UTILIZING MACHINE LEARNING TECHNIQUES IN PREDICTING JOB VIABILITY OF INFORMATION TECHNOLOGY PROGRAM GRADUATES

by

Caesar Jude Clemente

A Research Paper Submitted to the School of Computing Faculty of
Middle Georgia State University in
Partial Fulfillment of the Requirements for the Degree

DOCTOR OF SCIENCE IN INFORMATION TECHNOLOGY
MACON, GEORGIA
2023

Utilizing Machine Learning Techniques in Predicting Job Viability of Information Technology Program Graduates

Caesar Jude Clemente, *Middle Georgia State University, Caesar.clemente@mga.edu*

Abstract

Having a job immediately after graduation is the dream of every IT graduate. However, not everyone can achieve this outcome. The study's primary goal is to develop predictive models to forecast IT graduates' chances of finding a job based on factors such as academic performance, socioeconomic status, academic habits, and demographic data. Furthermore, the paper also seeks to identify the most influential predictors of the models. Ensemble machine learning algorithms such as bagging, boosting, and voting were utilized to develop the models, and an evolutionary optimization technique was used to identify the most relevant attributes. The results reflected the voting ensemble as the model achieving the highest accuracy (88.29%) followed by random forest (82.28%). The optimizer algorithm identified job placement, IT experience, degree, high school, final and last semester GPA, IT project research, study frequency, mother's educational level, sibling number, and living accommodation as the most influential predictors. Random forest also ranked first in the optimized models by garnering an 84.27 accuracy rating. The research results will greatly benefit educational institutions, school administrators, and educators by giving them a deeper insight into their job placement programs. IT graduating students can also use the research output to improve their job placement chances.

Keywords: machine learning, ensemble algorithms, bagging, random forest, voting, XGBoost

Introduction

Most educational institutions already have programs to aid students in their passage to the next phase of their life. Some conduct seminars to teach students how to develop better resumes or ace interviews (Guyon E.,2019). Some conduct job fairs where they invite their industry partners to talk to their students (Lee et al., 2019). During these events, students are considered prospective applicants; hence are given a chance to submit their resumes and, at times, be interviewed by prospective employers. Still, although not prevalent, some give job placement exams to evaluate students' job readiness (Clemente & Kwak, 2022). Regardless of the method, one thing is clear; the goal is to market the students and advocate for their job placement. It is in an educational institution's best interest to have a strong job placement program. However, this goal is not always achieved, as some graduates find themselves jobless months after getting their diploma. For instance, data from the statistical office of the European Union reports that the employment rate for students was only 83.4% in 2019 (Guo et al., 2020). While this statistic may be considered high, still an improvement is needed to decrease the remaining 16.6%. IT student graduates are similarly situated. Reports state that several IT graduates do not possess the industry's required skills and, because of this, find it challenging to find a job (Samantha & Poojah, 2020). Another study stated that 10.3% of computer students failed to get a job six months after graduation (Smith et al., 2018). Based on these reports, it is, therefore, in the best interest of educational institutions to forecast the job viability of their students. After all, having a high percentage of graduating students hired by the IT industry within a time frame close to their graduation date is an essential metric of a university’s success (Olayniyi & Aogi, 2022). Such an outcome can be seen as a reflection of the strength of the programs and the trust of their industry partners.

Purpose

Predicting student employability is not something new. It has been done using varying techniques (Mezhoudi et al., 2021). This study seeks to contribute to this research tradition by utilizing ensemble machine learning algorithms to forecast students' employability based on academic and non-academic factors. These factors are divided into four categories: demographics, academic results, socio-economic and student's academic habits, and inclinations. The study will also identify the most influential predictors from the models.

Research questions

Consistent with the purpose of the study, the following research questions are asked:

RQ1. What ensemble machine learning algorithms can be utilized to predict the job employment of IT graduating students based on demographic, socio-economic, academic performance, and academic experiences?

RQ2. What predictors are the most influential for each model?

The utilization of a wide array of attributes provided the research with a more comprehensive analysis of the factors affecting students' job placement. The decision to apply ensemble machine-learning techniques was motivated by the desire to develop better predictive models, as this approach is considered to be superior to traditional machine-learning strategy. (Nzuva & Nderu, 2019). The output of the research is expected to provide the following benefits.

1. To provide educational institutions with relevant insights on predicting the employability of IT students so that they can be used as input to improve IT programs.
2. To serve as an early warning system for students who are predicted to have difficulties finding a job, thus giving them a chance to improve.
3. To provide advisers, academic administrators, and instructors a means to have a more profound understanding of the student's performance, allowing them to design intervention techniques to help the students.

The research document is organized as follows. The paper starts with the introduction, which incorporates discourse on the problem statement, the purpose of the study, and research questions. After this, the review of related literature is discussed, followed by the proposed research methodology. The research methodology section includes a narrative regarding the phases of the research, the instrument used, and an explanation of how data was collected, analyzed, and turned into a dataset for the models. The methodology also contains a description of the machine learning algorithms utilized and the validation and optimization techniques used to improve the viability of the results. The presentation of the results and the discussion of findings follows next, and finally, the conclusion and recommendations.

Review of the Literature

Machine Learning Algorithms

At the core of every program is an algorithm. Computer programmers define algorithms as a series of steps that performs a specific task (Snyder, 2022). Algorithms that are converted to a set of instructions in programming code are how computers fulfill the objective of a task. This process is why we also describe computational thinking as algorithmic (Guler, 2021). On the other hand, we humans can do a task without specifying an algorithm. For example, we can easily recognize persons in a photograph or predict the weather by looking at the sky. For humans, arriving at a conclusion based on given parameters is seen as almost instantaneous. The question is, can we also train machines to make decisions like humans? The answer to this question is the central underpinning of machine learning (ML). Machine learning is a computer science discipline that aims to teach computers the ability to learn without constant programming intervention (Mahesh, 2018).

The history of machine learning is closely intertwined with artificial intelligence due to the connection of machine learning to improving the "intelligence" of machines (Choi et al., 2020). Artificial intelligence is defined as the capability of computers to simulate human intelligence in doing specific tasks (Volkmar et al., 2022). Based on this definition, ML is, therefore, an application of artificial intelligence (Choi et al., 2020).

Machine Learning Categories

The research literature divides ML into supervised, unsupervised, semi-supervised, and reinforcement learning. The following is a discussion of each category.

Supervised Learning**:** Supervised learning utilizes labeled datasets to train ML algorithms. ML, under supervised learning, uses training data to detect data patterns and relationships and outputs a classification or prediction label. The trained model is then presented with test data to evaluate the model's accuracy (Sarker, 2021). Supervised Machine learning can do classification and regression tasks (Sarker, 2021). An example of a classification task is classifying spam and not spam emails, and an example of a regression task is predicting an employee's salary.

A standard supervised ML algorithm is a decision tree. Decision trees present options and results in a tree format. Decision trees are composed of decision nodes where the data splitting happens and leaves, which serve as the outcome or decision. Decision trees can be used to predict true or false questions and continuous data types (Mahesh, 2018).

Unsupervised Learning**:** While supervised learning focuses on labeled datasets, unsupervised learning deals with data that are neither labeled nor classified (Dridi, 2022). Unsupervised learning algorithms will try to find the structure according to the similarities and differences of the data. In other words, unsupervised learning lets the data speak for itself. An example of unsupervised learning is the K-means algorithm. The K-means algorithm aims to partition the dataset into distinct none overlapping clusters. The data points inside the cluster belong to only one group. The less variation within each cluster, the more similar the data points are (Dridi, 2022).

Semi-supervised**:** Semi-supervised algorithms are considered a hybrid as they combine the features of supervised and unsupervised learning to develop a prediction model (Sarker, 2021). Semi-supervised operates on labeled and unlabeled data. Similar to supervised and unsupervised, algorithms belonging to this category can be used in regression and classification problems (Mahesh, 2018). An example of a semi-supervised algorithm is a self-training classifier. A self-classifier will work with both labeled and unlabeled data. First, the self-classifier will be trained with the portion of the labeled dataset. Once this is done, the unlabeled data is then fed to the model. The predicted labels based on the unlabeled points are then added to the training set. The process is repeated until all unlabeled data are integrated (Mahesh, 2018).

