ISSN 2347 - 3983
Volume 8. No. 5, May 2020
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter102852020.pdf
https://doi.org/10.30534/ijeter/2020/102852020

Predicting Students' Employability using Support Vector Machine: A SMOTE-Optimized Machine Learning System

Cherry D. Casuat[1], Enrique D. Festijo[2], Alvin Sarraga Alon[3]
[1]Technological Institute of the Philippines, Philippines, ccasuat.cpe@tip.edu.ph
[2]Technological Institute of the Philippines, Philippines, enrique.festijo@tip.edu.ph
[3]Technological Institute of the Philippines, Philippines, aalon.cpe@tip.edu.ph

ABSTRACT

The graduates in every institution reflect the skills developed and competencies acquired by the tudents through the education offered by the institution that is suitable in the companies. Employability of graduates becomes one of the performance indicators for higher educational institutions (HEIs). Therefore, it is important to accentuate the employability of graduates. This is the reason why this research is being carried out. This study involved twenty-seven thousand (27,000) information consist of three thousand (3000) observations and twelve (12) features of student's mock job interview evaluation results (MJI), on-the-job training (OJT) student's performance rating and general point average (GPA) of students enrolled in the on-the-job training course of SY 2015 to SY 2018. To address the issue in imbalance datasets where the minority class, the researchers used synthetic minority over-sampling technique (SMOTE) were applied in this study to address the issue in imbalanced datasets Six learning algorithms with SMOTE were used such as Decision Trees (DT), Random Forest (RF), and Support vector machine (SVM), K- Nearest Neighbor (KNN), Logistic Regression (LR) to understand how students, get employed. The six algorithms were evaluated through the performance matrix as accuracy measures, precision and recall measures, f1-score, and support measures. During the experiments, Support Vector Machine (SVM) obtained 91.22% inaccuracy measures which were significantly better than all of the learning algorithms, DT 85%, RF 84%. The learning curve produced during the experiment displays the training error results which were above the one for validation error while the validation curve displays the testing output where gamma was best at 10 to 100 in gamma 5. This concludes that the model produced with SVM was not under fitted and over-fit. This study is very promising which leads the researchers to be motivated to enhance the process and to validate the produced predictive model for further study.

Key words : Employability prediction system, Decision trees, K-nearest neighbour, Logistic regression, Naïve Bayes, Random Forest, SMOTE, Support Vector Machine

1. INTRODUCTION

One of the main challenges of schools and universities today is providing programs that are aligned with the policy of commission on higher education (CHED) and delivering consistent outcomes that can be accepted not only in the Philippines but also in other countries. The higher educational institutions shift their paradigm from just simply educating the students to developing life-long learning skills. The CHED in the Philippines initiated reforms for education sectors through the conversion of the traditional way of teaching known as teacher-centered to an outcomes-based education (OBE) curricula that are commonly known as student-centered, as per CHED-CMO No. 46 s. 2012 [1]. The higher education institutions' programs offering and syllabus were based on the principles of OBE by which accentuate the type of delivery of its services to the students. To strengthen and assist the country by producing graduates with critical thinking, behavioral and life-long learning skills and competencies aligned with institutional learning outcomes, industry desired values, and international standards [2].

Graduate readiness study has also been conducted by other scholars, where the aim of their paper is,(1) to investigate student's experience when it comes to essential skills acquired in the university ; (2), student perception when it comes to potential job role when they graduate [3]. According to a recent survey by the Malaysian Ministry of Education assessing "(youth)" unemployment and graduates, In particular, just 53 percent of the 273,373 graduates find jobs within six months of graduation in 2015, 24 percent of graduates have no job after graduation and 18 percent were engaged in continuing education. That is why, as mentioned, only 53% of students were employed because of "the discrepancy between the education provided at the universities and the skills needed by the industry. According to the researchers in Malaysia and China, most of the university curriculum where they conducted their studies reflect the current skill requirements of the industry [4]. In a Alsore researchers have studied graduate employability. Research utilizing data mining and modeling methods has been carried out that highlights particular computing and data management challenges. [5]. Using the data analytics, the results were evaluated by (1) monitoring the job status of graduates by giving them prompts and invitations; (2) encouraging them to maintain track of the position they

wanted; and (3) determining the jobs approaches function well, especially in the sector-specific region. [6]. Unfortunately, the complexity of the workplace and the advent of modern technologies have shifted, and the varied demands. of consumers demand the definition of "globally competitive" for employable graduates. It also challenges the capacity of universities to satisfy the need for graduates who are suitable for the job in the industry [7]. These concepts of career development and employability lead college education to evaluate their program offerings and to test their effectiveness and congruence with the needs of the sectors to deliver qualified students who can quickly be consumed by industries.

The assessment of being suitable for the job often follows the theory of human capital, in which individuals' personal and technical growth are called assets in human capital that acts as factors for their degree of employment and personal earnings. Graduates will then make substantial improvements in their intellectual resources to allow their potential employers more marketable [8]. However, It is increasingly important for individuals to retrain or develop new knowledge and abilities to address the requirements of the rapidly growing workforce and the multifaceted entrepreneurial environment. They will develop their professional abilities, values, and work experience and adapt to the changing labour market requirements.

