


Figure 51.SVM Curve

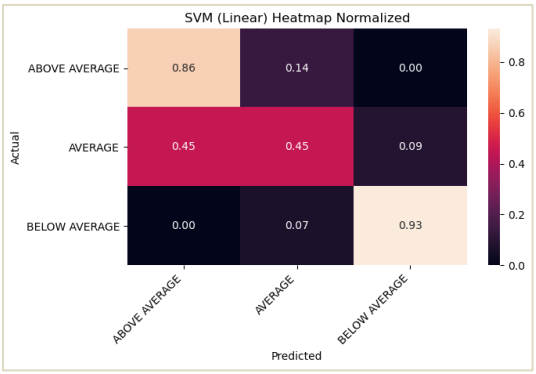


Figure 50.OHE SVM Heatmap (Normalized)

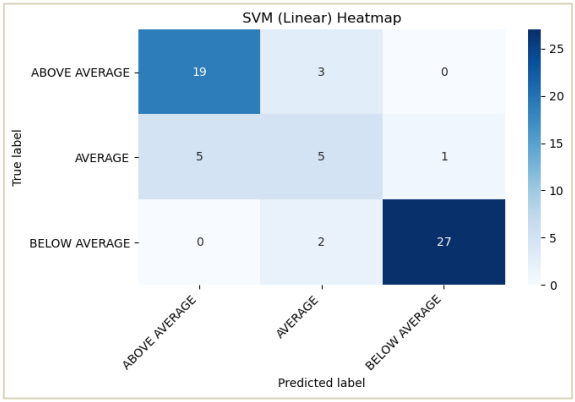


Figure 52.OHE SVM Heatmap

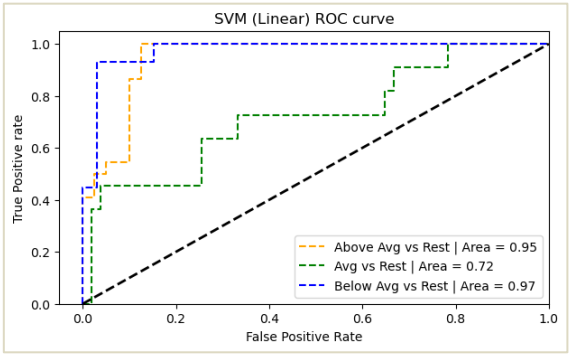


Figure 53.OHE SVM Curve

Random Forest

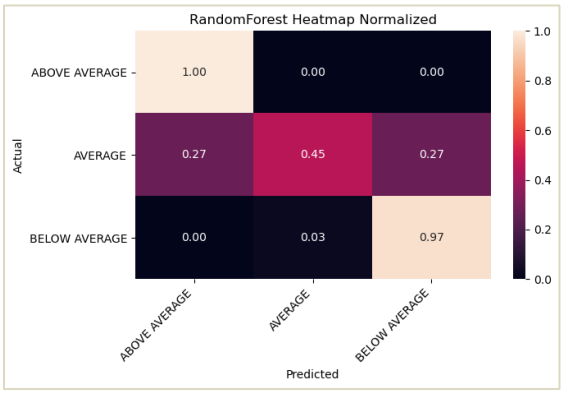


Figure 55.Random Forest (Normalized)

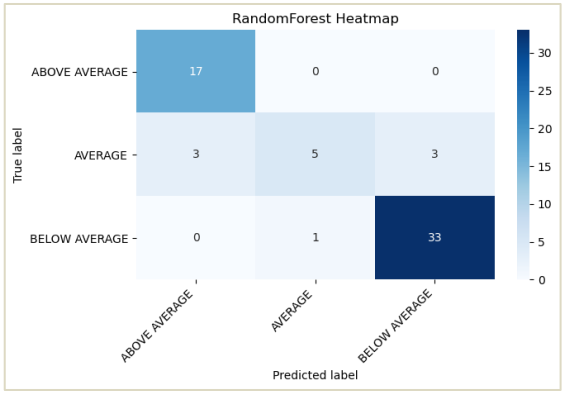


Figure 54.Random Forest Heatmap

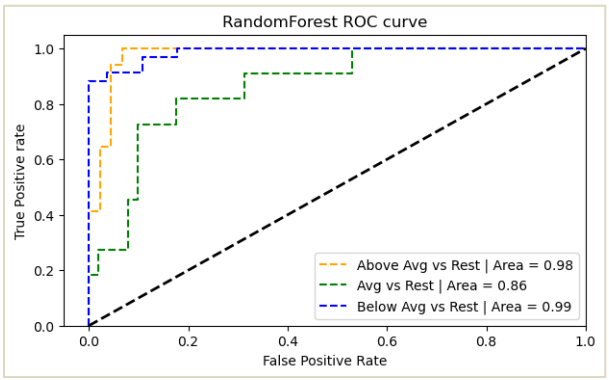


Figure 57.Random Forest Curve

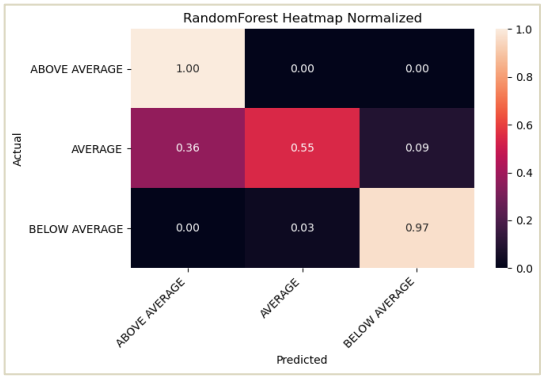


Figure 56.OHE Random Forest Heatmap(Normalized)

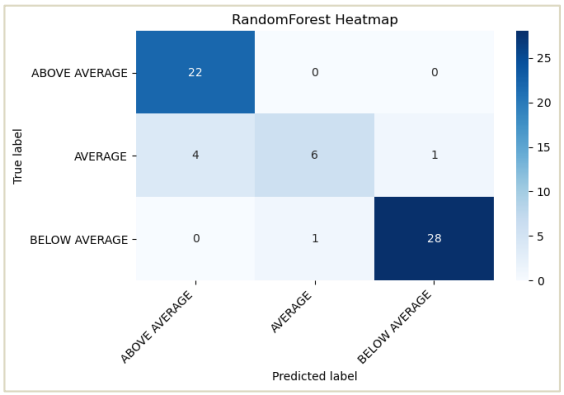


Figure 59.OHE Random Forest Heatmap

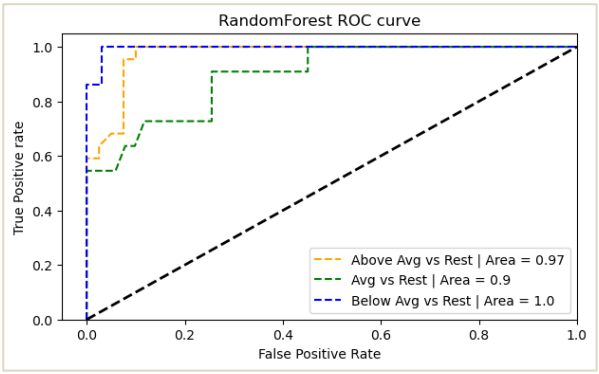


Figure 58.OHE Random Forest Curve

Neural Network

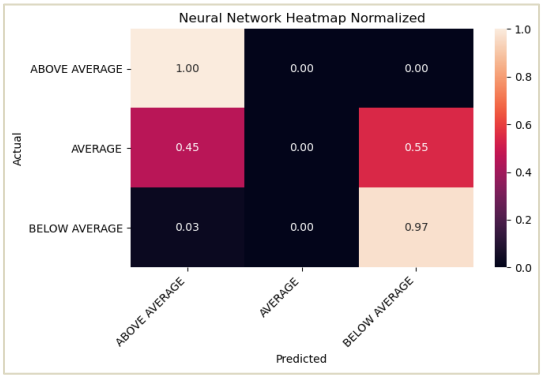


Figure 60. Neural Network Heatmap (Normalized)

