Enhancing Prediction of Employability of Students: Automated Machine Learning Approach

Jamee Shahriyar
Faculty of Computing
Universiti Teknologi Malaysia
Johor, Malaysia
shahriyar@graduate.utm.my

Johanna Binti Ahmad
Faculty of Computing
Universiti Teknologi Malaysia
Johor, Malaysia
johanna@utm.my

Noor Hidayah Zakaria
Faculty of Computing
Universiti Teknologi Malaysia
Johor, Malaysia
noorhidayah.z@utm.my

Goh Eg Su
Faculty of Computing
Universiti Teknologi
Malaysia Johor, Malaysia
eg.su@utm.my

2022 2nd International Conference on Intelligent Cybernetics Technology & Applications (ICICyTA) | 979-8-3503-9913-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICICyTA57421.2022.10038231

***Abstract*—Enhancing the employability of students has recently become an important strategic goal for most institutions of higher education. With education becoming increasingly employment-oriented, a university's reputation might suffer greatly if a significant proportion of its graduates are unable to find work. Universities generally store huge amounts of student data such as student profiles, student academic records, and student behavioral records. Due to rapid scientific developments in Big Data Analytics and Machine Learning (ML), such data can be analyzed effectively to bring about great returns in predicting the employability of students. Several initiatives and research publications have developed their own ML models for predicting employability, but each has its own set of challenges and inadequacies. The primary purpose of this study is to investigate the usage of Automated Machine Learning (AutoML) as a method to increase the accuracy of prediction for the employability of students and to reduce the complexities involved in choosing the best model and corresponding hyper-parameters for a given student dataset. This research carries out AutoML using the tool Auto-Sklearn which automates the model selection and hyperparameter optimization stages of the ML pipeline. Experiments are performed where ML models are trained using Decision Tree algorithm, Gaussian Naive Bayes algorithm, Multilayer Perceptron, K-Nearest Neighbor and AutoML and the performance metrics accuracy of prediction and Matthew's Correlation Coefficient (MCC) are used to determine the best ML method and the most important employability factors. This research acknowledges that a model suitable for the dataset of one higher studies institution might not be suitable for other higher study institutions with different datasets, which is evident even in the literature for employability prediction, where, different studies corroborate different models to be the best.**

Keywords—employability prediction model, machine learning pipeline, classification, automated machine learning

I. Introduction

A massive survey launched in 2018 found that only 4 out of 10 students felt well prepared for a professional career [1]. Even for the students who do feel well prepared, their actual professional skills and work ethics may not be up to the level that is demanded by employers. The 2018 Future Workforce Survey found that though 77% of freshly graduated students felt confident in their professionalism and work ethic only 43% of employers felt that freshly graduated students are sufficiently competent in their professionalism and work ethic [1]. For a university, if a significant proportion of its students are unable to find jobs after graduation, the reputation of the university will be greatly affected. On the other hand, if the university can predict the students who will be unprepared to tackle the job world and then use that knowledge to take additional care of these students by providing them career counseling, enrolling them into different Professional Employability courses, and so on, the university can greatly improve the employability of the students and at the same time increase its reputation. In recent years, predictive analysis has greatly depended on machine learning (ML) to make predictions in areas such as education management, business decision-making, and such. Several research papers have generated their own ML models for employability, but each have their own complexities and gaps (reviewed in II. Literature Review).

This study proposes the use of Automated Machine Learning (AutoML) to decrease the complexity and human time involved in generating the optimal model. AutoML is a technique that automatically selects the best algorithm or combined algorithm from a generally wide collection of available algorithms and then automatically determines the best hyperparameter settings for each algorithm for optimal performance [2]. This technique has received validation in a number of researches. Zeineddine, Braendle and Farah [3] have used AutoML to enhance the prediction of student performance by utilizing student data features that are available from before students start their new academic program. AutoML can add significant value in the area of employability prediction, however, a review of the literature in the area reveals that there a scarcity of observable work using AutoML. This research will carry out automated machine learning using the tool Auto-Sklearn which automates the model selection and hyperparameter optimization stage in the ML pipeline. Furthermore, this study will use the accuracy of prediction and Matthew's Correlation Coefficient (MCC) as metrics to judge the success of the proposed model over the ML models from past literature.