Reinforcement Learning (RL)**:** Reinforcement learning is centered on the theory of decision-making. The learning is based on an optimal behavior to acquire maximum reward (Sarker, 2021). The behavior is calibrated based on the interactions and observations in the environment. The RL algorithm must discover the series of actions that will bring the maximum reward through trial and error (Liu et al., 2020). The decision taken by the algorithm is measured based on an immediate and delayed reward system (Liu et al., 2020).

Ensemble Machine Learning Algorithms

Ensemble machine learning techniques combine ML models to arrive at a better predictive model (Wen & Hughes, 2020). Ensemble learning methods are classified into three categories: bagging, stacking, and boosting. The following is a discussion of each category. The voting ensemble has similarities with bagging and stacking.

Bagging**:** Bagging is also known as bootstrap aggregation. The name refers to the two steps performed in this technique: bootstrapping and aggregation (Odegua, 2019). Bootstrapping is the part where a random sampling method is applied to the dataset using the replacement procedure. The replacement means that if data is chosen, it will be returned to the training dataset for possible reselection. Data, therefore, can be selected multiple times (Khan et al., 2019). The data is then fed to base learners, called ensemble members. The predictions generated by these members are aggregated using voting or averaging. The expectation is that the result will be more accurate and reduce the variance significantly (Wen & Hughes, 2020).

One of the popular bagging ML algorithms frequently used in research is the random forest classifier. Random forest first creates multiple decision trees, serving as the "forest" (Ali et al., 2012). Each tree is created using the bootstrapping sampling method described in the previous paragraph. The algorithm then selects the best feature from a random subset of features generated. Based on this random subset of features, the algorithm will utilize these variables to split each node. The overall process results in significant variance reduction, enhanced accuracy, and a more stable model (Ali et al., 2012).

Boosting: If bagging thrives by combining weak learners all at once to develop a more stable model, boosting techniques also aims to achieve this outcome by making model predictors learn from each other mistakes (Odegua, 2019). The weak base learners are arranged in sequential order and processed one after another. First, the boosting algorithm allocates the same weights to the data sample. It will then generate the first predictive model which will become the first base learner Data is fed to the base learner for the initial predictions. The algorithm will then assess the prediction results and assign weights based on the model's performance (Nzuva & Nderu, 2019). These weights are then passed to the next learner and the process is again executed until errors are below the specified threshold.

Adaptive boosting or AdaBoost is an example of a boosting algorithm. AdaBoost implementation initially assigns equal weights to the datasets and adjusts the weight in each boosting iteration (Chengsheng et al., 2017). Like what was described in the previous paragraph, it assigns more weight to classification errors and corrects them in the next iteration The goal is to reach a point in which the residual error is within the acceptable threshold (Nzuva & Nderu, 2019). AdaBoost is one of the earliest boosting algorithms developed and its ability to adapt and self-correct makes it a popular ML option.

Stacking: The learners in bagging and boosting are usually homogeneous that is all the models were developed based on the same ML. In stacking, however, different models are executed in parallel, and the results are combined by a meta-learner (Wen & Hughes, 2020).

This characteristic of stacking allows it to capitalize on the strengths of the different ML models. Each model is expected to learn a part of the problem. The final model called the meta learner is stacked above the other base learners. The meta-learner takes the inputs of the sub-models and learns what is the most optimal way to combine the predictions of the base learners (Odegua, 2019).

Voting**:** Voting is a type of ensemble machine-learning technique that trains various base models similar to stacking and then aggregates the predictions of each base learner (Jindal et al., 2022). The goal is to improve the final model performance. The base models will vote based on two methods: hard or soft voting. Hard voting utilizes the highest number of votes from the models, while soft voting uses the largest sum of probabilities from the participating models (Manconi et al., 2022). Voting can also be used in regression problems by computing the average of the contributing models (Jindal et al., 2022).

Voting and bagging trained their base models in parallel. Unlike bagging, however, voting can combine different types of learners. Both voting and stacking aggregate their predictions as part of the last step; however, in voting, user-specified weights are used to combine the learners, while in stacking, the aggregation is done by another learner.

Previous Studies on Student Job Placement Prediction

The topic of predicting student job placement is not something new. There is a plethora of studies conducted for this purpose. In one paper, the author collected data from 515 information technology (IT) students (Piad, 2018). Data collected were demographic profile (gender, age, and location) and academic performance such as cumulative weighted average. The author used five algorithms for model generation. These classifiers were Naïve Bayes, J48, SimpleCart, Logistic regression, and Chaid. The results showed Chaid got 76.3, which was the highest accuracy rate among the classifiers. The study also identified the three most dominant factors that have a direct impact on IT employability. These factors are IT core subjects, IT professional subjects, and gender.

In another study, the authors highlighted the significant challenges students face in finding a job after graduation and aimed to develop a predictive framework to help students in their job search(Guo et al., 2019). The study incorporated in their model the different employment biases to further enhanced their output. There were 2,133 participants in the dataset, and all graduated from a Chinese University in 2017. Students in the set were from 64 different majors. The dataset captured demographic data such as hometown, gender, and nation, academic performance data such as scores, and credit and employment data such as employment status and company information. The authors then identified the variables for the bias analysis, which are gender, nationality, administrative level of hometown, and enrollment status. The authors relied on neural network algorithms, particularly long short-term memory (LSTM) variants, to constructively improve their model. They initially started with LSTM, which resulted in 86 percent accuracy but produced a low F1 score of 46 percent. Incorporating other techniques such as dropout, generative adversarial networks (GAN), and new loss with LSTM improved the model by increasing the accuracy to 88% and the F1 score to 81 percent. The author also emphasized that the final model outperforms baseline models such as LSTM and XGboost. The study did not identify the most influential predictor among the factors.

In Rao et al. (2019), the authors focused on real-time university data from four major engineering disciplines. Data included are academic performance, extracurricular activities, internships, and students who took massive online courses (MOOC). As additional input, the authors consulted companies to determine the essential skills students need to get placed. In addition, the authors also gave weights to each of their chosen features. For example, internship, extracurricular details, and MOOC were given 20% each, while academic performance was awarded 40%. The output of the model does not just predict if the student would be placed; it would also determine which company category the student will have more chances to find a job. The authors categorized the companies into best, good, and average. The model also advised on which areas the students need to improve. The authors implemented support vector machines (SVM), K-nearest neighbor (KNN), and artificial neural networks (ANN). The results showed that ANN achieved 99.02% using Tanh activation. ANN got the highest result, with SVM ranking the lowest (95.12%). ANN also scored the highest based on precision, F1, and sensitivity analysis.

Aside from developing a predictive model for Master of Business Administration (MBA) students' employment, one study scrutinized the association of demographics on the student's job placement (Kumar et al., 2021). The authors also examined the impact of gender and MBA specialization on salary, the association between gender and MBA specialization and placement percentage, and the correlation between degree stream and placement status. Finally, they developed models to predict job placement based on significant features. The authors used an existing dataset from Kaggle. Features they utilized came from demographic data such as gender, academic performance results, placement test results, salary offer, degree specialization, and streams. For developing the predictive model, the authors used support vector machine (SVM), random forest (RF), extreme gradient boosting (XGB), gradient boosting (GB), and logistic regression (LR).

The results demonstrated that there was no bias on gender in terms of the salary offered to students. There was also no impact on salary with regard to MBA specialization. The same finding was observed on the impact of gender on placement exam scores. The MBA streams were also found to be statistically insignificant concerning placement status. The only experiment that was found to be statistically significant was the MBA specialization about placement status. Regarding the predictive model, SVM achieved the highest accuracy rate of 90%. GB got only 80%, ranking as the lowest accuracy rate. The authors also extracted the most essential features of the model. The variables, work experience, and SSC test results were found to be influential in getting a job. In addition to the models predicting job placement, the authors also produced a model to predict the gender of a placed student. Based on the result, the random forest got the highest accuracy rate, achieving an 88 percent accuracy in predicting the gender of the placed student.

Katkar et al. (2019) used board examination data as data fuel for their predictive models of job placement. The classifiers they employed were decision trees, random forest, and multilayer perceptron. The innovation that the authors suggested in this research is to conduct the prediction in the first year of engineering. According to the authors, this would give more time for students to improve their skills. The results demonstrated multilayer perceptron achieved the highest accuracy rating, 76.06%, based on the 10th and 12th state board exams. Decision tree attained an 80% accuracy for job placement based on the 10th and 12th CBSE exams.