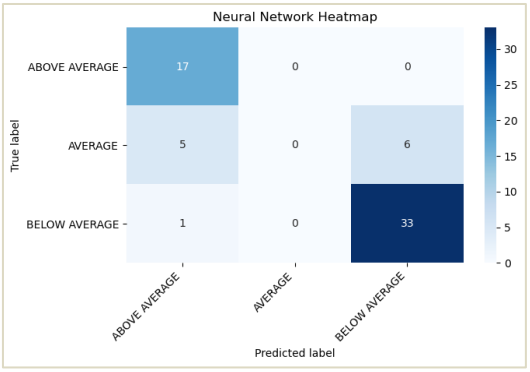


Figure 61.Neural Network Heatmap

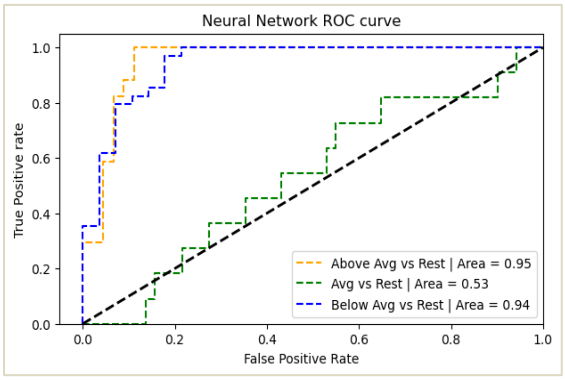


Figure 63.Neural Network Curve

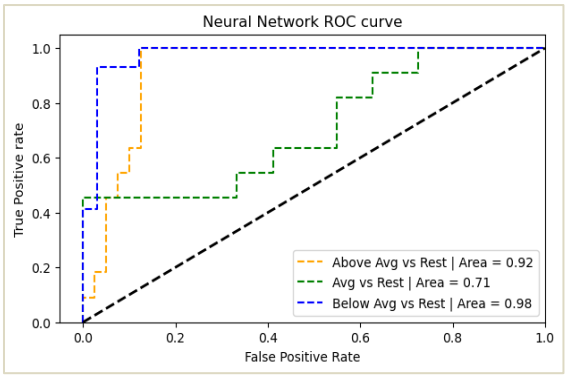


Figure 62.OHE Neural Network Heatmap (Normalized)

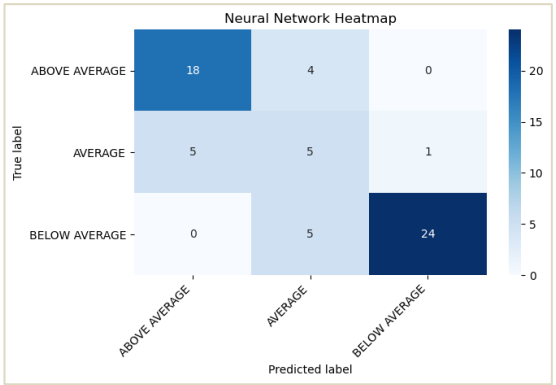


Figure 64.OHE Neural Network Heatmap

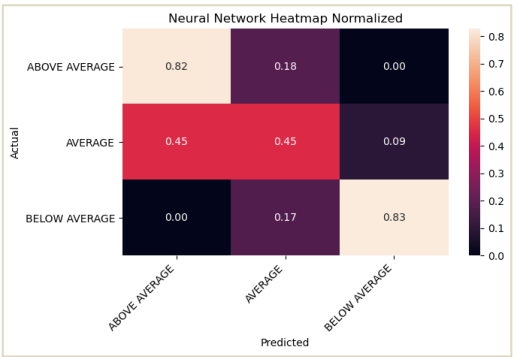


Figure 65.OHE Neural Network Curve

Naïve Bayes (Complement)

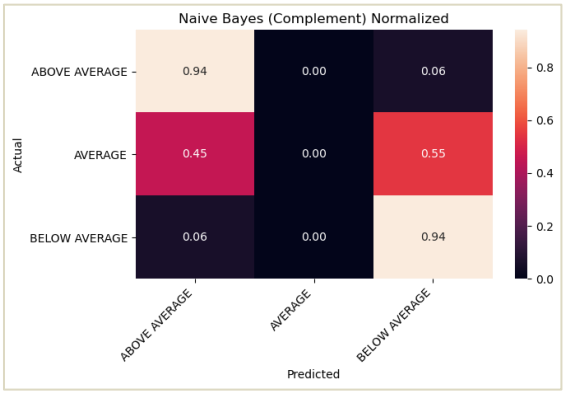


Figure 67.Naive Bayes Heatmap (Normalized)

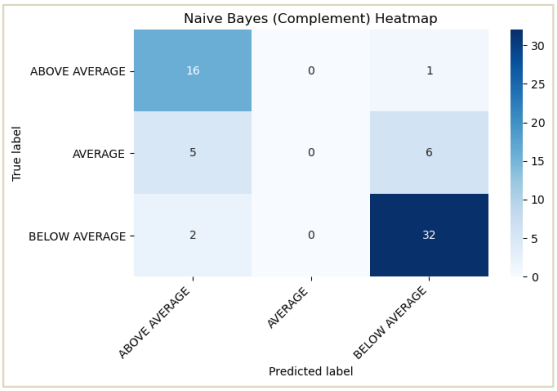


Figure 66.Naive Bayes Heatmap

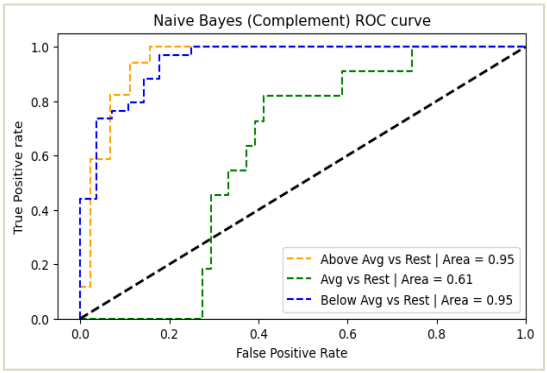


Figure 68.Naive Bayes Curve

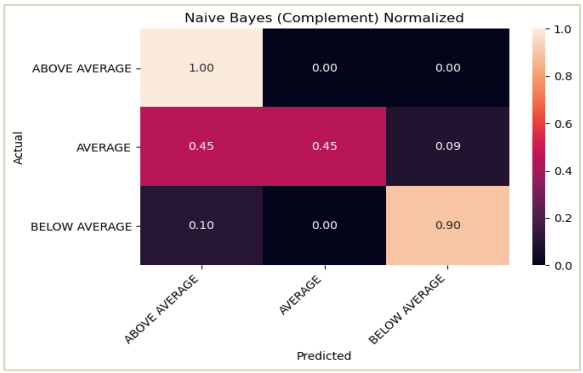


Figure 69. OHE Naïve Bayes Heatmap (Normalized)

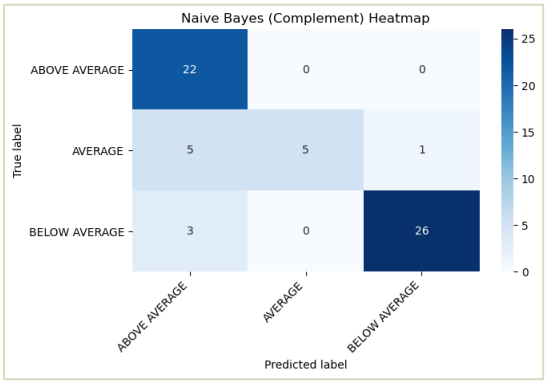


Figure 70.OHE Naive Bayes Heatmap

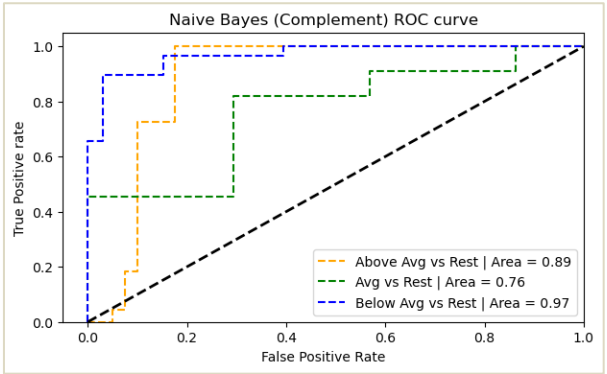


Figure 71.OHE Naive Bayes Curve

Appendix B System Development Test Cases

Project Name	Western Mindanao State University Online CET Application System
Module Name	Login Module
Created By	Theo Jay G. Sanson, Jane Stephanie J. Domingo