II. Literature Review

The topic of employability prediction has garnered the attention of researchers for well over a decade. The focus of these studies have mainly been on two fronts: find the best ML model for employability prediction and identifying the most important factors that effect employability. Mishra, Kumar and Gupta [7] have utilized a dataset containing student profile data, student academic data and student psychometric data to predict MCA (Masters of Computing Application) student employability with an accuracy of prediction of 70.19%. They performed their research on a number of classification algorithms and deemed the decision tree to be the best because of its accuracy and ease of comprehensibility. Piad et al. [9] have used a dataset containing key features such as gender, IT core subjects and IT professional courses to predict IT employability with an accuracy of 78.4%. Khadilkar and Joshi [11] have used resumes containing data such as years of experience, GPA, achievements, and publications to train a Gaussian Naive

Bayes model and predict employability with an accuracy of 89%. Khadilkar and Joshi [11] also used other classification models but for their dataset, the Naive Bayes algorithm yielded the highest accuracy of prediction. Linsey [15] in her research explored if key employability signals such as major, GPA, co-curricular activities, and internships can predict if a student secures fulltime employment prior to graduation. She concluded that, the highest accuracy of prediction value for employment prior to graduation (73%), can be obtained with Multilayer Perceptron model. Bharambe et al. [13] experimented with several different algorithm to assess the employability of students and have obtained decent accuracy values for employability prediction. TABLE I summarizes the comparison of ML techniques from past research predicting employability.

TABLE I. TABLE OF COMPARISON FOR ML TECHNIQUES FROM PAST RESEARCH PREDICTING EMPLOYABILITY

Past research	Points of Comparison	Decision Tree	Naive Bayes	K-Nearest Neighbor	Multilayer Perceptron
(Mishra et al., 2016)	Accuracy	70.19%	62.87%	-	70. 64%
	Effectiveness	Mishra et al deemed it best model in the study.	Did not perform very well in study	Not used	Performed really well but long model build time
(Piad *et al.*, 2016)	Accuracy	74.95%	75.33%	-	-
	Effectiveness	Had mid-level performance in study	Had mid-level performance but slightly more than Decision Tree.	Not used	Not used
(Khadilkar and Joshi, 2017)	Accuracy	85%	89%	76%	-
	Effectiveness	Second highest accuracy in study	Highest accuracy in study	Did not perform well	Not Used
(Bharambe *et al.*, 2017)	Accuracy	97%	-	91%	-
	Effectiveness	Second highest accuracy in study	Not used	Had mid-level performance in study.	Not used
(Linsey, 2021)	Accuracy	72%	-	-	73%
	Effectiveness	Second highest accuracy in study	Not used	Not used	Linsey deemed it to be the best model in the study.

A. Factors Affecting Employability

Employability refers to an individual's ability to "gain and maintain a job in a formal organization" [4]. Numerous studies into the employability skills and qualifications that employers demand from job applicants have shed light on some of the features that strongly affect employability. These features are generally referred to as employability skills. Every year, The National Association of Colleges and Employers [5] published a list of rank-ordered employability skills for undergraduate college students desired by employers from various organizations representing 20 industries. The top employability skills include teamwork &collaboration, critical thinking/problem solving, professional work ethic, leadership, career management and information technology application. Rosenburg, Heimler, and Morote [6] in their research on basic employability skills highlighted eight skills: management skills, leadership skills, basic literacy and numeracy skills, work ethic, information technology skills, interpersonal (communication) skills, systems thinking skills and critical thinking skills. Mishra et al., [7] in their research on predicting employability determined that solely academic attributes do not lead to an accurate forecast of a student's employability. They emphasized the importance of psychometric attributes for greater prediction accuracy and better analysis of students' performance.

B. ML Methods Used in Past Research

Decision tree makes predictions by learning decision rules inferred from the training data features and generating a tree-like structure. Nodes that split further into sub-nodes are called decision nodes. Each internal node corresponds to an attribute and acts as a branching condition, each branch of the tree is an attribute value and each leaf node represents a class label (the final outputs of classification) [8]. Traversal from root to leaf is driven by the inputs as different values of the features will trigger different branching and will result in different paths through the tree to the leaf (output class value).