Like the previous study, Huynh et al. (2020) used neural networks in their experiment on student job prediction. The dataset came from another research and consisted of data from online finding job sites. Four deep neural network models were utilized by the authors, namely, TextCNN, Bi-GRU-LSTM-CNN, and Bi-GRU-CNN. The authors combined the four models, forming an ensemble set up with majority voting as the mechanism to increase the efficiency of the final predictive model. The authors also mentioned that they allocated 10% of the dataset to testing, 20% to validation, and 70% to training. The metrics used to measure the model were accuracy, precision, recall, and F1-score. In presenting the results, the authors showed the individual model metrics before the ensemble. The findings revealed that the ensemble method outperformed the individual models, gaining 72.70 accuracy, 72.83 precision, 72.59 recall, and 72.71 F1-

measure. The authors also emphasized that their model can be used to analyze applicants' resumes and cover letters and predict applicants' job viability.

Harihar and Bhalke (2020) mainly used academic performance features to develop their models. They utilized classifier algorithms for their research, such as Naïve Bayes (NB), Multilayer Perceptron (MLP), logistic model tree (LMT), minimal sequential optimization (SMO), and simple logistic and logistic classifiers. The data set comprised 1000 records collected from the placement data of a college. The researchers also used tenfold cross-validation to generate the model.

The results showed MLP, SMO, simple logistics, and LMT performing very well by gaining more than 95%. LMT was shown to have the maximum accuracy ranking, gaining a score of 99.5%. The study did not provide additional information on the characteristics of the participants in the dataset.

In a recent study by Muraina et al. (2022), the authors used the graduating students' GPA to predict student employability. The machine learning algorithms utilized were decision tree, support vector machine, and K-nearest neighbor. The authors found that the decision tree has the highest accuracy, precision, recall, and F-measure, garnering scores of 89, 90, 89, and 89 percent. Support vector machine got the lowest accuracy, precision, recall, and F-measure, getting only 45, 30, 45, and 35 percent. The study did not identify which factor is the most influential predictor.

The proposed study will be differentiated from the other research due to the following factors. First, it will concentrate only on IT graduates. Second, it will employ a wide array of factors that has the potential to affect IT job employability. These factors include academic performance, socioeconomic, academic experiences, and demographic data. Third, the study will utilize ensemble machine-learning techniques to generate the models. Fourth, aside from presenting the model results, the research will also determine the most influential predictors, and lastly, the study has an actual recipient, in this instance, the university where the participants graduated. The result of this study will be highly significant to the university's IT department.

Research Methodology

The research was divided into three phases. Phase one was data gathering. An instrument was developed and distributed to get the data from participants. Phase two was centered on data preparation. The collected information was prepared for model generation. Phase three is the model building and presentation of results. Details for each phase are discussed in the succeeding paragraph.

The Instrument

The questions in the survey were designed to fulfill the data requirements mandated by the research questions. While the researcher developed the survey questions to satisfy the research objectives, the survey items' general concept was partially based on the variables of other studies that predicted students' academic performance (Yılmaz & Sekeroglu, 2019). In the current setting, however, job-related and IT degree questions were added. Furthermore, the prediction targeted was job placement, not academic performance. There are four data categories: demographic, socio-economic, academic performance, and academic experiences. Demographic data includes gender, type of degree, concentration, semester of graduation, age range, job placement recipient, previous IT experience and ethnicity. Next would be the academic performance data comprising of high school GPA, GPA of the last semester before graduation, GPA upon graduation, and coding grade. Academic experience includes scholarship status, attendance of IT seminars, the relevance of IT projects or research, internship completion, studying frequency, and class attendance.

Socioeconomic factors include the romantic partner's presence, accommodation type, transportation mode, financial support, mother and father's educational attainment, the number of siblings, and the parent's marital status.

Lastly, the questionnaire asked essential questions about job placement, such as when they found a job after graduation. In answering some questions, the participants were instructed to limit the start of the time frame to the last year before graduation. For example, in answering the question if they have a romantic partner while studying, the time frame would be if they had a partner within the year before graduation. The instrument was constructed in survey monkey and distributed to the participants through email.

Target Subjects

As stated in the research question, the research examined various factors and their impact on job placement. The target participants of the study will be graduates with an IT degree from institutions of higher education in the United States and Canada. The scope will cover both bachelors and post-graduate degrees. Answering the survey will be anonymous. No name will be attached to the questionnaire.

Data Preparation

First, the data was collated. In preparation for model generation, the data was cleansed. Data cleaning involved handling missing values and removing duplications. The items were also tested for correlations. If there are columns with high correlations with other columns, such features will be removed. The threshold for the correlation coefficient will be a coefficient greater than .9. A more detailed discussion about the formation of the final dataset can be found in the results section.

Model Generation

Metrics

Classifier metrics are essential to determine the ability of the model to predict and serve as criteria for selecting the best among a variety of models (Hossin & Sulaiman, 2019). The following metrics were utilized to evaluate the generated models of the study.

Confusion Matrix- The confusion matrix is a table that visualizes the model predictions against the actual values (Kulkarni et al., 2020). In other words, it is the summary of the performance of the classifier model. The following is an example of a confusion matrix.

Table 1: *Confusion Matrix*

	Predicted class	
Positive	True positive (TP)	False negative (FP)
Negative	True negative (TN)	True negative (TN)

True positives and negatives are the quantity of items classified correctly by the classifier as positive or negative. False positives and negatives are the number of items classified incorrectly. The following are the summary metrics that can be derived from the confusion matrix.

Accuracy- This metric refers to the ratio of the correct predictions over the total number of items in the dataset. Below is a formula to compute the accuracy (Kulkarni et al., 2020).

$$TN + TP/ TP + FP + TN + FN$$

Precision- The ratio of correctly classified items to the total of predicted items in the positive class (Hossin & Sulaiman, 2019). The formula for computing precision is the following:

$$TP/ TP + FP$$

Recall- The proportion of correctly identified positive cases over the actual positive items. The following is the formula for computing recall (Hossin & Sulaiman, 2019).

$$TP/ TP + TN$$

F1 Score-F1 score incorporates precision and recall by computing their harmonic mean (Hossin & Sulaiman, 2019).

$$\text{F1 score} = 2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})$$

Machine Learning Algorithms

As stated in the research question, the study aimed to utilize ensemble machine learning algorithms to produce the models. The following were the selected algorithms per ensemble category.

Bagging

Random forest (RF)**:** To understand the random forest, we must begin by discussing decision trees. Decision trees graphically represent the choices and their possible consequences (Sarica et al., 2017). Each tree node represents the conditions that allow the tree to split into branches. The end of the branch is called a leaf representing a prediction or classification. Decision trees are trained using the classifications and regression tree (CART) algorithm. The quality of splits in decision trees is measured by using metrics such as Gini impurity, information gain, and means square of error (MSE) (Tangirala, 2020). Decision trees can be susceptible to bias and overfitting. However, their output can be significantly improved when combining multiple decision trees (Zhao et al., 2021). This combination of more than one decision tree is what gives the random forest its major advantage. Random forest uses an ensemble of decision trees to train the data. Random forest utilizes the bagging method. Unlike decision trees which consider all the possible node splits, RF selects only a subset of the variables. The selection of these features enables RF to generate uncorrelated decision trees (Wang et al., 2020). Overall, RF improves precisions and reduces the chances of overfitting.

Boosting

Gradient Boosting Trees (GBT): Unlike random forests where the decision trees are generated at the same time, gradient-boosted trees are built sequentially. The succeeding trees are added to improve the predictive power of the previous trees. This method is precisely what boosting is all about. Every new learner fits into the residual of the previous model. The residuals are measured through a loss function such as logarithmic loss for classification tasks. The goal of gradient boosting is minimizing the gradient of the loss function; thus, the added learner's target outcome is to minimize the gradient of error. The final model combines the results, generating a strong learner (Hagiwara et al., 2022).

XGBoost: XGboost is based on the gradient boosting tree technique. The difference is that instead of trees being built in sequence, trees are built in parallel. XGboost was developed to make models faster and improve their performance. In gradient-boosted trees, the stopping criterion is based on the negative loss at the point of the split. XGBoost, on the other hand, uses the maximum depth parameter as its stopping criterion. The algorithm also has a built-in cross-validation method which makes the specifications of the exact number of iterations automatic (Chen & Guestrin, 2016).