Test Case Id	Test Scenario	Test Case	Pre-Condition	Test Steps	Test Data	Expected Result	Post Condition	Actual Result	Status
TC_Login_001	Validate User Credentials	Enter Valid Data on both username and password fields	Valid Registered System Account	1. Enter Username 2. Enter Password 3. Press Login	Valid username Valid password	Alert "Login Successful!"	Redirect to Guidance Index Page	Same as Expected	P
TC_Login_002	Validate User Credentials	Enter Valid Data on username only	Valid Registered System Account	1. Enter Username 2. Enter Password 3. Press Login	Valid username Invalid password	Alert "Password Incorrect!"	Redirect back to login page.	Same as Expected	P
TC_Login_003	Validate User Credentials	Enter Valid Data on password only	Valid Registered System Account	1. Enter Username 2. Enter Password 3. Press Login	Invalid username Valid password	Alert "No record of given username!"	Redirect back to login page.	Same as Expected	P
TC_Login_004	Validate User Credentials	Enter Invalid Data on username and password fields	Valid Registered System Account.	1. Enter Username 2. Enter Password 3. Press Login	Invalid username Invalid password	Alert "No record of given username!"	Redirect back to login page.	Same as Expected	P

Project Name	Western Mindanao State University Online CET Application System
Module Name	Apply CET Module
Created By	Theo Jay G. Sanson, Jane Stephanie J. Domingo

Test Case Id	Test Scenario	Test Case	Pre-Condition	Test Steps	Test Data	Expected Result	Post Condition	Actual Result	Status
TC_Apply_001	Apply for CET as Student	Enter Valid Data on User Form	No Pre-condidition required	1. Click Apply 2. Fill Out Form Completely 3. Click Submit	Valid Information	Redirect to Application Page with Valid Prediction	Redirect to Student Application Page	Same as Expected	P
TC_Apply_002	View Student Applications	View Student Applications	User Logged In	1. Click View Applications 2. Select an application 3. Click View Application.	No data required	Redirect to Application Page with Data	Redirect to Application View Page	Same as Expected	P
TC_Apply_003	Print Student Applications	Generate Report for Student Application	User Logged In. Viewing Student Application	1. Click Print Application	No data required	Redirect to Application Page with Data in Print Format	Redirect to Application Print View Page	Same as Expected	P
TC_Apply_004	Validate Student Application Form	Erase Data on some required fields.	Valid Registered System Account. Apply Button Clicked.	1. Click Apply 2. Fill Out Form Incompletely 3. Click Submit	No data required	Alert “Please Fill out the Required Forms (Highlighted in Red)”	Required empty Fields are Highlighted Red.	Same as Expected	P

Project Name	Western Mindanao State University Online CET Application System
Module Name	Examination Scheduling Module
Created By	Theo Jay G. Sanson, Jane Stephanie J. Domingo

Test Case Id	Test Scenario	Test Case	Pre-Condition	Test Steps	Test Data	Expected Result	Post Condition	Actual Result	Status
TC_Exam_001	Create Examination.	Enter Valid Data on Exam Schedule Creation Form	User Logged In	1. Click Apply 2. Fill Out Form Completely 3. Click Submit	Valid Information	Redirect to View Examination Page	Examination Page should reflect inputted data.	Same as Expected	P
TC_Exam_002	View Examination	View Examination	User Logged In	1. Click View Examinations 2. Select an examination 3. Click View examination	No data required	Redirect to Application Page with Data	Redirect to Application View Page	Same as Expected	P
TC_Exam_003	Update Examination	Enter Valid Data on Exam Schedule Update Form	User Logged In. Viewing Examination	1. Click Edit 2. Change Schedule of Examination 3. Click Submit	Valid Information	Redirect to View Examination Page	Examination Page should reflect inputted data.	Same as Expected	P
TC_Exam_004	Generate Examination Report	Generate Report for Student Application	User Logged In. Viewing Examination	1. Click Print Examination	No data required	Redirect to Examination Page with Data in Print Format	Redirect to Examination Print View Page	Same as Expected	P

Appendix C Source Code

Software Source Code

Web Application Source Code - https://github.com/ZuluDoggo/ml_website

ML Model Creation Source Code - https://github.com/ZuluDoggo/model_development

Data Model Fitting Code Snippet

```
import pandas as pd
from sklearn.svm import SVC
from sklearn.model_selection import train_test_split

df = pd.read_csv('one_hot_data.csv') #Set Filepath for CSV
X = df.drop(columns={
      "sum_raw_categorical",
      "sum_raw_scores",
      "Selected",
      "sum_raw_binary"
      }, axis=1)
y = df['sum_raw_categorical']
x_train, x_test, y_train, y_test= train_test_split(X,y, test_size=0.2,
      random_state=2)
sv_model = SVC(kernel="linear",probability=True)
sv_model.fit(x_train, y_train)
```

Data Model with Web Application Integration Code Snippet

```
loaded_model = pickle.load(open('finalized_model.sav', 'rb'))
columns = [
      'school_category_numerical=1',
      'school_category_numerical=2',
      'school_type_numerical=1',
      'school_type_numerical=2',
      'school_type_numerical=3',
      'district=1',
      'district=2',
      'track_numerical=1',
      'track_numerical=2',
      'track_numerical=3',
      'strand_numerical=1',
      'strand_numerical=2',
      'strand_numerical=3',
      'strand_numerical=4',
      'strand_numerical=5',
      'strand_numerical=6',
      'strand_numerical=7',
      'strand_numerical=8',
      'strand_numerical=9',
      'strand_numerical=10',
      'english_number',
      'english_average',
      'math_number',
      'math_average',
      'science_number',
      'science_average',
      'sum_ems',
      'ems_average_number',
      'ems_average'
]

df = pd.DataFrame(index=['student'],columns=columns)
tempStudent = Student.objects.get(pk=tempForm.id)
tempSchool = School.objects.get(pk=tempStudent.school.id)
```

