Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 32 No. 9s (2025)

Machine Learning using Recursive Feature Elimination of Students' Data in Singapore During Pre-Covid and Covid

[1]Ts. Dr. Amna Saad, [2]Prof. Dr. Cordelia Mason , [3]Fareed Kaleem Khaiser
[1]*Malaysian Institute of Information Technology (MIIT) , Universiti Kuala Lumpur* Kuala Lumpur, Malaysia, amna@unikl.edu.my, Orcid ID: 0000-0002-8553-4348
[2]Universiti Kuala Lumpur Business School, *Universiti Kuala Lumpur* , Kuala Lumpur, Malaysia cordelia@unikl.edu.my, Orcid ID: 0000-0003-0712-0809

[3]Malaysian Institute of Information Technology (MIIT) , Universiti Kuala Lumpur, Kuala Lumpur, Malaysia, fareed.kaleem@s.unikl.edu.my, Orcid ID: 0000-0002-1574-0285

Article History:

Received: 12-01-2025

Revised: 15-02-2025

Accepted: 01-03-2025

Abstract:

Student career management is a critical activity in the education sector. Machine learning method, Recursive Feature Elimination (RFE), has been used to find key features for prediction tasks with the help of Students' data. Using RFE, which is a machine learning technique, massive datasets are analyzed, and the most pertinent features are chosen for predictions, increasing computing speed and accuracy. RFE has been combined with several machine learning techniques for feature selection, such as support vector machines (SVM), decision trees, and random forests, and many more. This study focuses on comprehending students' data and evaluates the effect on graduate employability based on predictive power of Machine Learning algorithms. Post COVID, the data imply that the COVID-19 epidemic has significantly hampered the graduate employment for governments, corporations, and the people. The study offers potential answers in analyzing the independent variables that affect the employability aspects using Machine Learning Algorithm's performances in dealing with pre-COVID and COVID datasets.

Keywords: recursive feature elimination, undergraduates' employability, machine learning, model performance, classification, work placements, predictive analytics.

I. INTRODUCTION

To get the greatest outcomes in a given machine learning assignment, choosing the ideal subset of features from a raw dataset becomes more and more important as dataset sizes increase (Jeon & Oh, 2020). One common method for feature selection is recursive feature removal. RFE's popularity can be attributed to its simplicity of setup and operation, as well as its ability to identify the features (columns) in a training dataset that have a higher probability of being effective in accurately predicting the target variable (Brownlee, J., 2020). The amount of features to select and the algorithm to help with feature selection are two important setup decisions while using RFE. It is feasible to look into both of these hyper parameter configurations, even though they do not significantly affect the approach's efficacy (Brownlee, 2020).

A restricted and efficient feature (variable) selection process is necessary for the development of a classification model. Overfitting problems are frequently caused by high-dimensional datasets, which

hinders the creation of reliable models. Moreover, these datasets often produce models with low classification accuracy and need a significant amount of processing power and storage capacity. This was referred regarded as the "curse of dimensionality". To tackle these issues, a representative subset of features must be selected (Jeon & Oh, 2020).

Feature selection is used in many areas to choose the best subset of features, including image processing, biology, health, finance, manufacturing, and production. A feature selection method called Recursive Feature Elimination (RFE) chooses the optimal feature subset by considering the learning model and classification accuracy. The weakest feature that reduces "classification accuracy" is systematically removed once a classification model has been developed using conventional RFE. Recently, a unique RFE approach was proposed that picks the least significant features for deletion by evaluating "feature (variable) relevance" using a support vector machine (SVM) model instead of "classification accuracy". Other classification models with integrated feature assessment processes, including random forests (RFs) and gradient boosting machines (GBMs), can also be applied using this method (Jeon & Oh, 2020).

A training dataset can be used to teach a classifier the feature weights, or the relative relevance of each feature. Following the weighted ranking of each feature, the feature with the lowest weight value was eliminated. The classifier was then retrained using the leftover data until its feature set ran out. Ultimately, all features can be rated using the feature-importance-based RFE technique. It has been demonstrated that by mitigating the shortcomings of filter and wrapper approaches, this embedded feature selection method performs better (Jeon & Oh, 2020).