The Gaussian Naive Bayes classification algorithm is based on Bayes' Theorem, $P(A|B) = (P(B|A) * P(A))/P(B)$. Therefore, it is a probabilistic algorithm that makes use of conditional probability computations to make predictions. Naive Bayes makes an assumption that all features of the dataset are independent of each other, that is, the probability of each feature given class will not affect the probability of any other feature given class. This is shown as (1):

$$P(X|C) = \prod_{i=1} P(X_i|C) \quad (1)$$

where $X = (X_1, X_2, \ldots, X_n)$ is a feature vector and C is a class [10]. To perform the classification of a data point x, containing input values $[x_1, x_2, \ldots, x_n]$ the algorithm computes the conditional probability of being in a particular class C given x. It does this calculation for the other class labels as well. Finally, the class label with the highest probability of occurrence for data point x will be deemed the predicted class [3]. The formula used to calculate the probability of occurrence of class C using data point x, where $x = (x_1, x_2, \ldots, x_n)$ is as (2) follows:

$$P(C|x) = \frac{P(C)}{P(x)} * \prod_{i=1}^{n} P(x_i|C) \quad (2)$$

K-nearest neighbor (KNN) algorithm classifies a data element based on the dominant class of its K-nearest neighboring elements within a training set [3]. KNN takes all the data points and utilizes a distance-based approach to compare the closeness between the gathering points of the training and testing data. A specific function, such as the Euclidean or Manhattan distance function, is used to calculate the distance between two data points. An example of one such distance function, the Euclidean function is (3):

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (3)$$

The parameter k in KNN, refers to the number of nearest neighbors to include in the majority voting process. When the model encounters an unlabeled data point, it calculates the distance to the K nearest neighbors and assigns the unlabeled data point to the class with the most training occurrences among the K nearest neighbors [12].

Multilayer Perceptron (MLP) algorithm is modeled after the human brain and tries to mimic the neural connections and interactions in the human brain [14]. MLP uses a mathematical function $F(x)$ to simulate a single brain neuron and layers functions based on a network model to simulate the interconnections among neurons [3]. Signals traveling through neurons are mathematically represented as a combination of inputs $x = (x_1, x_2, \ldots, x_n)$ and weights that form a weighted sum and then pass through an activation

function to generate an output. This function can be represented as (4):

$$F(x) = S(\sum_i^n w_i F(x_i)) \quad (4)$$

Weight is determined by training the network using past data [3]. S is the activation function, which decides whether a neuron should be activated or not and derives output within a specific range of values. The function induces non-linearity in a neuron's output. Some widely used activation functions include linear function, sigmoid function, RELU function, and softmax function. MLP is a multi-layered hierarchical model. There are three sorts of layers in MLP: the input layer, a set of middle layers, and the output layer.

C. Gaps in Existing Research

Each study finds a different ML algorithm to be the best for predicting employability, resulting in inconclusive evidence for the optimal model. The accuracy of prediction for a ML model is highly dependent on the input dataset [16]. Therefore, a model suitable for the dataset of one higher studies institution might not be suitable for other higher study institutions with different datasets. This is evident in the literature for employability prediction. Khadilkar and Joshi [11] in their research, obtained the highest accuracy of prediction value for employability with the Naive Bayes algorithm (89%) while, Mishra, Kumar, and Gupta [7], in their research, obtained the lowest accuracy of prediction value for student employability with the same algorithm (62.87%). This is problematic, as being able to predict, with sufficient accuracy, the employability of a university's students, with its dataset, will serve the greatest benefit to the university in finding students who require more help or in finding sections of the curriculum that require more focus. This further emphasises the necessity of always performing model selection using various models when trying to find the most accurate model for predicting employability for a given dataset. However, given that there are numerous ML model types and model architectures and that each of them has unique hyperparameters that must be adjusted in order to achieve higher prediction accuracies, the task is extremely difficult, time-consuming, and frequently error-prone and may benefit from automation.

D. Automated Machine Learning (AutoML)

AutoML automatically selects the best algorithm or combined algorithm from a generally wide collection of available algorithms and then automatically determines the best hyperparameter settings for each algorithm for optimal performance [2]. AutoML loops through all the available ML algorithms and the possible hyperparameter settings for each algorithm. Model training and model testing are performed each time and based on the comparison of the highest prediction values, probabilistic averaging is applied to decide whether to discard the prior model and maintain the current one, to discard the current model, or to combine the current model with the existing model. Thus, the output of AutoML is the optimal algorithm or combination of algorithms (ensemble). This is represented in Fig. 1.