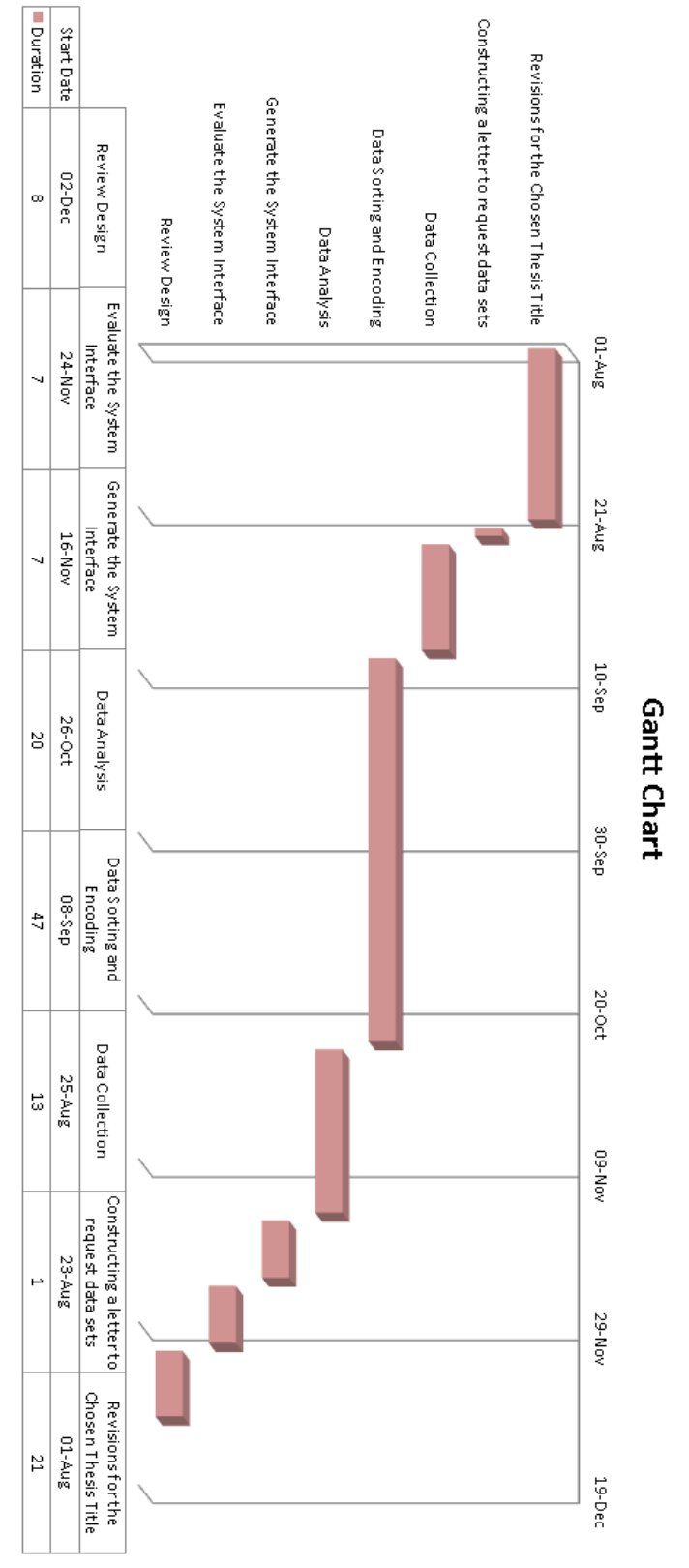
```
school_category = tempSchool.school_type
SCHOOL_CATEGORY = [ "Public", "Private" ]
for i in range(2):
      if school_category == SCHOOL_CATEGORY[i]:
            df['school_category_numerical='+str(i+1)] = 1
      else:
            df['school_category_numerical='+str(i+1)] = 0

school_type = tempSchool.school_category
SCHOOL_TYPE = [ "University", "College","School","Other",]
for i in range(3):
      if school_type == SCHOOL_TYPE[i]:
            df['school_type_numerical='+str(i+1)] = 1
      else:
            df['school_type_numerical='+str(i+1)] = 0

school_district = tempSchool.district
SCHOOL_DISTRICT = [("District I", "District I"),("District II","District II
("Other","Other"),]
for i in range(2):
      if school_district == SCHOOL_DISTRICT[i]:
            df['district='+str(i+1)] = 1
      else:
            df['district='+str(i+1)] = 0

student_strand = tempStudent.strand
STRAND = [
      'ABM',
      'STEM',
      'HUMSS',
      'GAS',
      'TVL - Information Communication and Technology',
      'TVL - Home Economics',
      'TVL - Other',
      'TVL - Food Related',
      'Industrial Arts',
      'Arts & Design / Sports',
]
for i in range(10):
      if student_strand == STRAND[i]:
            df['strand_numerical='+str(i+1)] = 1
      else:
            df['strand_numerical='+str(i+1)] = 0
```

```
student_track = ''
if student_strand == 'ABM' or student_strand == 'STEM' or student_strand ==
'HUMSS' or student_strand == 'GAS':
      student_track = 'Academic'
elif student_strand == 'Arts & Design / Sport':
      student_track = 'Sports and Arts'
else:
      student_track = 'TVL'
TRACK = ['Academic','TVL','Sports and Arts']
for i in range(3):
      if student_track == TRACK[i]:
            df['track_numerical='+str(i+1)] = 1
      else:
            df['track_numerical='+str(i+1)] = 0

SUBJECT_TYPE = [
      "English",
      "Math",
      "Science",
      "Other",
]

all_subjects = Subject.objects.all()
english_subject_list = []
math_subject_list = []
science_subject_list = []

for tempSubject in all_subjects:
      if tempSubject.subject_type == 'English':
            english_subject_list.append(tempSubject.id)
      elif tempSubject.subject_type == 'Math':
            math_subject_list.append(tempSubject.id)
      elif tempSubject.subject_type == 'Science':
            science_subject_list.append(tempSubject.id)

student_english_grades = []
student_math_grades = []
student_science_grades = []
master_subject_list = list(SubjectAssignment.objects.filter(student_id__exa
tempForm.id))
```

```
for subject_assignment in master_subject_list:
      if subject_assignment.subject.id in english_subject_list:
            student_english_grades.append(subject_assignment.value)
      elif subject_assignment.subject.id in math_subject_list:
            student_math_grades.append(subject_assignment.value)
      elif subject_assignment.subject.id in science_subject_list:
            student_science_grades.append(subject_assignment.value)

english_total = 0
for i in range(0, len(student_english_grades)):
      english_total = english_total + int(student_english_grades[i])
if len(student_english_grades) != 0:
      english_average = english_total / len(student_english_grades)
else:
      english_average = 0

math_total = 0
for i in range(0, len(student_math_grades)):
      math_total = math_total + int(student_math_grades[i])
if len(student_math_grades) != 0:
      math_average = math_total / len(student_math_grades)
else:
      math_average = 0

science_total = 0
for i in range(0, len(student_science_grades)):
      science_total = science_total + int(student_science_grades[i])
if len(student_science_grades) != 0:
      science_average = science_total / len(student_science_grades)
else:
      science_average = 0

sum_ems = len(student_english_grades) + len(student_math_grades) +
      len(student_science_grades)
ems_average_number = sum_ems / 3
ems_average = (english_average + math_average + science_average) / 3
```

```
df['english_number'] = len(student_english_grades)
df['english_average'] = english_average
df['math_number'] = len(student_math_grades)
df['math_average'] = math_average
df['science_number'] = len(student_science_grades)
df['science_average'] = science_average
df['sum_ems'] = sum_ems
df['ems_average_number'] = ems_average_number
df['ems_average'] = ems_average

X = df
prediction = loaded_model.predict(X)
tempStudent.predicted_performance = prediction[0]
tempStudent.save()
```

Appendix D **Development Timeline**



	Review Design	Evaluate the System Interface	Generate the System Interface	Data Analysis	Data Sorting and Encoding	Data Collection	Constructing a letter to request data sets	Revisions for the Chosen Thesis Title
Start Date	02-Dec	24-Nov	16-Nov	26-Oct	08-Sep	25-Aug	23-Aug	01-Aug
Duration	8	7	7	20	47	13	1	21