In order to evaluate the effectiveness of the algorithms, we have carefully chosen the most crucial features required for predictive analytics, in this study.

II. LITERATURE

Employability of graduates has been predicted and prescribed using data-mining algorithms (Héritier et al., 2023). Nevertheless, the studied literature does not specifically address Recursive Feature Elimination (RFE). Researchers Aam et al. (2022), Renato et al. (2022), and Norfarahzatul et al. (2022) employed a range of data mining techniques, such as ensemble models, machine learning algorithms, and classification algorithms like Random Forest, Neural networks, decision trees, logistic regression, support vector machines, and Naïve Bayes. These methods have been applied to employability analysis and prediction, employability factor identification, and model accuracy comparison (Aniss et al., 2020). Despite being a widely utilized feature selection method in data mining, RFE is not mentioned by name in the abstracts that are supplied. Consequently, more investigation is required to examine the application of RFE in gauging the employability of graduates.

III. RESEARCH METHODOLOGY

A research methodology is derived based on the following phases which are important part of the conceptual framework from forming the Idea & understanding the organization to the results and discussion of the case studies. Specifically for the results machine learning method is used to drill into data to determine the predictive power of the algorithms.

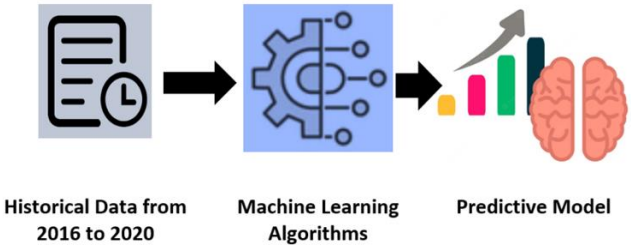

Figure 1 – Research Methodology for the Conceptual framework

The following stages, which are crucial to the conceptual framework, are the basis for a research technique. They range from idea formation and organization comprehension to case study outcomes and discussion. In particular, the predictive power of the algorithms was ascertained by delving further into the data using the machine learning approach.

We may further enhance this model by applying our previously developed conceptual framework, which can serve as the basis for the methodological framework of the machine learning-based prediction model.

This study builds on a previous investigation, which provides a brief overview of the study's conceptual framework for the Job Matching Model. Four independent variables make up the framework, which is based on the profile of an undergraduate student: major course, cumulative GPA, advantage point for extracurricular activities, and internship data. The institution plays a crucial role in determining whether an undergraduate student is successful in landing an internship with a Partnering Company (Industry).

In the event that an undergraduate is unemployed, they have two options: either they use Corrective Actions to obtain the three essential 21st Century Skills listed by the Institute of Higher Learning—Customer Orientation, Collaboration, and Communication—as well as the Course of Competencies (CoCs); or they get support from SG Skills Future to learn new skills, apply for jobs, and modify their resumes in order to apply for jobs and land a job. Undergraduates may need to acquire new competency-based abilities as a result of the procedure if their attempts prove unsuccessful. Every level of the remedial action involves the use of the feedback system to ensure that the undergraduate succeeds in finding employment.

The government, an additional entity in the updated framework, contributes significantly economically by funding 90% of upskilling and reskilling initiatives for Singaporeans and Permanent Residents as part of government initiatives aimed at nation-building.

IV. DATA ANALYSIS

The experiment showed how important feature selection is while processing classification data. Feature selection becomes more and more important, especially for datasets with a large number of variables and features. It increases the accuracy and performance of categorization by removing unnecessary variables. Chen et al. (2020) addressed the significance of feature selection with four main points. Reduce the amount of parameters in the model first, then increase training speed, reduce overfilling by increasing generalization, and break free from the dimensionality curse. Evaluations and comparisons of the effectiveness and precision of the K-nearest neighbors (KNN), Random Forest

(RF), and Support Vector Machines (SVM) classification models were also carried out. The optimal classifier is the model with highest accuracy. The highest kappa values and percentage accuracy are obtained by the use of recursive feature elimination (RFE) (Chen, R. C., et al., 2020).