LightGBM: LighGBM also belongs to the gradient-boosting ensemble category based on decision trees. LightGBM generates decision trees that grow leaf-wise. This process limits the tree depth, which in turn helps prevent overfitting. The leaf-wise growth allows the tree to grow vertically instead of the usual horizontal process. LightGBM also uses the histogram-based technique, where data is divided into bins. The bins are the ones that are used to compute the gain and split the data. LightGBM is also capable of feature bundling, which reduces data dimensionality. As for sampling, LightGBM uses the gradient-based one-side sampling (GOSS) method. This method gives more weight to data points that have bigger gradients. Some of the data points with small gradients are then randomly removed (Ke et al.,2017)

Voting

Literature advises that voting should have diverse base learners to capitalize on their advantages (Van Rijn et al., 2017). Given these relevant recommendations, the base learners of the ensemble would be K-nearest neighbor (KNN), Naïve-Bayes, Random Forest, Support Vector Machines, and Deep learning. It will also include the algorithms XGboost and LightGBM.

K-nearest neighbor (KNN): KNN algorithm is based on the idea that similar things are usually found near each other (Uddin et al., 2022). KNN uses a majority voting mechanism to assign a class label in classification problems. The computation of the distance is required before a classification is made. The usual metric for measuring distance in KNN is the Euclidean distance. KNN also belongs to the ML category of "lazy learning" models as it does not undergo a training stage. All calculations are done when a prediction is generated. KNN is one of the widely used ML due to its simplicity and adaptability (Uddin et al., 2022).

Naïve-Bayes: Naïve Bayes assumes that the presence of a characteristic of a class is unrelated to other characteristics of that class (Chen et al., 2021). In real life, however, an object's multiple characteristics contribute to that object's identity. The naïve-Bayes assumption that characteristics do not correlate with each other makes the algorithm "naïve." Naïve Bayes belongs to the category of a probabilistic algorithm as it calculates the probability of a feature based on prior knowledge related to that feature.
(Chen et al., 2021)

Deep Learning*:* Deep learning comes from neural network algorithms. Neural networks mimic the human brain in terms of artificial neurons called nodes (Alzubaidi et al., 2021). These node layers contain an input layer, one or more hidden layers, and an output layer. Deep learning is a neural network with three or more hidden layers. This algorithm transforms its inputs in a nonlinear manner and creates a statistic model as an output. The output represents what it learned from the previous iteration. This process continues until the output reaches an acceptable threshold of accuracy. The number of processing layers the data must undergo makes the algorithm "deep" (Alzubaidi et al., 2021).

SVM: Support vector machines come from the family of kernel-based algorithms. SVM divides data into different categories by finding the hyperplane, which is a line that separates the data. The distance between the data points and the hyperplane is called the margin. The algorithm will maximize the distance between the classes to increase the margin; thus, if such maximization is achieved, the probability of correct classification increases (Jun, 2021).

***Logistic Regression*:** Originating from the field of statistics and incorporated into machine learning algorithms, logistic regression predicts the probability of the event's occurrence. The prediction serves as the dependent variable based on independent variables. The output of the dependent variable is exemplified as 0 or 1. The computation is a logit transformation that is applied to the probability of success divided by the probability of failure (Boateng & Abaye, 2019).

Evaluation and Optimization Techniques

The study used K-10-fold cross-validation to evaluate the predictive ability of the generated models. The 10-fold was chosen due to its proven better results in previous studies (Nti et al., 2021). Under this technique, the dataset is divided into a K number of folds. The K, in this case, is the number of the subset of data from the dataset (Berrar, 2019). For example, if there are ten groups, the model is trained on nine groups and tested on the remaining group. The process is repeated until each subset serves as a validation set. This technique differs from the typical hold-out method, where the dataset is divided into two parts, the training and validation set. The advantage of k-fold validation is that it reduces bias and overfitting (Nti et al., 2021).

For the study feature selection, the research utilized an evolutionary operator belonging to the genetic algorithm category for feature selection and optimization. This algorithm mimics the process of natural evolution utilizing techniques such as mutation, selection, crossover, and inheritance (Schulte et al., 2021). Under this technique, an initial population first is generated and switched on with a probability metric. Different processes are then applied, such as enacting the mutation function, performing crossover, selecting, and mapping individuals according to fitness, and randomly drawing individuals based on probability (Gmbh, 2023). The researcher used RapidMiner as a tool to generate the models (RapidMiner Amplify the Impact of Your People, Expertise & Data, 2022).

Results and Discussion

The Dataset

A total of 834 participants initially filled up the survey. Twenty participants were not included in the final dataset due to partial responses. As stated in the previous paragraph, the survey's target audience was IT graduates with an IT degree. 88.32 percent of the dataset population came from the United States, and the remaining 11.68 percent are from Canada. The binomial classification for the job placement prediction utilized six months as the threshold for determining the prediction label of the participant. Graduates who found a job within the first six months after graduation were grouped into the "yes" class, and those who did not find an IT job yet or found a job after six months were grouped into the "no" class. The six months cut-off was based on research conducted by the University of Washington that states the average time for a college graduate to find their first employment is six months (Apalla, 2022). Applying the six months limit, 553 participants found a job within six months after graduation, and 261 participants failed to do so. According to the survey's demographic-related questions, 60.8 percent of participants got a bachelor or associate degree, while 39.2 percent possess a post-graduate degree. Forty concentrations were reported. The top two were the generic degrees in IT and computer science. These findings were followed by cybersecurity, software engineering, networking, and data analytics. The bottom list includes specialization in microcomputer technology, technical management, and blockchain, garnering only one each. Most participants reported graduating in spring (36%), and the winter graduates (8.7%) got the lowest score. 56.4 percent of the participants were male, 43 percent were female, and 6 percent chose the prefer not to say option. For ethnicity, 62.2 percent were Caucasians, 15.7 percent identified as African American, 11.8 percent as Asians, and 4.7 percent as Hispanic or Latino.

Moving on to the academic performance category, most participants reported having a high school GPA between 3.0 to 3.49 (39.9%), followed by 38.3 percent of participants belonging to the above 3.49 category. Six percent 6% of students had a high school GPA below 2.0. The GPA of the last semester before graduation reflected a higher trend, with more participants reporting a more than 3.49 GPA range (46.8%). It was seconded by those in the 3.0 to 3.49 range (36.2%). Only 4% belonged to the below 2.0 category. The GPA after graduation followed the same trend as the last semester before graduation GPA with 46.3 percent from the more than the 3.49 category and 37.3 percent from the 3.0 to 3.49 range. Only 2 participants reported having a GPA below 2.0. For the coding score grade range, more than fifty percent (50.7) reported having their programming grade belonging to the A level. The data also showed 37.2 percent on the B and 9.8 on the C level. 1.8 percent belonged to the D category, and only 3 participants selected the F level.

The next round of questions pertains to academic experience and habits. 54.7 percent stated they received some form of scholarship while studying. 38.7 percent reported studying daily, while only 2.8 percent declared studying only during the exam period. The largest portion of the participants (58%) attended IT seminars to improve their IT skills while studying. The next question was class attendance, 52.3 percent reported always attending classes, and only 1.5 percent picked the "rarely" option. Six hundred sixteen participants overwhelmingly stated that the IT project or research they did helped them find their job. Finally, 59.5 percent agreed that their internship experience aided them in obtaining employment.

The final set of questions relates to socioeconomic factors. 61.9 percent had romantic partners during the last year of their degree. For accommodation, 47.8 percent expressed they lived with their family while studying, 29.4 percent were renting, and 16 percent lived in a dorm. 55.8 percent reported using a private car while studying, and 21.9 percent used public transportation. Most participants (65.1%) received financial help while studying. 41.2 percent of the participants reported their mother's highest educational level is a bachelor's/associate, 28.6 percent chose a master's, and 23 percent high school. A similar movement can be observed for fathers' educational level; 33 percent reported a bachelor's, 31.4 percent master’s level, and 24.1 percent high school. Interestingly, based on the combined findings of the parent's educational level, more than fifty percent have a postsecondary degree. For the number of siblings, 34.9 percent reported having two siblings, 27.3 percent one, 15 percent three, and 9.8 percent none. Lastly, for the marriage status of the participants' parents while studying, a whopping 78.5 percent reported parents being married, 11.11 percent were divorced, 4.1 percent separated, and 2.1 percent one or both were deceased.