Appendix E **Screenshot of the System**

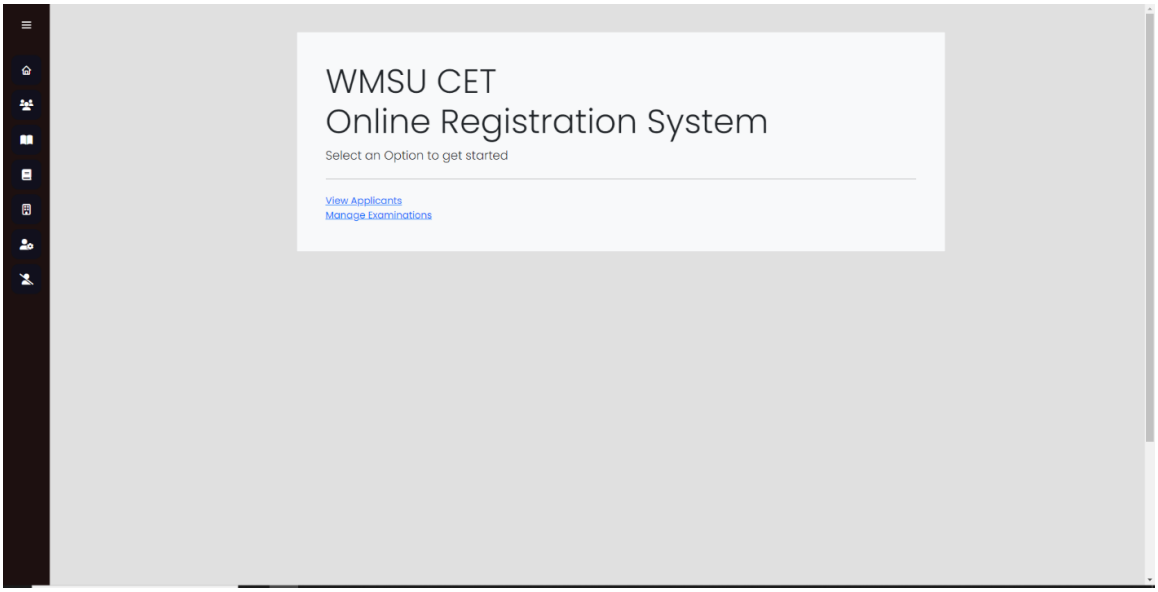


Figure 72 Dashboard

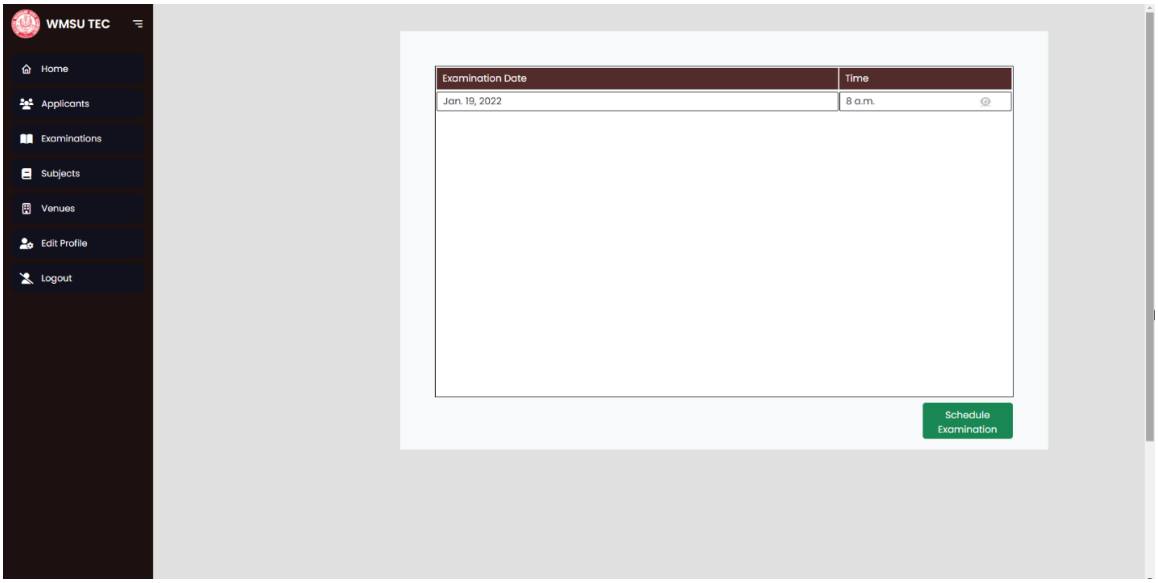


Figure 73 List of Examinations

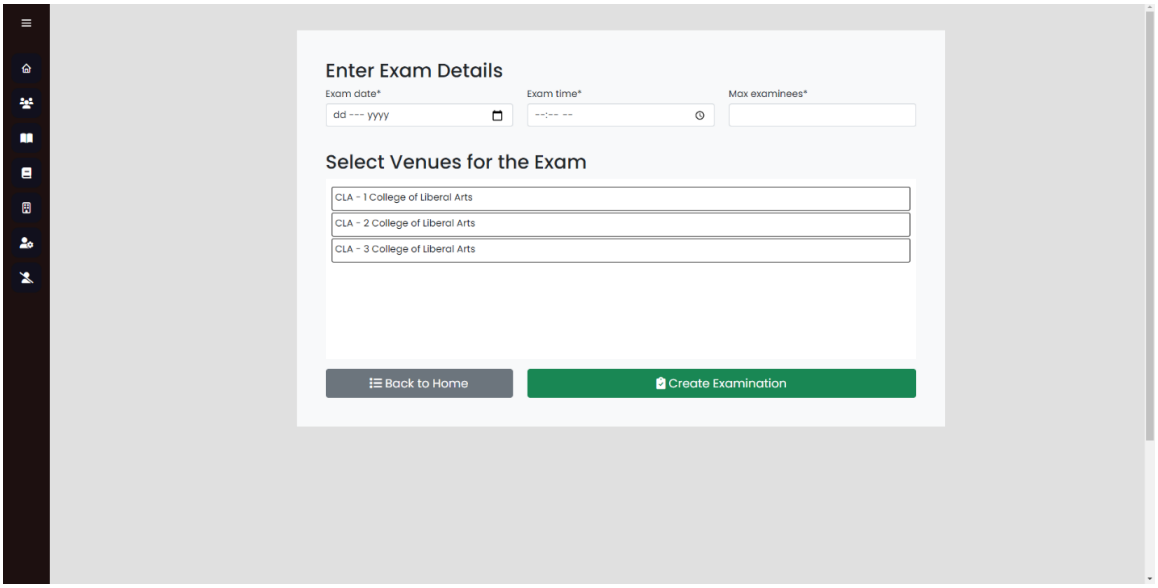


Figure 74 Adding New Examination

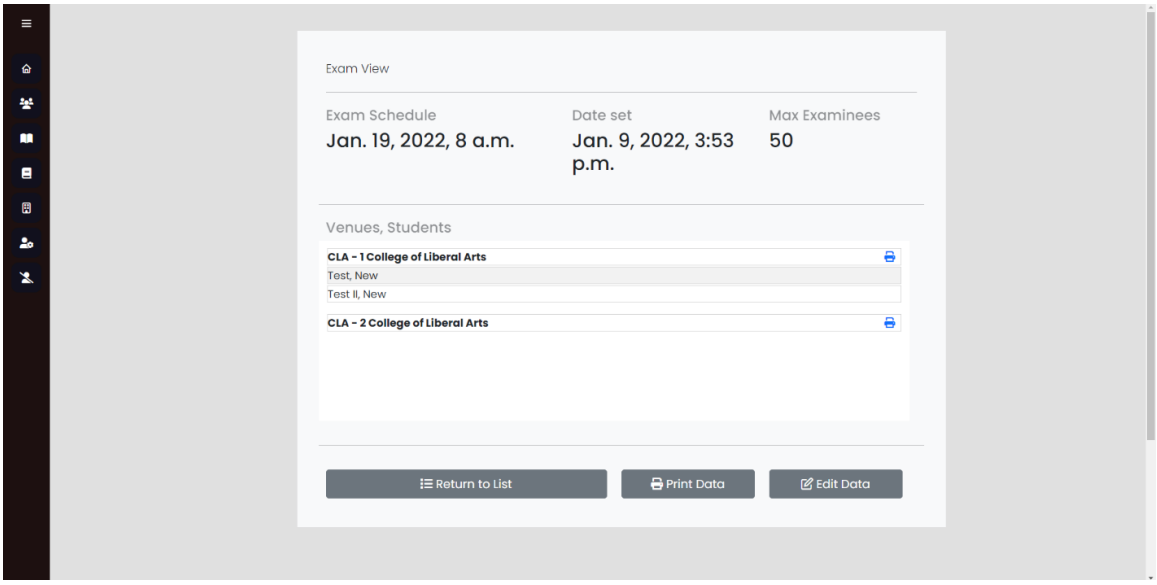


Figure 75 Examination Details