Multi-class classification was employed in this study's analysis to provide dependable results, which required pre-processing the data prior to cleaning. The employability target (dependent variable), CCA Advantage Points, and student extracurricular activity missing values were removed from the dataset.

After the datasets were gathered and filtered, data correlation and variance were noted, and the experiment's decision about which variables to include and exclude was made.

This study uses a lot of modeling since the data it generates is more reliable, consistent, and organized. RFE was used to assess the most important features selection factors in this investigation. The accuracy of the results and the speed at which the results may be computed are the most important factors in selecting a model.

In this case study, a confusion matrix is utilized to accomplish the following goals:

The quantity of accurate positive forecasts is known as True Positives (TP).

The quantity of accurate negative forecasts is known as True Negatives (TN).

The quantity of erroneous positive predictions is known as False Positives (FP).

The quantity of inaccurate negative forecasts is known as Negative Prediction Errors (FN).

The algorithms are crucial to this investigation since they produce the required outcomes. Artificial Neural Networks (ANNs), Decision Trees, Bagging Classifiers, Ada Boost, Random Forests, Extremely Random Forests, Cat Boost, and lightGBM are a few examples of techniques.

A receiver operating characteristic (ROC) curve is a graph that displays a classification model's True Positive Rate (TPR) and False Positive Rate (FPR) at each classification threshold. Regularly evaluating a logistic regression model with various classification criteria would be inefficient in terms of gaining points on the ROC curve. Thankfully, there is an effective sorting-based technique called the AUC that can be utilized to collect this data. The acronym AUC stands for "Area under the ROC Curve." The whole two-dimensional region from (0, 0) to (1, 1) is measured under the whole ROC curve (It brings integral calculus to mind.). The range of the AUC was 0 to 1. Classification: ROC Curve and AUC (n.d.) states that an AUC of 0 corresponded to a model that made 100% of its predictions incorrectly, and an AUC of 1 to a model that made 100% of its predictions correctly.

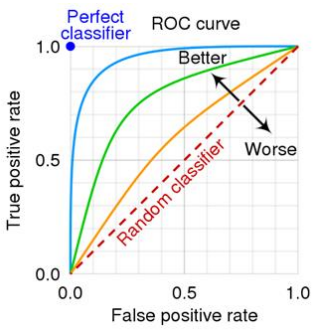

Figure 2 shows a "Better" and "Worse" classifier in ROC space (Source: Wikipedia)

The degree of agreement between the observed and predicted classes was then visualized using Cohen's kappa statistic, which was used to evaluate each model's performance.

V. RESULTS & DISCUSSION

This part displays the results and analysis from the training and experiments.

A. Perceived Employability Prediction (PRE-COVID):

The pre-processing techniques altered the original dataset to prepare it for analysis. First, the missing values in the dataset were removed. The missing values included students who dropped out of a security technology course at the Institute of Higher Learning (IHL) or did not finish their internships or extracurricular activities. Seven records—a total of 282 records—were removed from the original dataset at this point.

From the original dataset, repetitive variables were eliminated. For instance, the original dataset had variables for name, race, and student ID. Due to their assistance in the studies, all three of the variables were eliminated from the dataset. In addition, unnecessary variables were eliminated from the initial dataset. For instance, the first dataset had details on the Internship Company, title, and salary. Since they had no effect on the investigation, these variables were removed (Khaiser, F. K., Saad, A., & Mason, C., 2021).

Lastly, the dataset was made easier to analyze by making changes to the remaining variables and values. For the 282 entries in this investigation, a single Course Code was applied. The Security Technology department is reflected in the course code. The characteristics of selective programs, centers, and student organizations might involve additional computations and are outside the purview of this study. Every variable was changed to a numerical value in preparation for data cleaning. Moreover, the dataset was cleared of twenty percent of the records. Out of the 56 records, 20% assessed the machine learning models' performance through testing; as a result, those records were excluded from the machine learning model development process. There were 282 records in the final pre-processed dataset used for the analysis.