E. Ensemble Model

The usage of AutoML tools often results in the generation of a model consisting of a collection of algorithms, rather than a single algorithm, called an ensemble model. The output predicted from an ensemble model is decided via probabilistic averaging or voting of the results of individual algorithms in the model. Ensemble models are widely known to perform better than individual models [19]. They reduce model bias and variance in results [20]. Several studies have made use of ensemble models. Miguéis et al. [21] used Boosted trees (an ensemble model) to predict students' performance within their first year of studies with an accuracy of 95%. Zeineddine, Braendle and Farah [3] presented an ensemble model that performed better than a number of ML algorithms at predicting student failure using data available prior to students starting their new year.

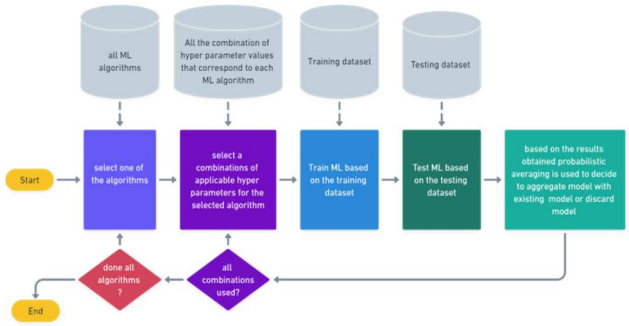

Fig. 1. Automated machine learning method.

III. Methodology

A. First Dataset (AMEO 2015 Dataset)

Considering the significance of several employability skills identified in prior studies, the first dataset used in this study for employability prediction was the Aspiring Minds' Employability Outcomes 2015 (AMEO 2015) dataset [22]. The AMEO 2015 dataset is a one-of-a-kind dataset that includes engineering graduates' demographic data, academic data, career outcomes (salaries, job titles, and job locations), as well as standardized evaluation scores in three key areas: cognitive capabilities, technical skills, and personality. The evaluation scores are obtained from the AMCAT employability test which is one of the most popular employment examinations, with thousands of people taking it each year. The original dataset has a total of 38 attributes and 3998 instances. The attribute characteristics are date-time, ordinal, and numeric data and the dataset possess missing values.

B. Second Dataset (Campus Recruitment Dataset)

The second dataset used in this study for student employability prediction was the Campus Recruitment student dataset. The dataset is about the on-campus placement of postgraduate students. It displays student academic data, student demographic data, and vital career outcome data such as work experience and employability test scores that influence a student's employability and placement. The Campus Recruitment dataset is a highly downloaded dataset in Kaggle and has seen use in several past research. Muhajir et al. [8] used this dataset in their study on enhancing classification algorithms on education-centric datasets using hyperparameter tuning. The original dataset has a total of 15 attributes and 215 instances. The attribute characteristics are ordinal and numeric data and the dataset possess missing values. The predicted class (y variable) for this dataset would be the placement status where students who received placement possessed the desired level of employability while others didn't.

C. Tools

This research hoped to explore the use of AutoML for employability prediction and needed an AutoML tool that could be used to reliably automate the model selection and hyperparameter optimization stages of the employability prediction ML pipeline. The tools that have been used most extensively in past research for AutoML are Auto-WEKA [26] and Auto-Sklearn [27]. Auto-Sklearn generally has better performance than Auto-WEKA [2]. Auto-Sklearn is also more time-efficient than Auto-WEKA. It uses SMAC3 (Sequential Model-based Algorithm Configuration), a re-implementation of SMAC that Auto-WEKA uses, to efficiently perform Bayesian Optimization [28]. As a result, Auto-Sklearn was selected for this research. This research also hoped to carry out ML classification using some of the proposed models from the existing literature to compare against the AutoML model. For this purpose, the extensively used ML python library for ML pipeline design and implementation, Scikit-learn was utilized [29].

D. Performance Evaluation Metrics

In order to evaluate the performance of different ML models this research utilizes the performance measures Accuracy and Matthew's Correlation Coefficient (MCC). Accuracy is used because of its extensive use in many ML-based research papers and its comprehensibility. However, accuracy is not reliable enough. It does not take into account all of the four categories of the confusion matrix in the final score calculation and is, therefore, highly affected by data imbalances. MCC considers the proportion of each of the categories of the confusion matrix in its formula and so is unaffected by data imbalances. It is therefore, much more reliable than accuracy as an evaluation metric. It works by treating the true class and predicted class as two binary variables and then calculates their Pearson product-moment correlation coefficient. The formula for MCC is also derived from the confusion matrix, and is as follows:

$$MCC = \frac{(TP\ x\ TP)-(FP\ x\ FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (5)$$

MCC: worst value = 0; best value = 1 or -1 (negative value represents negative correlation and the classifier's value needs to be reversed to get the ideal classifier).

where:

TP is number of true positive elements
TN is number of true negative elements
FP is number of false positive elements
FN is number of false negative elements

IV. Experiment

The experimental procedure was divided into three stages: Data Preprocessing, Model Training, and Model Evaluation, and was carried out in Jupyter Notebook using Python.