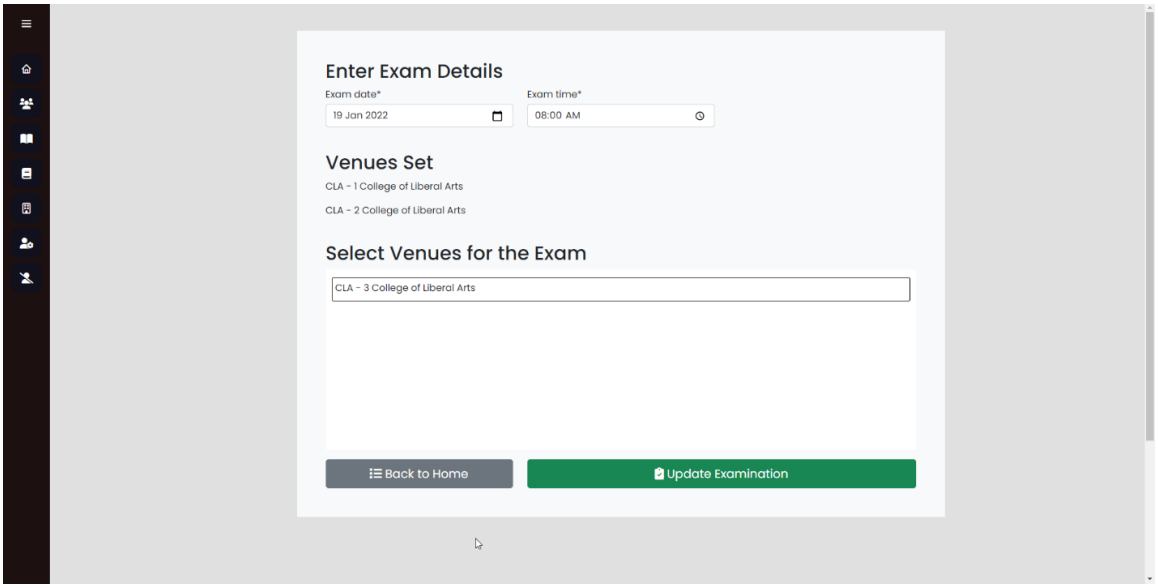


Figure 76 Edit Examination

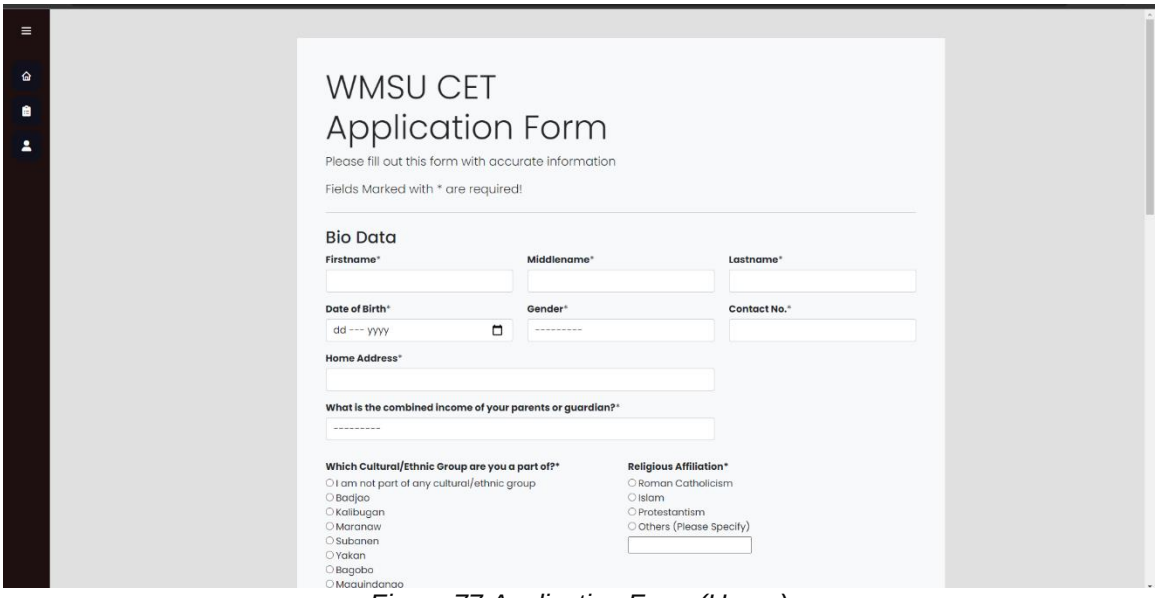


Figure 77 Application Form (Upper)

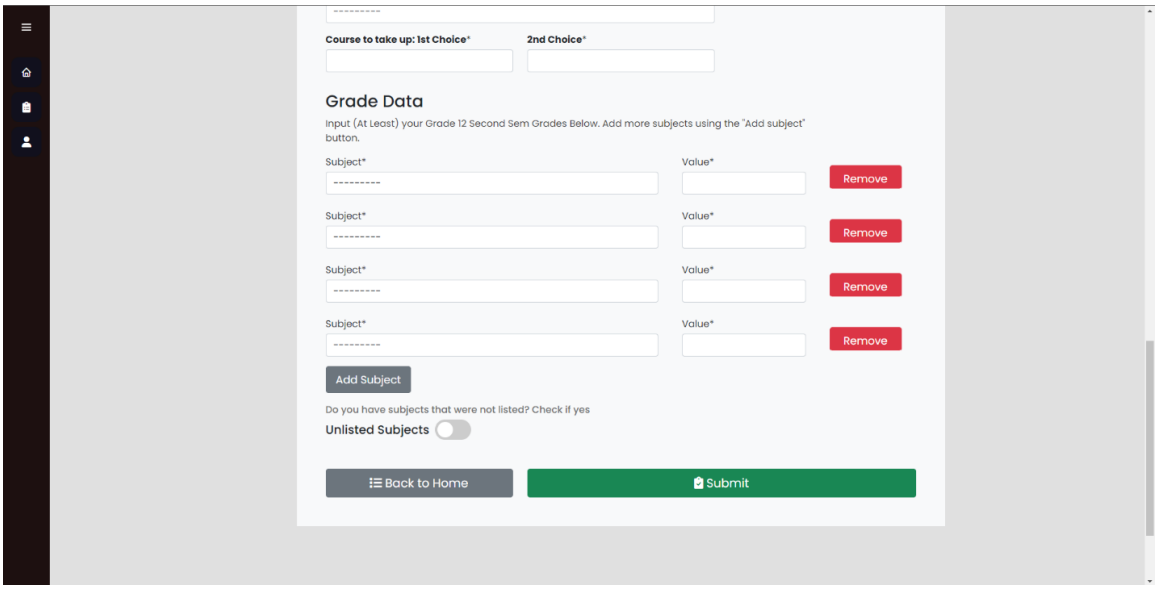


Figure 78 Application Form (Lower)

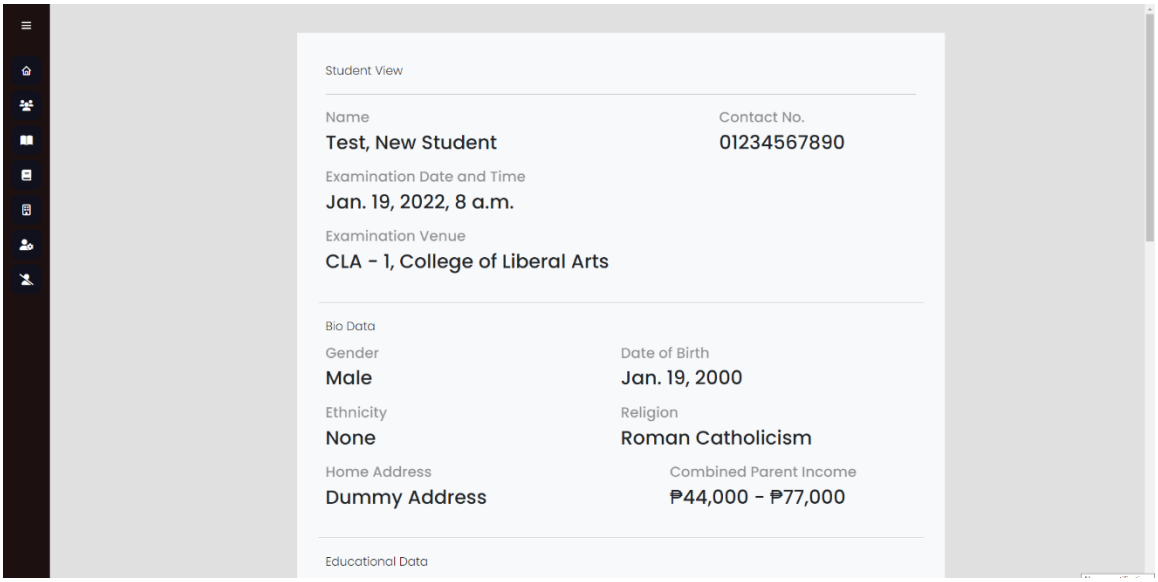


Figure 79 Student Data View (Upper)