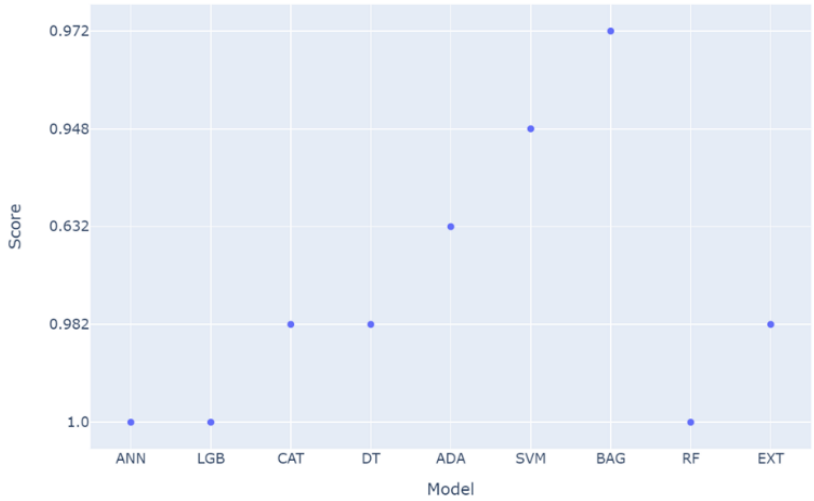

Figure 3 - Pre-Covid Overall Analysis Result

Three of the nine models (Random Forest, Light Gradient Boost, and Artificial Neural Networks) had particularly high performance.

The Random Forest (RF), Light Gradient Boost (LGB), and Artificial Neural Network (ANN) algorithms have outperformed the other six models. Next, the two selected models, Random Forest and Extra Trees, are to be cross-validated using the balanced accuracy scores (also known as macro average arithmetic), which are computed as the average of the accurate hits for each class.

B. Perceived Employability Prediction (COVID):

The original dataset was altered using pre-processing procedures to make the data ready for analysis. Initially, the dataset's missing values were eliminated. Students who failed to submit information to the Institute of Higher Learning (IHL) for their co-curricular activities or internships in a security technology course were among the missing values. In total, 201 entries were removed from the original dataset in this step.

The original dataset's repeating variables were then eliminated. For instance, the original dataset had variables for race, student name, and student ID. Since all three of these variables are not very helpful in the studies, they were all taken out of the dataset. Furthermore, unnecessary variables were eliminated from the original dataset in order to conduct this analysis. For instance, the original dataset included details about the title, pay, and Internship Company. These variables were eliminated because they are not pertinent to the study.