Based on the questionnaire, 27 variables were initially extracted. The dataset was then evaluated for multicollinearity. Multicollinearity is a phenomenon in which two or more explanatory variables are highly correlated. This high correlation might lead to inaccurate parameter estimates, and a decrease in the model's predictive power hence should be avoided (Chan et al., 2022). Based on the correlation matrix generated, no attributes were found to have a correlation higher than .9; thus, all attributes were retained. The dataset also exhibited imbalance classes, with the "yes" being the majority class (553) and the "no" the minority (261). The current research literature reports that an imbalance class should be avoided as it might affect the generalization ability of the predictive model (Fernandez et al., 2018). To improve this situation, the synthetic minority oversampling technique (SMOTE) was applied. SMOTE is a statistical technique that generates new instances from the existing minority class using an algorithm that selects an instance of the minority class and finds its K nearest neighbors (Fernandez et al., 2018). The final dataset after SMOTE application consists of 1106 examples (553 yes and 553 no), 26 attributes, and one predictive label.

The following figures show the matrix visualization of the correlations among the variables.

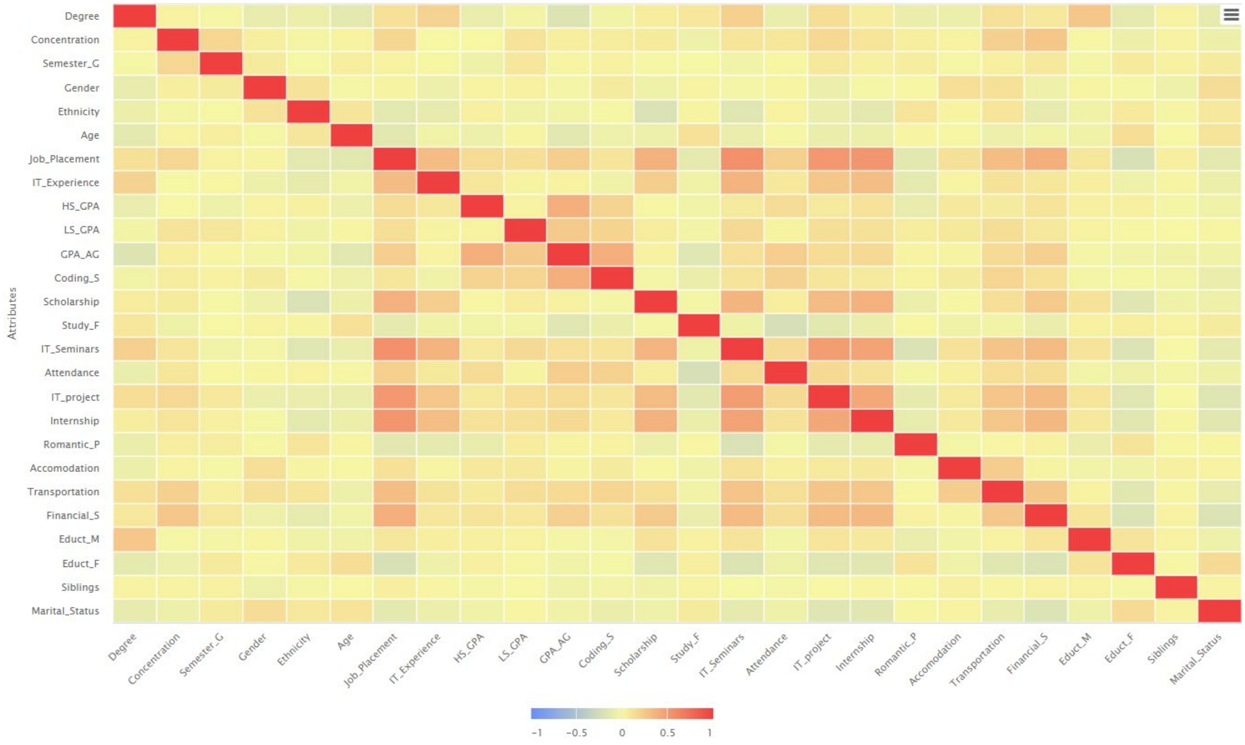

Figure 1: *Matrix Correlation Visual*

The visual correlation matrix reflects the strength of the correlations through the intensity of the color. Based on the legend, the more the color is closer to red, the higher the correlation. The figure above demonstrated that no attributes were highly correlated.

The discussion below will map the predictive models results to the research questions to prove that the experiment's outcomes satisfactorily fulfilled the research objectives.

Research Question 1: What ensemble machine learning algorithms can be utilized to predict the job employment of IT graduating students based on demographic, socio-economic, academic performance, and academic experiences?

The following table summarized the findings of the models.

Table 2: Summary Performance of Predictive Models

Ensemble Machine Learning Algorithm	Accuracy	F1
Random Forest	82.28	83.66
GBT	79.83	80.94
XGboost	79.93	80.26
LightGBM	78.40	79.38
Voting	88.29	88.50

The ML algorithms applied in the research all came from the ensemble category. The findings above demonstrated more than 75% accuracy, with the voting ensemble getting the highest score (85.59). The F score of each model also corroborates the accuracy ratings. Unsurprisingly, the voting ensemble topped the list, leveraging the "wisdom of the crowd" principle to generate an improved classification result (Luo & Liu, 2019, p. 1). Furthermore, the base learners chosen to be part of the voting ensemble are diverse according to best practices (Luo & Liu, 2022). Different models will generate different types of errors, and by combining the predictions of each model, there is a significant reduction in the overall error rate.

The random forest classifier reached second place in the accuracy metric. This result is similar to the previous research in which RF also scored the highest. (Kumar et al., 2021). The overall results of the experimentations, despite not reaching over 90, are still relevant and comparable to previous studies, some of which have achieved an accuracy score within the range of 76-88% (Paid, 2018; Guo et al., 2019; Katkar et al, 2019; Huynh et al, 2020). Furthermore, the output clearly illustrated that ensemble ML algorithms could effectively be used to develop job placement predictive models based on diverse factors.

Research Question 2: What predictors are the most influential for each model?

Most ensemble ML algorithms chosen for the research are tree-based models (Random Forest, XGboost, LightGBM). Tree-based models already utilize feature selection as the optimal feature is usually selected and used to split the data (Dubey, 2021). Nonetheless, as demonstrated in the results, the genetic optimizer algorithm managed to reduce the number of variables for each model. Feature selection is vital as it can decrease over-fitting, reduce training time, and improve accuracy (Chen et al., 2020). The selected features by the optimizer algorithm represent the most influential predictors. The following table shows the critical variables chosen by each model, organized by the survey category.

Table 3: Selected Features

	Demographic	Academic Performance	Academic Experience	Socio-economic
Random Forest	Semester Grad Age Job Placement IT Experience Gender Degree Ethnicity	HS GPA LS GPA GPA After Grad	IT Project/Research Scholarship Study Frequency IT seminars	Romantic Partner Accommodation Siblings Educt of Mother
GBT	Semester Grad Ethnicity Age Job Placement IT experience Degree Gender	HS GPA GPA After Grad LS GPA	IT project/Research Study Frequency	Accommodation Transportation Educt of Mother Siblings
XGboost	Age Ethnicity Job Placement IT Experience Degree	HS GPA GPA After Grad LS GPA	IT Project/Research Study Frequency IT Seminars	Accommodation Transportation Siblings Educt of Mother
LightGBM	Degree Gender Job Placement IT Experience	HS GPA LS GPA GPA After Grad Coding Grade	IT project/Research Study Frequency IT Seminars	Accommodation Transportation Educt of Mother Educt of Father Siblings

Under the demographic category, all the models contain the variables job placement, IT experience, and degree. The feature job placement refers to whether the graduate is a beneficiary of the job placement program of their educational institution. The tagging of job placement as one of the most influential predictors is to be expected as universities and colleges specifically designed their job placement programs to increase the job marketability value of a student. The inclusion of this feature highlights the importance of such programs in helping IT graduates land a job sooner. Previous IT experience as a critical feature also makes sense, as most employers prefer someone with an existing IT background to reduce their training costs.

The degree is also a critical factor. It is tempting to ask which level was the most influential as the research classified degrees into bachelor's and postgraduate degrees. Does having a postgraduate degree gives an IT student an edge in finding a job after graduation? Based on the result, we cannot concretely arrive at this conclusion as most of the dataset participants are at the bachelor's level. However, a cursory examination of data shows that out of 319 postgraduate students, 213 found a job within the first six months after graduation. This number is more than 50% and hence can be considered positive. The same trend applies to bachelor's level students. Out of 495, 340 found a job within the first six months. This determination is not to be construed as the sole defining factor, as we must consider the degree in combination with the other variables to make a prediction. However, there is evidence in the literature that found a positive correlation between a postgraduate degree and employability (Ali & Jalal, 2018). Some study even further posits that having a postgraduate degree improves job performance and can impact job viability (Hashmi et al., 2019).