A. Data Preprocessing

The original AMEO 2015 dataset had 38 attributes containing date-time, ordinal and, integer data. At first exploratory data analysis was conducted which helped to identify irrelevant attributes and missing values in the dataset. Then feature engineering was performed. Missing values in the 'Domain' column were imputed with column mean. Attribute columns that were irrelevant to the research were dropped. As the value ranges of the different numerical features in the dataset varied greatly, all the numerical features were standardized. As per the discussion in Section A of Chapter III, values in the salary column were transformed to create a new column "Preparedness" which represented two categories based on the following conditions: For each of the values in the salary column, if the value was equal to or exceeded 350,000 INR the value was assigned the category "prepared", else the value was assigned the category "unprepared". Finally, feature selection was performed. Independent variables that were highly correlated to each other, with an absolute Product moment correlation coefficient value greater than 0.85, were identified and one of the variables was removed. The final preprocessed dataset is summarized in TABLE II.

TABLE II. Final Preprocesssed AMEO 2015 Dataset

Type	Attribute_	Preprocessing Performed
Output (Predicted Class)	Preparedness	Binning performed to map salary values to the two categories "prepared" and "unprepared". Then categorical values were encoded to 0 and 1.
Input features	10percentage	Standardization performed.
	12percentage	Standardization performed.
	CollegeGPA	Standardization performed.
	English	Standardization performed.
	Logical	Standardization performed.
	Quant	Standardization performed.
	Domain	Imputation performed to remove missing values with column mean
	conscientiousness	Left as is.
	agreeableness	Left as is.
	extroversion	Left as is.
	neuroticism	Left as is.
	openness to experience	Left as is.

The original Campus Recruitment dataset had 15 attributes containing ordinal and integer data. At first exploratory data analysis was conducted which helped to identify irrelevant attributes in the dataset. Then feature engineering was performed. Features with categorical values were encoded to numerical values. Attributes irrelevant to the research were dropped. Finally, feature selection was performed. Independent variables that were highly correlated to each other, with an absolute Product moment correlation coefficient value greater than 0.85, were identified and one of the variables was removed. Input variables that had very low mutual dependency relative to the predicted y variable were removed. After preprocessing was performed the dataset was left with 9 attributes. The final preprocessed dataset is summarized in TABLE III.

TABLE III. Final Preprocesssed Campus Recruitment Dataset

Type	Attribute_	Preprocessing Performed
Output (Predicted Class)	Status	Categorical values were encoded to 0 and 1.
Input	10percentage	no change
	ssc_p	no change
	hsc_p	no change
	degree_p	no change
	workex	Categorical values were encoded to 0 and 1.
	etest_p	no change
	specialization	Categorical values were encoded to 0 and 1.

Type	Attribute_	Preprocessing Performed
	gender	Categorical values were encoded to 0 and 1.

B. Model Training

The second stage involved using the preprocessed dataset to train separate models for AutoML, Decision Tree, Gaussian Naive Bayes, KNN, and MLP. AutoML automatically selects the best ML algorithm or combination of algorithms and their corresponding best hyperparameter settings. However, for Decision tree, Naive Bayes, KNN and, MLP the best hyperparameter settings needed to be manually selected so that optimal models can be developed during model training. For each of the algorithms other than AutoML Randomized Search was utilized to determine the most optimum combination of hyperparameters for training. Randomized Search is a method for determining the most optimum combination of hyperparameters for a ML algorithm in regards to a particular dataset. It does this by selecting the best hyperparameter values from a list of hyperparameter values for a particular algorithm given to it.