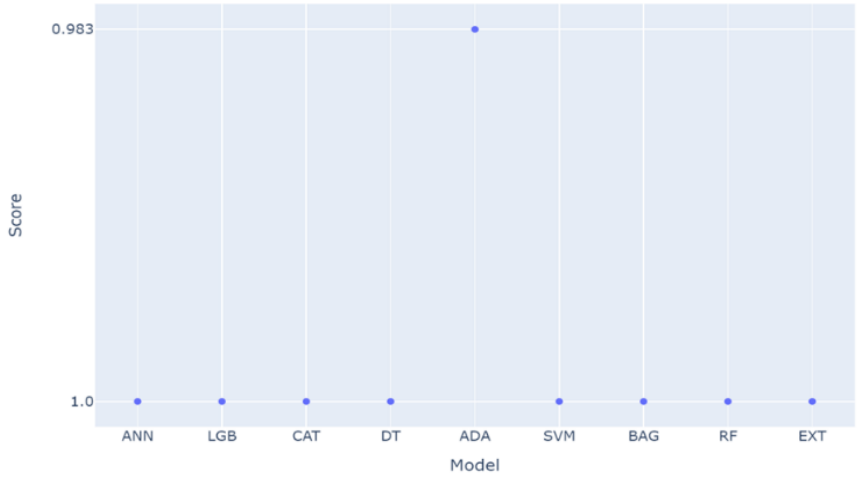

Figure 4 - Covid Overall Analysis Result

The next step involved cross-validating every model with the balanced accuracy scores (also known as macro average arithmetic)—which were determined by taking the average of the correct hits for every class. The Ada Boost model was outperformed by all of the techniques (Khaiser, F. K., Saad, A., & Mason, C., 2023).

All things considered, the outcomes produced both before and during the Covid can be combined in a table as:

Table 1. Algorithms' Performance based on pre-Covid and Covid data.

		Algorithms' Performance Based on Pre-Covid & Covid data		
S.No.	Algorithm	Pre-Covid - F1-Score	Covid - F1-Score	Similarity
1	ANN	1.0	1.0	YES
2	LGB	1.0	1.0	YES
3	CAT	0.982	1.0	SOMEWHAT
4	DT	0.982	1.0	SOMEWHAT
5	ADA	0.632	0.983	NO
6	SVM	0.948	1.0	SOMEWHAT
7	BAG	0.972	1.0	SOMEWHAT
8	RF	1.0	1.0	YES
9	EXT	0.982	1.0	SOMEWHAT

ANN	Artificial Neural Networks
LGB	Light Gradient Boost
CAT	Categorical Boost Classifier
DT	Decision Trees
ADA	Adaptive Boost Classifier
SVM	Support Vector Machine
BAG	Bagging Classifier
RF	Random Forest
EXT	Extreme Gradient Boost

Knowing about Recursive Features the discussion's primary goal is elimination through the application of various machine-learning algorithms, and it is based on the following concepts:

1. The value of feature selection using RFE

Eliminating superfluous or redundant predictors is required to choose the optimal statistical or machine learning model. These predictors degrade the performance of the model and raise the standard error of the predicted regression coefficients. It is possible that over-fitting occurs in simple statistical models, such as linear regression, when appropriate model optimization techniques, including feature selection and parameter adjustment, are neglected. According to Krzywinski et al. (2015), poor prediction capacity happens when the number of predictors is close to the sample size and the model is fitted to noise.

2. The choice of machine learning predictive models

Making the most use of machine learning-related classification approaches was essential due to the multiclass classification complexity in this study. The models that yielded the best predictions for the Pre-Covid and Covid institutional datasets were: lightGBM, AdaBoost, Random Forests, Decision Trees, Artificial Neural Networks (ANN), Keras-constructed neural networks, SVMs, AdaBoost, Cat Boost, and Extremely Random Trees.

3. Crucial determinants in machine learning

Using multiclass classification, only four predictors were found to be useful in predicting the dependent variable for the Pre-Covid and Covid's Institutional datasets. Predicting which independent variable will have the biggest impact on students' employability is also essential.

4. Implications of the Study for Theory and Practice

Prior to employing the chosen model to predict students' employability, this study employed a variety of machine-learning techniques to find accurate models. The next step involved using Recursive

Feature Elimination (RFE) to find relevant, important features or predictors. It offers understanding of the statistical processes and crucial elements to take into account when forecasting students' employability in postsecondary education. In order to prevent the selected model from performing too well, it is crucial to take into account pre-processing methods for feature selection, such as RFE and model accuracy verification with testing datasets as routine operations. The Institute of Higher Learning gained insights from the study's findings about what areas to prioritize when developing employability-related policies that will also help students in higher education reach higher skill levels. This was accomplished with the use of exact significant features utilized for additional research and a more accurate predictive model. The predictive model put forth by Hugo et al. L.S. (2019) is supported by our research.