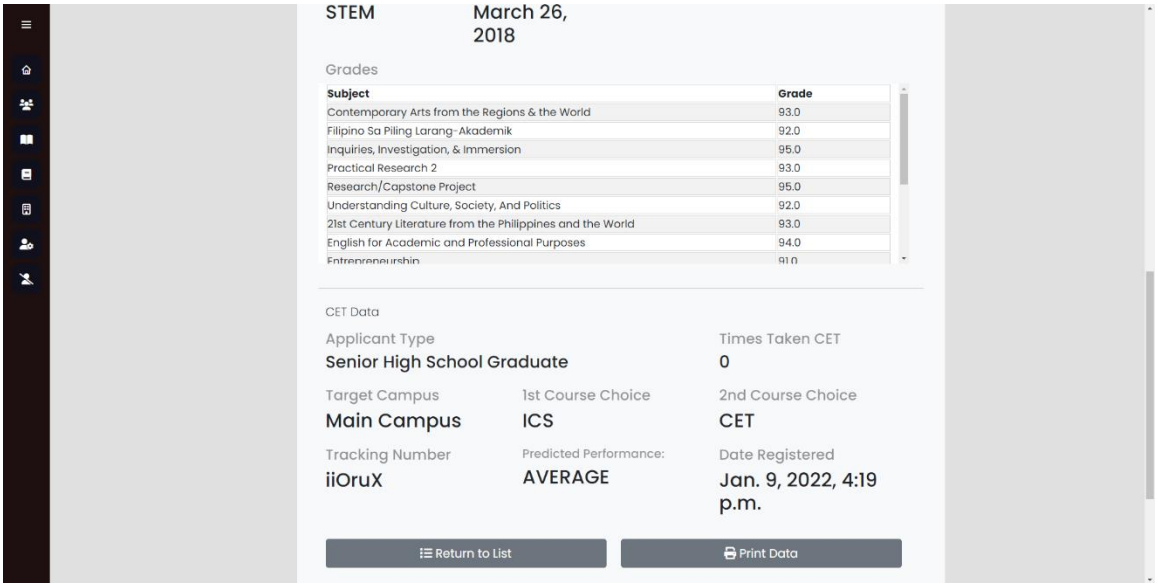
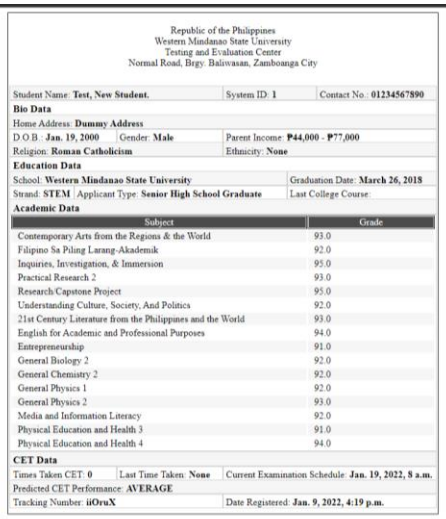


Figure 80 Student Data View (Lower)



Republic of the Philippines
Western Mindanao State University
Testing and Evaluation Center
Normal Road, Brgy. Baliwasan, Zamboanga City

Student Name: **Test, New Student.**	System ID: **1**	Contact No.: **01234567890**
Bio Data		
Home Address: **Dummy Address**		
D.O.B.: **Jan. 19, 2000**	Gender: **Male**	Parent Income: **₱44,000 - ₱77,000**
Religion: **Roman Catholicism**	Ethnicity: **None**	
Education Data		
School: **Western Mindanao State University**		Graduation Date: **March 26, 2018**
Strand: **STEM** | Applicant Type: **Senior High School Graduate**		Last College Course:
Academic Data		

Subject	Grade
Contemporary Arts from the Regions & the World	93.0
Filipino Sa Piling Larang-Akademik	92.0
Inquiries, Investigation, & Immersion	95.0
Practical Research 2	93.0
Research/Capstone Project	95.0
Understanding Culture, Society, And Politics	92.0
21st Century Literature from the Philippines and the World	93.0
English for Academic and Professional Purposes	94.0
Entrepreneurship	91.0
General Biology 2	92.0
General Chemistry 2	92.0
General Physics 1	92.0
General Physics 2	93.0
Media and Information Literacy	92.0
Physical Education and Health 3	91.0
Physical Education and Health 4	94.0

CET Data		
Times Taken CET: **0**	Last Time Taken: **None**	Current Examination Schedule: **Jan. 19, 2022, 8 a.m.**
Predicted CET Performance: **AVERAGE**		
Tracking Number: **iiOruX**		Date Registered: **Jan. 9, 2022, 4:19 p.m.**