C. Model Evaluation

The third stage was model evaluation, which included testing the accuracy and Matthew's Correlation Coefficient of each of the models using 10 fold cross-validation. Each of the trained models: Decision Tree model, Gaussian Naive Bayes model, KNN model, MLP model, and the autoML model was tested to determine the predicted outcomes which were then compared to the actual outcomes of the testing set, and the number of TF (true positive), FP (false positive), TN (true negative), and FN (false negative) cases for each model were determined. Finally, the TP, TN, FP, and FN values for each model were used to calculate the accuracy of prediction values and MCC values for each of the models.

V. Results

TABLE IV. Final Results for Both Datasets

Results AMEO 2015 Dataset		
Algorithm	**Accuracy (%)**	**MCC**
Decision Tree	61.67	0.19
Gaussian Naïve Bayes	69.85	0.27
K-Nearest Neighbor	71.25	0.22
Multilayer Perceptron	71.28	0.25
Automated Machine Learning	72.45	0.27
Results Campus Recruitment Dataset		
Algorithm	**Accuracy (%)**	**MCC**
Decision Tree	72.09	0.39
Gaussian Naïve Bayes	82.79	0.58
K-Nearest Neighbor	84.15	0.62
Multilayer Perceptron	84.19	0.62
Automated Machine Learning	84.65	0.63

For the AMEO 2015 dataset, the AutoML model achieved the highest accuracy for employability prediction, with a value of 72.45%. The AutoML model also tied with the Gaussian Naive Bayes model for the highest Matthew's Correlation Coefficient value, 0.27. In the case of the Campus Recruitment dataset the AutoML model achieved the highest accuracy for employability prediction, with a value of 84.65%, and also obtained the highest Matthew's Correlation Coefficient value of 0.63.

The list if algorithms that AutoML performed using auto-sklearn library is able to cycle through includes the Gaussian Naive Bayes, Decision Tree, KNN and MLP. These are some of the models that yielded the highest prediction results for employability prediction. AutoML method could access and combine multiple of these models when creating an ensemble thus yielding higher results than each of the individual models. In order to obtained the results for the Decision Tree, Gaussian Naive Bayes, KNN and MLP Random Search was utilized where list of possible hyperparameters of each of the models had to be provided. For AutoML hyperparameter optimization is automatic. This results in much less human time being utilized for a better outcome as shown by the experimental results. AutoML, however, does require more computation time, and by default and minimum of 30 mins is provided by auto-sklearn. Because human labour and time are so much more expensive than computational resources and time, and because computational time used for Auto-sklearn will replace human time that will span much longer, the trade-off is highly acceptable.

In the experiment, during the feature selection stage, mutual information was used to select only the features that resulted in the most significant change in the predicted outcome for each of the databases. The higher the mutual information value obtained for a particular feature the greater the dependency of the prediction outcome on the feature. For the Campus Recruitment Dataset, ssc_p (secondary school grade percentage) has been the most important employability factor with a mutual information value of 0.269, followed by hsc_p (higher secondary school grade pecentage) at 0.231. In third and forth place are degree_p (Bachelors Degree grade in percentage) and mba_p (MBA grade in pecentage) with mutual information values of 0.191 and 0.060 respectively. Thus, grades obtained by students in different academic examinations at different institutional levels can be assumed play a significant role in their future employability.

However, in the case of the AMEO 2015 dataset standardized evaluation scores in three key areas: cognitive capabilities, technical skills, and personality have claimed the top spots as the most important employability factors. Domain (technical ability in own domain specific subject) is the most important employability factor with a mutual information value of 0.071, followed by Quant (Quantitative ability) with value 0.066. Then onwards, nueroticism (how calm, happy, undisturbed and emotionally stable a person is) with a value of 0.059 and English (vocabulary, grammar and Comprehension analysis) with a value of 0.056 have taken the third and forth positions respectivily. As we go down the list more technical, cogitative and psychometric attributes follow on until we get to collegeGPA, 12percentage and 10percentage in the tenth, eleventh and twelfth positions respectively. Thus, in the study of employability prediction, demographic and academic attributes are not the only factors affecting employability and must not be given sole importance. Attributes once considered as secondary attributes such as personality traits, technical skills and cogitative capabilities possess an equally if not more important role as factors effecting employability and must be considered during research. Moreover, when universities are collecting data of students, such data must be recorded with dedication so that a more complete dataset with higher predictive potential can be obtained and better models can be trained from it. For both datasets, gender and subject specialization have yielded very low mutual information values and have occupied the very bottom positions of the lists. These factors, therefore have a much lower significance as employability factors than the aforementioned features.