High school GPA, last semester GPA, and GPA after graduation are considered critical factors under the academic performance category. This finding emphasizes the importance of grades for new IT graduates. Literature also corroborates that a higher GPA can help one find a job sooner (Sulastri et al., 2015). In recent research, GPA can even be used as a predictor of wages for graduates (Zou et al., 2022). Employers can use GPA as a recruitment input to evaluate competence and suitability to the position for a new IT graduate who has yet to acquire an IT skillset portfolio. Moreover, many employers consider having a good GPA as not just a reflection of academic excellence but a testament to a student's dedication and perseverance.

Another interesting point to highlight based on the results is the improvement in GPA from high school to college. Most participants belonged to the 3.0-3.49 category in high school, but during college, the trend improved, with the most significant chunk now consisting of participants with a GPA above 3.49. The connection between high school GPA and academic performance in college has been investigated and proven to have a positive correlation (Al-Asmar et al., 2021). High school GPA can be used as an indicator of success in academic achievement and job satisfaction (Al-Asmar et al., 2021). Surprisingly, the coding grade, which refers to the mark of the participants in the programming courses, can only be found in one model. This result is unanticipated as most of the first employment ventures by IT graduates are usually related to software development. However, this finding does not mean that coding proficiency is not essential. Most of the time, IT employers rely on customized technical exams as part of the recruitment process to evaluate applicants' programming skills.

The academic experience category identified study frequency and IT project/research as the most influential factors. These two features are present in all the models. The IT project/research refers to the final project an IT graduate does as part of their degree requirements. IT projects are usually seen as the culminating activity where students are supposed to apply everything they have learned. Its summative assessment value is, therefore, influential in representing the burgeoning skill of a new IT graduate. An IT project also prepares students for their transition to the industry. For example, in a paper by Adlemo (2022), he concluded that students' capstone project has a significant role in decreasing the skill gap from education to industry. Students are advised to leverage their school projects and make them part of the portfolio when they present themselves to their prospective employers. Projects are indeed an excellent way for new graduates to demonstrate that they possess the necessary skills needed for the job (Shurin et al., 2021).

Study frequency corroborates the academic performance findings. It is reasonable to expect that students who have good study habits not just get good grades but are more disciplined and reliable, aspects that a future employer would appreciate. Attending IT seminars to improve the skill set of IT students is in three models. While this feature is not present in all models, it is still relevant and worth mentioning. Joining IT seminars is not just for personal development; it can also be an activity that can expand a student's network and possibly affect their job prospects. For instance, attending research conferences can potentially connect students to job opportunities or provide strategic information that can lead to a job (Hauss, 2020).

The last category refers to the socio-economic factors. Three features are present in all models: accommodation, the mother's education level, and the number of siblings. 247 out of 553 of those who found a job after graduation declared themselves to be living with family. This finding, while a sizable portion, does not constitute the majority. Although 55% of those who were successful in their job hunt are not living with their family, we cannot conclude that this fact alone is a deciding factor for their job placement. Nevertheless, does family support impact job fulfillment? Based on the literature, evidence suggests that having good family dynamics affects job engagement (Karatepe, 2015). The keyword here is support. The current data result is inadequate to make a definitive conclusion about family support in job placement. Even though the majority of the yes category were not living with their family, it does not mean that they did not have a robust family support system. To provide more clarity on this situation, further study needs to be done.

As for the mother's highest educational attainment, the data showed that 397 out of 553 of those who were successful in their job search declared their mother's highest educational attainment is more than high school (244 bachelor's, 153 master's). This finding reflects that most of those who found a job within six months have a mother who is highly educated. The data underscores the importance of a parent's influence on a student's drive and motivation to succeed. This conclusion is supported by literature as data proves that a parent's educational attainment encourages children's academic success through example, expectations, and cognitive stimulation. (Davis-Kean et al., 2021)

Lastly, for the siblings, 200 out of 553 of the yes class declared to have a sibling of 2. This finding constitutes the most considerable portion of the yes class, followed by those who stated having a sibling of 1 (158 participants). The same trend can be seen for the no class. The data shows that the majority of the participants in both yes and no classes have at least 1 or 2 siblings. Researchers have investigated the impact of siblings on the success of a person. Research suggests that when a youth attends college, it is more likely that his or her siblings will attend too. (Smith, 2020). Siblings can be an excellent source of resources and support to help students navigate college life (Waugaman, 2022). A well-rounded student with the emotional support of their family will be more confident in their job application and thus have better chances of finding one.

Five features were not selected by any of the models. These are concentration, financial aid while studying, marital status of parents, attendance, and internship. Concentration refers to the specialization of an IT student. This finding indicates that having a generic IT or computer degree does not prevent an IT graduate from venturing into any IT niche in the industry. The internship factor was also expected to play an essential role in job placement because it is considered a good preparation for students' transition to the workplace. The importance of internships in student career preparation has been emphasized by previous research (Galbraith & Sunita, 2020). Why was internship not considered an influential factor? This omission of internship might be because the participants who answered the survey already considered this a previous IT experience. Although the questionnaire emphasized that an internship happens before graduating, some participants might have concluded that it can be classified as an official IT job.

Below is the comparison of the accuracy before and after the application of the feature optimization algorithm.

Table 4: Predictive Model Before and After Comparison

Machine Learning Algorithm	Accuracy Before Optimization	Accuracy After Optimization
Random Forest	82.28	84.27
GBT	79.83	80.38
XGboost	78.93	79.66
LightGBM	78.40	78.94

Leading the list is random forest, with an accuracy of 84.27 percent. All the models showed improvement in their accuracy levels. The above table evidently demonstrates that a subset of features can be extracted using an optimizer algorithm, and new and improved models can be generated based on the extracted features. Moreover, feature reduction helps train the model faster.

Conclusion and Recommendations

Every IT graduate wants to become economically productive as soon as they graduate, yet according to a recent report, 41% of college graduates may need help finding a job (Apalla,2022). The paper's main objectives are to develop a predictive model to forecast the job placement of IT graduates based on various factors and to identify the critical factors that significantly affect the models' outcomes. Twenty-seven questions based on four different factors, such as demographic, academic performance, academic experience, and socioeconomic, were developed and administered to IT graduates. Ensemble machine learning algorithms such as random forest, GBT, XGboost, LightGBM, and voting were utilized to produce the models. Finally, an evolutionary algorithm that can optimize selection was applied to determine the most influential predictors from the model. The results showed that the voting ensemble achieved the highest accuracy with a rate of 88.29 percent, followed by the random forest, which garnered an accuracy score of 82.28. The genetic algorithm utilized to determine the relevant features identified the features that are the most influential predictors of the model. These features are job placement, IT experience, degree, high school, Final and last semester GPA, IT project research, study frequency, mother's educational level, sibling number, and living accommodation. Finally, the models were redeveloped, but this time utilizing only the identified relevant features. All models showed an improvement from their previous accuracy rates, with the random forest attaining the highest percent (84.27)

The output of the study is expected to support educational institutions' mandate to improve students' job placement after graduation. Specifically, it will aid students in knowing their prediction of finding a job, thereby adapting possible changes to improve their chances. Academic administrators, instructors, curriculum designers, and advisers can also use the model to gain a more profound insight into the type of students they have and institute the necessary reforms to their job placement programs. Finally, for educational institutions, the model can serve as a litmus test of the effectiveness of their program offerings, thereby giving them an idea of the areas, they need to improve. The significant attributes identified can be utilized to improve the institution's approach to job placement. For instance, the model highlighted the importance of grades and study frequency on job prediction. Institutions can focus on these two aspects and include them in improving their job placement programs.

The research findings have produced many "whats" but not the "whys.” Future research can further examine these areas. For instance, while it did emphasize living with family as a relevant feature, the output does not reveal why this is the case. Siblings are also an area that requires a closer look and further explore its impact on job placement. A qualitative investigation of the relationship of these socioeconomic factors to the job viability of IT graduates will give a more in-depth insight into this area of inquiry. Future research can also add additional variables to the feature set. For example, the role of IT certifications was not included in the questionnaire. Other socioeconomic factors such as parents' occupation, inflation rate, or student debt can also be added and investigated.