Figure 81 Student Report Generation Output

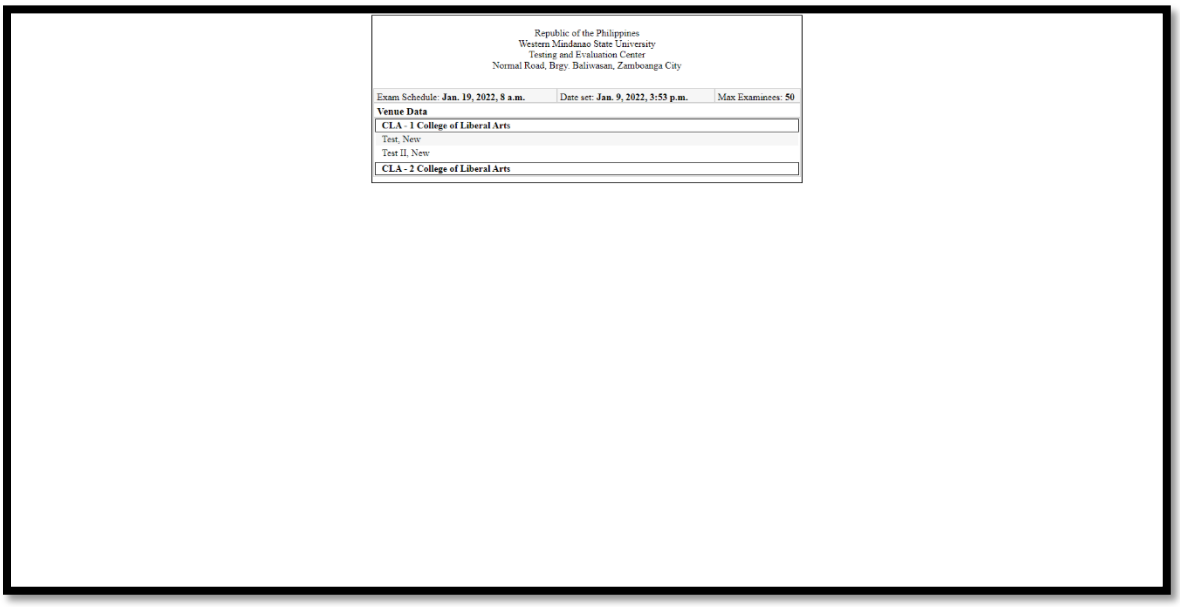


Figure 82 Exam Report Generation Output

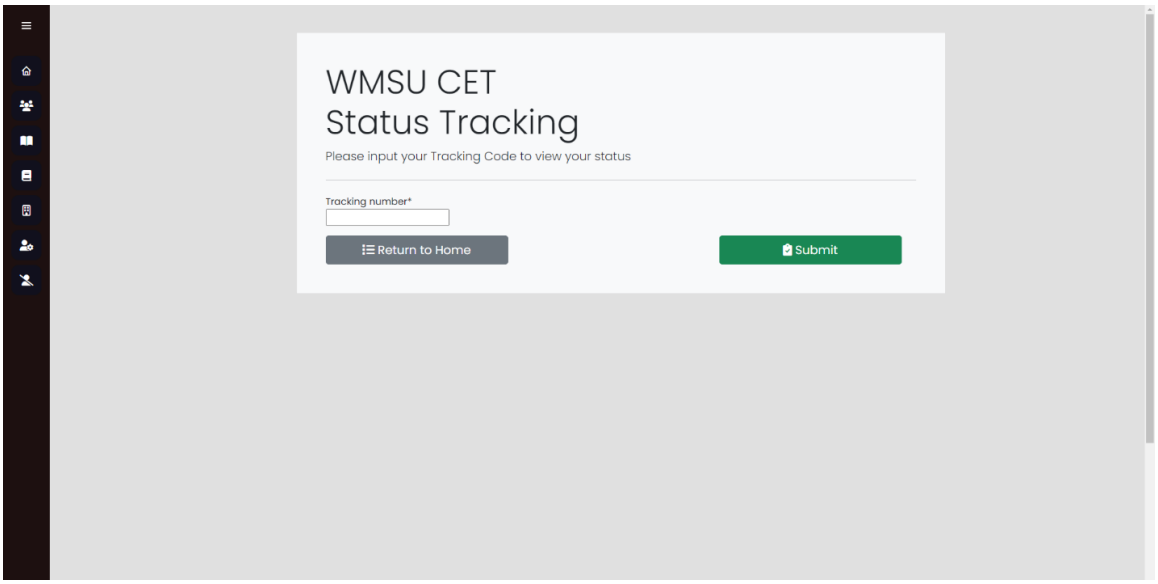


Figure 83 Student Status tracking form

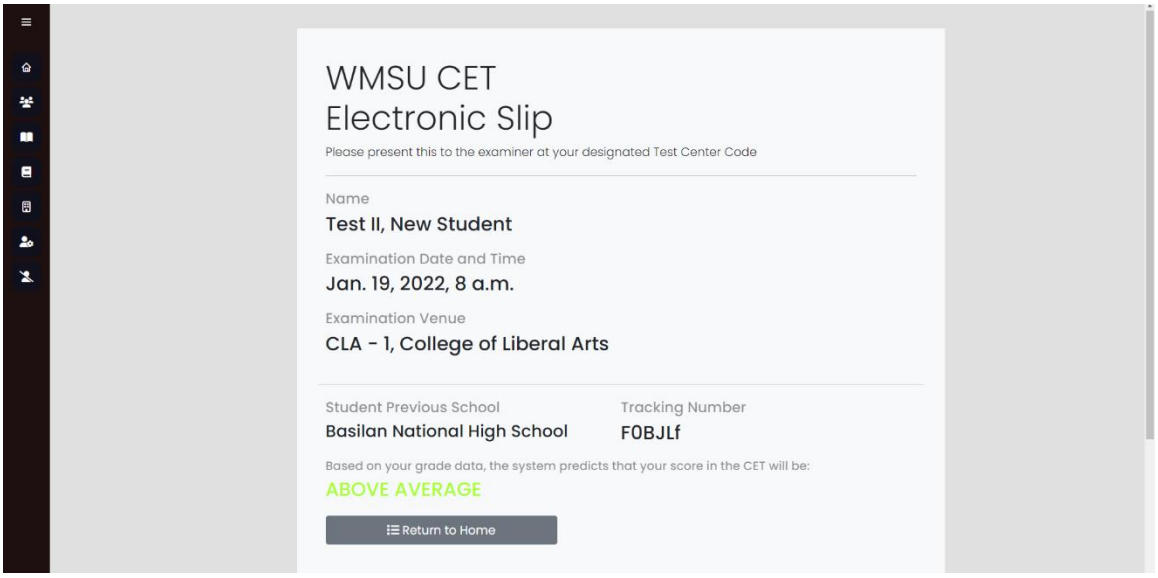


Figure 84 Student Electronic Slip View

Appendix F Curriculum Vitae

Theo Jay M'Lleno Jay G. Sanson
(+63) 977 208 4969
tjmileno@gmail.com
Estrada Drive, Brgy. San Roque, Zamboanga City, Philippines

I. PERSONAL PROFILE

A resourceful, rational, and logic-oriented Bachelor of Science in Computer Science graduate of Western Mindanao State University with internship experience in an Artificial Intelligence company, proficient with problem-solving skills that are inherent in critical thinkers such as analysis, decision making, and communication ability committed to solving problems and improving the community through digital software solutions.

II. PERSONAL DETAILS

Gender:	**Male**
Date of Birth:	**January 19, 2000**
Place of Birth:	**Zamboanga City, Philippines**
Nationality:	**Filipino**
Marital Status:	**Single**
Present Address:	**Estrada Drive, Brgy. San Roque, Zamboanga City, Philippines**

III. RELATED EXPERIENCE

Intern
SENTI.AI, Philippines (July 2021 – October 2021)
Duties and Responsibilities:
·Assistance in Creation of Technological Solutions
·Development of software code for digital programs
·Analysis of Data for Machine Learning or Artificial Intelligence Algorithms

IV. EDUCATIONAL BACKGROUND

COLLEGE

Bachelor of Science in Computer Science
Western Mindanao State University
August 2018 - June 2022

HIGHSCHOOL

Integrated Laboratory High School
Western Mindanao State University
2012 – 2018

V. SKILLS .

TECHNICAL SKILLS
·Proficient in the following Programming Languages including C++, C#, Java, Python, HTML, CSS, PHP, JQuery, and Javascript
·Capable of Developing apps for various Operating Systems such as Windows, Android, and iOS
·Video Editor proficient in Adobe Premiere Pro and After Effects
·Accomplished Sound Editor/Manager proficient with NCH Wavepad and FL Studio
·Proficient with MS Office Suite (Word, Powerpoint, Excel)
·Has a 95 WPM Typing Speed
·Specializes in Backend Software Development
·Proficient in Web Development
·Capable Debugger

SKILLS WITH DATA
·Proficient in software creation, computer programming, and coding
·Proficient with solving logic-based problems.
·Is able to utilize various forms/techniques of Data Analysis to gain insight from data sets.
·Can present data and insights in an easily digestible manner through Data Visualization

SKILLS WITH PEOPLE
·Interpersonal Communication
·Highly proficient in English, fluent in Tagalog.
·Able to express ideas and thoughts both comprehensively and concisely
·Confident public speaker
·Experienced in managing teams, well-versed with leadership skills
·Proficient in delegating tasks to the right people based on their ability

VI. CERTIFICATIONS

·Received Certificate of Recognition for inclusion in Dean's List with a GPA of 1.6304, S.Y. 2018 – 2019 (October, 2019)

·Received Certificate of Recognition for Best Software Engineering Project (June, 2021)

·Received Certificate of Recognition for Most Outstanding Software Developer (June, 2021)

VII. ACHIEVEMENTS

·Participant,
Young Leaders for Reslience Program: Enterprise Design Thinking Workshop
-IBM Philippines
-September 6 – 7, 2019
·Google Cloud Fundamentals: Core Infrastructures (Developing Apps for Google Cloud Platform)
-Coursera
-Certified August 09, 2021
·Getting Started With Application Development (Developing Apps for Google Cloud Platform)
-Coursera
-Certified August 12, 2021
·Securing and Integrating Components of your Application (Developing Apps for Google Cloud Platform)
-Coursera
-Certified August 13, 2021

Jane Stephanie J. Domingo
095576528515
janestephanied@gmail.com
A.Alvarez Drive, Talon-Talon, Zamboanga City

I. PERSONAL PROFILE

A Practical, Fearless, and Passionate Bachelor of Science in Computer Science Major in Software Engineering Graduate of Western Mindanao State University with a GPA of 1.95 skilled in leadership and Front End Designing, aims to help the community solve a complex problems by developing software/system.

II. PERSONAL DETAILS

Gender:	**Female**
Date of Birth:	**December 05, 1999**
Place of Birth:	**Zamboanga Doctors Hospital**
Nationality:	**Filipino**
Marital Status:	**Single**
Present Address:	**A. Alvarez Drive, Talon-Talon, Zamboanga City**

III. RELATED EXPERIENCE

Intern
WMSU Library
June 2021 - August 2021
Duties and Responsibilities:
***Encoding the Past Theses and List them in descending order by year**
***Designing Facebook Post for Book Week 2021**
***Organizing the Theses Books in order**
***Had a Little Knowledge about Chattbot**

IV. EDUCATIONAL BACKGROUND

COLLEGE

Bachelor of Science in Computer Science
Western Mindanao State University
August 2021 - June 2022

HIGHSCHOOL

Senior High School
GAS - Nursing
Western Mindanao State University
With Honor (92%)
June 2016 - March 2018

Junior High School
Zamboanga City High School (Main)
June 2012 - March 2016

ELEMENTARY

Tetuan Central School
June 2006 - March 2012

V. SKILLS .

TECHNICAL SKILLS

a. Platform
Windows 10

b. Other Tools/Software
Proficient in Adobe Photoshop
Microsoft Family
Proficient in Adobe Premiere

c. Expertise
Web Development (HTML & CSS)
Layout Design

SKILLS WITH DATA

a. Documentation
Proficient in MS Word

SKILLS WITH PEOPLE

a. Interpersonal Communication
Confident Speaker
Fluent in English, Tagalog and Chavacano

b. Project Version Control
Time Management
Critical Decision Making
Capable of managing a Team
Equipped with leadership skills

VI. CLUBS/ORGANIZATION

GENDER CLUB, ICS
Western Mindanao State University
AUDITOR
AUGUST 2021-JUNE 2022

VIII. INTEREST/HOBBIES

Layout Designers for Invitations, Tarp and Labels
Social Media Manager
Cooking Dishes & Baking
Creating Make up looks and filming it
Sewing tops, thrift flip

.

Appendix G Certificate of Proofreading

Republic of the Philippines
Western Mindanao State University
College of Computing Studies
DEPARTMENT OF COMPUTER SCIENCE
Zamboanga City

Date: April 20, 2022

Certificate of Proofreading

Vicenta Princess C. Gozum

This document certifies that the Thesis listed below has been proofread for appropriate English language usage, grammar, punctuation, and spelling by a professional native English–speaking editor.

Authors: Theo Jay M'Lleno G. Sanson and Jane Stephanie J. Domingo

Thesis Title: Predicting Western Mindanao State University College Entrance Test Scores Based on Student Profile and Senior High School Grades Using Data Mining Techniques

Princess

Vicenta Princess C. Gozum

Proofreader