VI. Conclusion

The findings of this study will add to the body of knowledge in the subject of student employability prediction. Specifically, it relies on AutoML to enhance the prediction of employability of students. One of the gaps in previous literature was inconclusive evidence for the best model, as the performance of past models varied greatly with different student datasets. In this research, experiments were conducted using two different unrelated student datasets containing different data such as student demographic data, academic data, career outcomes, cognitive capabilities, technical skills, and personality data. AutoML was able to obtain the highest accuracy and MCC values for both datasets. Accordingly, researchers and education institutions are encouraged to adopt AutoML in their search for an optimal student employability prediction model. Furthermore, when conducting future employability prediction research, secondary attributes such as personality traits, technical skills, and cognitive abilities must be strongly taken into account because they play an equally important role—if not a more significant one—as factors affecting employability. Future studies will also try to use AutoML to further investigate the employability factors that institutions of higher studies must focus on to bring about the biggest change in the employability of students.

Acknowledgment

The authors humbly acknowledge the UTM Quick Win Research Grant, R.J130000.7751.4J564, funded by Universiti Teknologi Malaysia, Malaysia.

References

[1] T. Gressling, 84 Automated machine learning. 2020. doi: 10.1515/9783110629453-084.

[2] H. Zeineddine, U. Braendle, and A. Farah, "Enhancing prediction of student success: Automated machine learning approach," Comput. Electr. Eng., vol. 89, no. October 2019, p. 106903, 2021, doi: 10.1016/j.compeleceng.2020.106903.

[3] R. Hogan, T. Chamorro-Premuzic, and R. B. Kaiser, "Employability and Career Success: Bridging the Gap Between Theory and Reality," Ind. Organ. Psychol., vol. 6, no. 1, pp. 3–16, 2013, doi: 10.1111/IOPS.12001.

[4] "Career Readiness Competencies: Employer Survey Results." https://www.naceweb.org/career-readiness/competencies/career-readiness-competencies-employer-survey-results/ (accessed Jun. 27, 2022).

[5] S. Rosenberg, R. Heimler, and E.-S. Morote, "Basic Employability Skills: A Triangular Design Approach".

[6] T. Mishra, D. Kumar, and S. Gupta, "Students' employability prediction model through data mining," Int. J. Appl. Eng. Res., vol. 11, no. 4, pp. 2275–2282, 2016.

[7] D. Muhajir, M. Akbar, A. Bagaskara, and R. Vinarti, "Improving classification algorithm on education dataset using hyperparameter tuning," Procedia Comput. Sci., vol. 197, pp. 538–544, 2021, doi: 10.1016/j.procs.2021.12.171.

[8] K. C. Piad, M. Dumlao, M. A. Ballera, and S. C. Ambat, "Predicting IT employability using data mining techniques," 2016 3rd Int. Conf. Digit. Inf. Process. Data Mining, Wirel. Commun. DIPDMWC 2016, no. July, pp. 26–30, 2016, doi: 10.1109/DIPDMWC.2016.7529358.

[9] E. P. F. Lee et al., "An ab initio study of RbO, CsO and FrO (X2∑+; A2∏) and their cations (X3∑-; A3∏)," Phys. Chem. Chem. Phys., vol. 3, no. 22, pp. 4863–4869, 2001, doi: 10.1039/b104835j.

[10] N. Khadilkar and D. Joshi, "Predictive Model on Employability of Applicants and Job Hopping using Machine Learning," Int. J. Comput. Appl., vol. 171, no. 1, pp. 37–41, 2017, doi: 10.5120/ijca2017914966.

[11] X. Luo, J. XuYu, and Z. Li, "Advanced Data Mining and Applications: 10th International Conference, ADMA 2014 Guilin, China, December 19-21, 2014 Proceedings," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 8933, no. December, 2014, doi: 10.1007/978- 3-319-14717-8.

[12] Y. Bharambe, N. More, M. Mulchandani, R. Shankarmani, and S. G. Shinde, "Assessing employability of students using data mining techniques," 2017 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI 2017, vol. 2017-Janua, no. October, pp. 2110–2114, 2017, doi: 10.1109/ICACCI.2017.8126157.

[13] James Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani, "An Introduction to Statistical Learning with Applications in R Related papers," 2013.