References

Adlemo, A. (2022). The capstone project's role in transitioning to industry for recently graduated software engineers [Thesis]. Jönköping University, Sweden. https://www.diva-portal.org/smash/get/diva2:1681749/FULLTEXT01.pdf

Al-Asmar, A. A., Oweis, Y., Ismail, N. H., Sabrah, A. H. A., & Abd-Raheam, I. M. (2021). The predictive value of high school grade point average to academic achievement and career satisfaction of dental graduates. BMC Oral Health, 21(1). https://doi.org/10.1186/s12903-021-01662-5

Ali J., Khan R.U, Ahmad N., & Maqsood I. (2012). Random Forests and Decision Trees. International Journal of Computer Science Issues, 9(5), 272–278. http://ijcsi.org/papers/IJCSI-9-5-3-272-278.pdf

Ali, M., & Jalal, H. (2018). Higher Education as a Predictor of Employment: The World of Work Perspective. Bulletin of Education and Research, 40(2), 79–90. http://files.eric.ed.gov/fulltext/EJ1209685.pdf

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., & Farhan, L. (2021). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. Journal of Big Data, 8(1). https://doi.org/10.1186/s40537-021-00444-8

Apalla, J. (2022, November 13). Average Time to Get a Job After Graduation in 2023. Degree Planet. https://www.degreeplanet.com/average-time-to-get-a-job-after-graduation/

Boateng, E. Y., & Abaye, D. A. (2019). A Review of the Logistic Regression Model with Emphasis on Medical Research. Journal of Data Analysis and Information Processing, 07(04), 190–207. https://doi.org/10.4236/jdaip.2019.74012

Berrar, D. (2019). Cross-Validation. Encyclopedia of Bioinformatics and Computational Biology, 542–545. https://doi.org/10.1016/b978-0-12-809633-8.20349-x

Chan, J. Y. L., Leow, S. M. H., Bea, K. T., Cheng, W. K., Phoong, S. W., Hong, Z. W., & Chen, Y. L. (2022). Mitigating the Multicollinearity Problem and Its Machine Learning Approach: A Review. Mathematics, 10(8), 1283. https://doi.org/10.3390/math10081283

Chen, H., Hu, S., Hua, R., & Zhao, X. (2021). Improved naive Bayes classification algorithm for traffic risk management. EURASIP Journal on Advances in Signal Processing, 2021(1). https://doi.org/10.1186/s13634-021-00742-6

Chen, T., & Guestrin, C. (2016). XGBoost. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. https://doi.org/10.1145/2939672.2939785

Chengsheng, T., Huacheng, L., & Bing, X. (2017). AdaBoost typical Algorithm and its application research. MATEC Web of Conferences, 139, 00222. https://doi.org/10.1051/matecconf/201713900222

Choi R. Y., Coyner, A. S, Kalpathy-Cramer J., Chiang M.F., & Campbell, P. (2020). Introduction to Machine Learning, Neural Networks, and Deep Learning. Translational Vision Science & Technology, 9(2), 14. https://doi.org/10.1167/tvst.9.2.14

Clemente, C., & Kwak, M. (2022, October 5-8), Utilizing Data Science and Analytics in Predicting Campus Placement [Paper Presentation], 62nd IACIS Annual Conference, Las Vegas, Nevada, United States

Davis-Kean, P. E., Tighe, L. A., & Waters, N. E. (2021). The Role of Parent Educational Attainment in Parenting and Children's Development. Current Directions in Psychological Science, 30(2), 186–192. https://doi.org/10.1177/0963721421993116

Dridi, S. (2022). Unsupervised Learning - A Systematic Literature Review. Preprint. https://doi.org/10.31219/osf.io/kpqr6

Dubey, A. (2021, December 7). Feature Selection Using Random forest - Towards Data Science. Medium. https://towardsdatascience.com/feature-selection-using-random-forest-26d7b747597f

Fernandez, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. Journal of Artificial Intelligence Research, 61, 863–905. https://doi.org/10.1613/jair.1.11192

Galbraith, D., & Sunita, M. (2020). The Potential Power of Internships and the Impact on Career Preparation. Research in Higher Education, 38. https://files.eric.ed.gov/fulltext/EJ1263677.pdf

Gmbh, R. (2023). Optimize Selection (Evolutionary) - RapidMiner Documentation. https://docs.rapidminer.com/latest/studio/operators/modeling/optimization/feature_selection/optimize_selection_evolutionary.html

Guler, C. (2021). Algorithmic Thinking Skills without Computers for Prospective Computer Science Teachers. Kuramsal Eğitimbilim, 14(4), 570–585. https://doi.org/10.30831/akukeg.892869

Guo, T., Xia, F., Zhen, S., Bai, X., Zhang, D., Liu, Z., & Tang, J. (2020). Graduate Employment Prediction with Bias. Proceedings of the AAAI Conference on Artificial Intelligence, 34(01), 670–677. https://doi.org/10.1609/aaai.v34i01.5408

Guyon, Emily (2020). Marketing U: Preparing Students to Succeed in the Job Search Process. Undergraduate Review, 15, 103-115.https://vc.bridgew.edu/undergrad_rev/vol15/iss1/12

Hagiwara, Y., Shiroiwa, T., Taira, N., Kawahara, T., Konomura, K., Noto, S., Fukuda, T., & Shimozuma, K. (2022). Gradient Boosted Tree Approaches for Mapping European Organization for Research and Treatment of Cancer Quality of Life Questionnaire Core 30 Onto 5-Level Version of EQ-5D Index for Patients with Cancer. *Value in Health*. https://doi.org/10.1016/j.jval.2022.07.020

Harihar, V., & Bhalke, D. (2020). Student Placement Prediction System using Machine Learning. SAMRIDDHI: A Journal of Physical Sciences, Engineering and Technology, 12(SUP 2), 85-91. https://doi.org/10.18090/samriddhi.v12iS2.17

Hashmi, F., Ameen, K., & Soroya, S. (2019). Does postgraduate degree make any difference in job performance of information professionals? Library Management, 41(1), 14–27. https://doi.org/10.1108/lm-07-2019-0042

Hauss, K. (2020). What are the social and scientific benefits of participating at academic conferences? Insights from a survey among doctoral students and postdocs in Germany. Research Evaluation, 30(1), 1–12. https://doi.org/10.1093/reseval/rvaa018

Huynh, T.V., Nguyen, K.V., Nguyen, N. L., & Nguyen, A. G. (2020). Job Prediction: From Deep Neural Network Models to Applications. 2020 RIVF International Conference on Computing and Communication Technologies (RIVF). https://doi.org/10.1109/rivf48685.2020.9140760

Hossin, M., & Sulaiman, M.N. (2019). A Review on Evaluation Metrics for Data Classification Evaluations. CERN European Organization for Nuclear Research - Zenodo. https://doi.org/10.5281/zenodo.3557376

Jindal, S., Sachdeva, M., & Kushwaha, A. K. S. (2022). Performance evaluation of machine learning based voting classifier system for human activity recognition. Kuwait Journal of Science. https://doi.org/10.48129/kjs.splml.19189

Jun, Z. (2021). The Development and Application of Support Vector Machine. Journal of Physics: Conference Series, 1748(5), 052006. https://doi.org/10.1088/1742-6596/1748/5/052006

Karatepe, O. M. (2015). The effects of family support and work engagement on organizationally valued job outcomes. Tourism: An International Interdisciplinary Journal, 63(4), 447–464. http://hrcak.srce.hr/149998?lang=en

Katkar V., Iyer, S., Kemkar, C., & Kolangara, N. (2019). Early Placement Prediction System for Engineering Students of Indian Universities. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3419952

Ke, G, Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. & Liu, T. (2017). LightGBM: a highly efficient gradient boosting decision tree. Neural Information Processing Systems, 30, 3149–3157. https://papers.nips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

Khan, S. S., Ahmad, A., & Mihailidis, A. (2019, December 23). Bootstrapping and multiple imputation ensemble approaches for classification problems. Journal of Intelligent &Amp; Fuzzy Systems, 37(6), 7769–7783. https://doi.org/10.3233/jifs-182656

Kulkarni, A., Chong, D., & Batarseh, F. A. (2020). Foundations of data imbalance and solutions for a data democracy. Data Democracy, 83–106. https://doi.org/10.1016/b978-0-12-818366-3.00005-8