5. Limitations of the Case Study

It is crucial to remember that the outcomes of employing machine learning models for prediction can differ based on the dataset's properties, including the quantity of predictors employed, and how the model's parameters are adjusted. To further support similar conclusions, more comparable studies with larger or equal datasets are needed. Because the current study was done at a single Institution of Higher Learning, its findings cannot be extended to institutions outside of Singapore. Thus, utilizing deep learning, future study may examine the underlying predictors of employability for institutions with diverse histories.

Ultimately, out of all the criteria we discussed, concentration was the most relevant and vital one. Other little or non-significant variables' influence shouldn't be entirely ignored.

Acknowledgment

In the first place, I would want to thank God, my family, and Drs. Cordelia Mason and Amna Saad, my supervisors, for their continuous support in pushing me to publish journal papers, which is really helping. I also thank my university, Universiti Kuala Lumpur, for sponsoring this work.

REFERENCES

1. Aam, Alamsyah (2022). Measuring Potential Employability of Being English Literature Graduates. doi: 10.4108/eai.14-8-2021.2317608.
2. Aniss, Moumen., El, Houcine, Bouchama., Younes, El, Bouzekri, El, Idirissi (2020). Data mining techniques for employability: Systematic literature review. doi:10.1109/ICECOCS50124.2020.9314555.
3. Brownlee, J. (2020). Recursive Feature Elimination (RFE) for Feature Selection in Python, https://machinelearningmastery.com/rfe-feature-selection-in-python/ .
4. Chen, R. C., Dewi, C., Huang, S. W., & Caraka, R. E. (2020). Selecting critical features for data classification based on machine learning methods. Journal of Big Data, 7(1), 52.
5. Classification: ROC Curve and AUC (n.d.). Google for Developers. https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc.

6. Héritier, Nsenge, Mpia., Lucy, Waruguru, Mburu., Simon, Nyaga, Mwendia (2023). Applying Data Mining in Graduates' Employability: A Systematic Literature Review. International Journal of Engineering Pedagogy (iJEP), doi: 10.3991/ijep.v13i2.33643.
7. Hugo L.S. (2019). Predicting Employment through Machine Learning. https://www.naceweb.org/career-development/trends-and-predictions/predicting-employment-through-machine-learning/.
8. Jeon, H., & Oh, S. (2020). Hybrid-Recursive Feature Elimination for Efficient Feature Selection. Applied Sciences, 10(9), 3211. MDPI AG. Retrieved from http://dx.doi.org/10.3390/app10093211.
9. Khaiser F. K., Saad, A. & Mason, C. (2023), "Systematic Review of Qualitative and Quantitative Studies on Perceived Employability of Graduates," 2023 17th International Conference on Ubiquitous Information Management and Communication (IMCOM), Seoul, Korea, Republic of, pp. 1-8, Doi: 10.1109/IMCOM56909.2023.10035634, IEEE.
10. Khaiser, F. K., Saad, A., & Mason, C. (2021). Framework for Future Employment Prospects of Undergraduates in Singapore through Multi-class Classification Prediction using Machine Learning Algorithms. ICare ’21 – Government Universities Linked International Conference on Advancing and Redesigning Education – Thriving in Times of Global Change, Kuala Lumpur Malaysia., 9789672880059. http://press.utp.edu.my/index.php/icare21/?utm_medium=poster
11. Krzywinski, M, Altman N. (2015). Multiple linear regression. Nat Methods.;12(12): 1103–1104. pmid:26962577.
12. Norfarahzatul, Asikin, Zakari., Mohamad, Zuber, Abd, Majid., Sheerad, Sahid. (2022). Systematic Literature Reviews of Marketability and Employability of Graduates. International Journal of Academic Research in Economics and Management Sciences, doi: 10.6007/ijarems/v11-i1/12278.
13. Renato, R., Maaliw., Karen, Anne, Quing., Ace, C., Lagman., Bernard, Ugalde., Melvin, A., Ballera., Michael, Angelo, D., Ligayo. (2022). Employability Prediction of Engineering Graduates Using Ensemble Classification Modeling. doi: 10.1109/CCWC54503.2022.9720783.