[14] Autoridad Nacional del Servicio Civil, "済無 No Title No Title No Title," Angew. Chemie Int. Ed. 6(11), 951–952., no. December, pp. 2013–2015, 2021

[15] A. Jha, A. Chandrasekaran, C. Kim, and R. Ramprasad, "Impact of dataset uncertainties on machine learning model predictions: The example of polymer glass transition temperatures," Model. Simul. Mater. Sci. Eng., vol. 27, no. 2, p. 24002, 2019, doi: 10.1088/1361-651X/aaf8ca

[16] L. Tuggener et al., "Automated Machine Learning in Practice: State of the Art and Recent Results," Proc. - 6th Swiss Conf. Data Sci. SDS 2019, pp. 31–36, 2019, doi: 10.1109/SDS.2019.00-11.

[17] M. Martin Salvador, M. Budka, and B. Gabrys, "Automatic Composition and Optimization of Multicomponent Predictive Systems with an Extended Auto-WEKA," IEEE Trans. Autom. Sci. Eng., vol. 16, no. 2, pp. 946–959, 2019, doi: 10.1109/TASE.2018.2876430.

[18] A. Lacoste et al., "Agnostic Bayesian Learning of Ensembles," vol. 32, 2014.

[19] M. J. Kim, S. H. Min, and I. Han, "An evolutionary approach to the combination of multiple classifiers to predict a stock price index," Expert Syst. Appl., vol. 31, no. 2, pp. 241–247, Aug. 2006, doi: 10.1016/J.ESWA.2005.09.020.

[20] V. L. Miguéis, A. Freitas, P. J. V. Garcia, and A. Silva, "Early segmentation of students according to their academic performance: A predictive modelling approach," Decis. Support Syst., vol. 115, pp. 36–51, Nov. 2018, doi: 10.1016/J.DSS.2018.09.001.

[21] A. Pawha and D. Kamthania, "Quantitative analysis of historical data for prediction of job salary in India - A case study," J. Stat. Manag. Syst., vol. 22, no. 2, pp. 187–198, 2019, doi: 10.1080/09720510.2019.1580900.

[22] V. Aggarwal, S. Srikant, and H. Nisar, "AMEO 2015-A dataset comprising AMCAT test scores, biodata details and employment outcomes of job seekers," Proc. 3rd ACM IKDD Conf. Data Sci. CODS 2016, pp. 2015–2017, 2016, doi: 10.1145/2888451.2892037.

[23] J. Mohd Abdul Kadir, N. Naghavi, G. Subramaniam, and N. A'amilyn Abdul Halim, "Unemployment among Graduates - Is there a Mismatch?," Int. J. Asian Soc. Sci., vol. 10, no. 10, pp. 583– 592, 2020, doi: 10.18488/journal.1.2020.1010.583.592.

[24] "Engineer Salary in India - Average Salary." https://in.talent.com/salary?job=Engineer (accessed Jun. 27, 2022).

[25] L. Kotthoff, C. Thornton, H. H. Hoos, F. Hutter, and K. Leyton- Brown, "Auto-WEKA 2.0: Automatic model selection and hyperparameter optimization in WEKA," J. Mach. Learn. Res., vol. 18, pp. 1–5, 2017, doi: 10.1007/978-3-030-05318-5_4.

[26] M. Feurer, K. Eggensperger, S. Falkner, M. Lindauer, and F. Hutter, "Auto-Sklearn 2.0: Hands-free AutoML via Meta-Learning," no. July, pp. 0–18, 2020, [Online]. Available: http://arxiv.org/abs/2007.04074

[27] A. Truong, A. Walters, J. Goodsitt, K. Hines, C. B. Bruss, and R. Farivar, "Towards automated machine learning: Evaluation and comparison of AutoML approaches and tools,0209

[28] A. Truong, A. Walters, J. Goodsitt, K. Hines, C. B. Bruss, and R. Farivar, "Towards automated machine learning: Evaluation and comparison of AutoML approaches and tools," Proc. - Int. Conf. Tools with Artif. Intell. ICTAI, vol. 2019-Novem, no. 2017, pp. 1471–1479, 2019, doi: 10.1109/ICTAI.2019.00209.

[29] A. Jović, K. Brkić, and N. Bogunović, "An overview of free software tools for general data mining," 2014 37th Int. Conv. Inf. Commun. Technol. Electron. Microelectron. MIPRO 2014 - Proc., pp. 1112–1117, 2014, doi: 10.1109/MIPRO.2014.6859735.