Kumar, D., Verma, C., Singh, P. K., Raboaca, M. S., Felseghi, R. A., & Ghafoor, K. Z. (2021). Computational Statistics and Machine Learning Techniques for Effective Decision Making on Student's Employment for Real-Time. Mathematics, 9(11), 1166. https://doi.org/10.3390/math9111166

Lee, M. J., Lee, P., & De Villa-Lopez, B. (2019). Hospitality and Tourism Career Fairs: How important are they and how well do they work? Journal of Teaching in Travel & Tourism, 19(4), 326–340. https://doi.org/10.1080/15313220.2019.1592061

Liu, S., See, K. C., Ngiam, K. Y., Celi, L. A., Sun, X., & Feng, M. (2020). Reinforcement Learning for Clinical Decision Support in Critical Care: Comprehensive Review. Journal of Medical Internet Research, 22(7), e18477. https://doi.org/10.2196/18477

Luo, T., & Liu, Y. (2022). Machine truth serum: a surprisingly popular approach to improving ensemble methods. Machine Learning. https://doi.org/10.1007/s10994-022-06183-y

Mahesh, B. (2018). Machine Learning Algorithms-A Review. International Journal of Science and Research (IJSR), 9(1), 381–386. https://www.researchgate.net/profile/Batta-Mahesh/publication/344717762_Machine_Learning_Algorithms_-A_Review/links/5f8b2365299bf1b53e2d243a/Machine-Learning-Algorithms-A-Review.pdf?eid=5082902844932096

Manconi, A., Armano, G., Gnocchi, M., & Milanesi, L. (2022). A Soft-Voting Ensemble Classifier for Detecting Patients Affected by COVID-19. Applied Sciences, 12(15), 7554. https://doi.org/10.3390/app12157554

Mezhoudi, N., Alghamdi, R., Aljunaid, R., Krichna, G., & Düştegör, D. (2021). Employability prediction: a survey of current approaches, research challenges and applications. Journal of Ambient Intelligence and Humanized Computing.

Muraina, I. O., Agoi, M. A., & Omorojor, B. O. (2022). Forecasting Students' Job Placement using Data Science Paradigms. International Journal of Education and Learning Systems, 7. https://www.iaras.org/iaras/filedownloads/ijels/2022/002-0003(2022).pdf

Nti, I. K., Nyarko-Boateng, O., & Aning, J. (2021). Performance of Machine Learning Algorithms with Different K Values in K-fold Cross Validation. International Journal of Information Technology and Computer Science, 13(6), 61–71. https://doi.org/10.5815/ijitcs.2021.06.05

Nzuva, S., & Nderu, L. (2019). The superiority of the ensemble classification methods: A comprehensive review. Journal of Information Engineering and Applications. https://doi.org/10.7176/jiea/9-5-05

Odegua, R. (2019, March). An Empirical Study of Ensemble Techniques (Bagging, Boosting and Stacking). [Paper Presentation]. Deep Learning IndabaX, 2019, Nairobi, Kenya

Olayniyi I & Agoi M. (2022, January 21-23). Data Science Techniques in Predicting Future Job Placement of Students After Graduation [Paper Presentation]. 7th INTERNATIONAL ZEUGMA CONFERENCE ON SCIENTIFIC RESEARCH, Gaziantep, Turkey

Piad, K. C. (2018). Determining the Dominant Attributes of Information Technology Graduates Employability Prediction using Data Mining Classification Techniques. Journal of Theoretical and Applied Information Technology, 96(12), 3780–3790. http://www.jatit.org/volumes/Vol96No12/15Vol96No12.pdf

RapidMiner Amplify the Impact of Your People, Expertise & Data. (2022, October 19). RapidMiner. https://rapidminer.com/

Rao, A. S., Kumar S V, A., Jogi, P., Bhat K, C., Kumar B, K., & Gouda, P. (2019). Student Placement Prediction Model: A Data Mining Perspective for Outcome-Based Education System. International

Journal of Recent Technology and Engineering (IJRTE), 8(3), 2497–2507. https://doi.org/10.35940/ijrte.c4710.098319

Samantha, J., & Poojah, M. (2020, May). Student Placement Chance Predictor. JETIR, 7(5), 1011–1105. https://www.jetir.org/papers/JETIR2005453.pdf

Sarica, A., Cerasa, A., & Quattrone, A. (2017). Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer's Disease: A Systematic Review. Frontiers in Aging Neuroscience, 9. https://doi.org/10.3389/fnagi.2017.00329

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Computer Science, 2(3). https://doi.org/10.1007/s42979-021-00592-x

Schulte, R. V., Prinsen, E. C., Hermens, H. J., & Buurke, J. H. (2021). Genetic Algorithm for Feature Selection in Lower Limb Pattern Recognition. Frontiers in Robotics and AI, 8. https://doi.org/10.3389/frobt.2021.710806

Shurin, A., Davidovitch, N., & Shoval, S. (2021). The Role of the Capstone Project in Engineering Education in the Age of Industry 4.0 - A Case Study. The European Educational Researcher, 4(1), 63–84. https://doi.org/10.31757/euer.414

Smith, C. M. (2020). In the Footsteps of Siblings: College Attendance Disparities and the Intragenerational Transmission of Educational Advantage. Socius: Sociological Research for a Dynamic World, 6, 237802312092163. https://doi.org/10.1177/2378023120921633

Smith, S., Taylor-Smith, E., Smith, C., & Webster, G. (2018). The impact of work placement on graduate employment in computing: Outcomes from a UK-based study. International Journal of Work-Integrated Learning, 19(4), 359–369. https://eric.ed.gov/?id=EJ1199461

Snyder, L. (2022). Fluency with Information Technology (Third Custom Edition for Southern New Hampshire University) (Skill, Concepts, & Capabilities) (5th ed.). Pearson.

Sulastri, A., Handoko, M., & Janssens, J. (2015). Grade point average and biographical data in personal resumes: predictors of finding employment. International Journal of Adolescence and Youth, 20(3), 306–316. https://doi.org/10.1080/02673843.2014.996236

Tangirala, S. (2020). Evaluating the Impact of GINI Index and Information Gain on Classification using Decision Tree Classifier Algorithm*. International Journal of Advanced Computer Science and Applications, 11(2). https://doi.org/10.14569/ijacsa.2020.0110277

Uddin, S., Haque, I., Lu, H., Moni, M. A., & Gide, E. (2022). Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction. Scientific Reports, 12(1). https://doi.org/10.1038/s41598-022-10358-x

Van Rijn, J. N., Holmes, G., Pfahringer, B., & Vanschoren, J. (2017). The online performance estimation framework: heterogeneous ensemble learning for data streams. Machine Learning, 107(1), 149–176. https://doi.org/10.1007/s10994-017-5686-9

Volkmar, G., Fischer, P. M., & Reinecke, S. (2022). Artificial Intelligence and Machine Learning: Exploring drivers, barriers, and future developments in marketing management. Journal of Business Research, 149, 599–614. https://doi.org/10.1016/j.jbusres.2022.04.007

Wang, Y., Zhang, Y., Lu, Y., & Yu, X. (2020). A Comparative Assessment of Credit Risk Model Based on Machine Learning case study of bank loan data. Procedia Computer Science, 174, 141–149. https://doi.org/10.1016/j.procs.2020.06.069

Waugaman, G (2022), A Qualitative Exploration of the Role Older Siblings Play in the College-Going Experiences of Younger First-Generation College Students, All Dissertations. 3152. https://tigerprints.clemson.edu/all_dissertations/3152

Wen, L., & Hughes, M. (2020, May 25). Coastal Wetland Mapping Using Ensemble Learning Algorithms: A Comparative Study of Bagging, Boosting and Stacking Techniques. Remote Sensing, 12(10), 1683. https://doi.org/10.3390/rs12101683

Yılmaz, N., & Sekeroglu, B. (2019). Student Performance Classification Using Artificial Intelligence Techniques. Advances in Intelligent Systems and Computing, 596–603. https://doi.org/10.1007/978-3-030-35249-3_76

Zhao, L., Lee, S., & Jeong, S. P. (2021). Decision Tree Application to Classification Problems with Boosting Algorithm. Electronics, 10(16), 1903. https://doi.org/10.3390/electronics10161903

Zou, T., Zhang, Y., & Zhou, B. (2022). Does GPA matter for university graduates' wages? New evidence revisited. PLOS ONE, 17(4), e0266981. https://doi.org/10.1371/journal.pone.0266981