Most published researches and studies used data mining techniques to predict employability. Some of the techniques were Tree of Decision, Naïve Bayes, and Vector Machine Support [9]. Often used in data mining techniques are the Logistic Regression, K-Nearest Neighbor, Random Woodland, SVM (Linearsvc), Quadratic Discriminant Analysis (QDA), and Multi-class Ada Boosted [10]. The application of Machine learning when it comes to forecasting employability is in the infancy period. They compared numerous algorithms in the analysis carried out by Ohio University where the datasets used were from business education. The aforementioned research will not find the datasets regarding the mismatch [11].

This paper seeks to establish a machine learning method to forecast the employability of the applicant and to examine the signs of their skill set. This paper is in the production stage of a model focused on machine learning to forecast the employability of students. The researchers were inspired to perform the study in the light of emerging areas such as operational intelligence or instructional analytics to strengthen and encourage certain ability sets found that will lead to the enhanced jobs of engineering students.

2. METHODOLOGY

The methodology of this proposed method was divided into three-phase such as Data collection, Preprocessing, Training of datasets which handling imbalance datasets were highlighted in the study.

2.1 Contextual Diagram

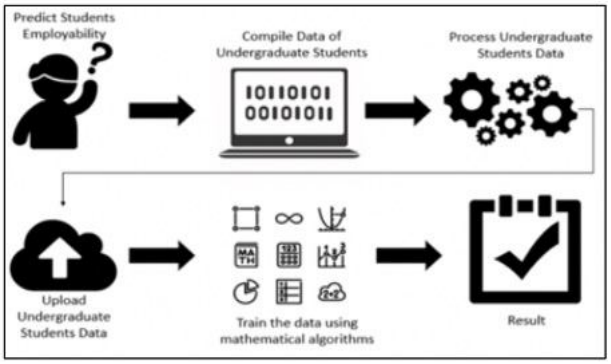

Figure 1: Contextual Diagram

This figure 1 shows how the system intends to work in predicting the employability of undergraduate students. First, it is important to know the objective is to predict the student's employability. Next is compiling all the student's data, in this stage the datasets collected will be cleaned and normalized and then merged different datasets. The datasets will be trained using the six algorithms, the best model that will be created will be used in the system that will be developed.

2.2 Datasets Collection

Table 1: Students' Employability Datasets

Feature Number	*Description*
1	General Appearance
2	Manner of Speaking
3	Physical Condition
4	Mental alertness
5	Self-Confidence
6	Ability to present ideas
7	Communication skills
8	Student performance rating
9	General point grade
10	Student's name
11	Student program
12	Student number

The datasets were collected from different agencies in the university which consists of Mock job Interview Results consist of three thousand (3000) observations and twelve (12) features, Student Performance Rating of the OJT students collected by the On-The-Job Training (OJT) Faculty In-charge and General Point Average from the Registrar's Office. The datasets collected need to be normalized and cleaned. The datasets that were collected were compliant with the Data Privacy Act of the Philippines.

2.3 Preprocessing of Datasets

The preprocessing stage consists of cleaning the first of the datasets. The researchers used data normalization where each attribute or column was filled with the median values when there is a missing value on attributes or columns. For row or number of observations were filled with the mean of that number of observations or row when there is a missing value for row [10]. Then merging of the cleaned datasets to create a

consolidated dataset that comes from a mock-job interview, OJT student performance rating, a general percentage grade

2.4 Training the Datasets

The proponents trained the datasets using the 70-30 splitting of datasets. The learning algorithms such as SVM, Decision Trees, Random Forest, Logistic Regression, Naïve Bayes, and KNN. Based on the training conducted the SVM got the highest accuracy of 92.22% of the entire used algorithm, which means that the Support Vector Machine was the best-created model. The Support Vector Machine (SVM) analyzes the data for classification analysis.

2.2.1Handling Imbalanced Datasets

The synthetic minority over-sampling technique (SMOTE) which applies the k-nearest neighbor algorithm that chose, combines and generates the synthetic samples in the nearest space. The algorithm takes the vectors of the attributes and its closest neighbors, measures the difference between such paths. It is multiplied and added back to the feature by a random number (0, 1). SMOTE algorithm is a pioneer and SMOTE is a basis for several other algorithms [12]. In this study, SMOTE was used to addressed the issues of imbalanced datasets in employability datasets where the majority class is employable.

2.5SOFTWARE DESIGN

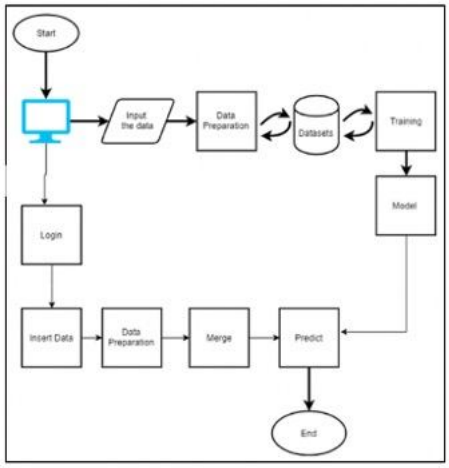

Figure 2: System Architecture

Figure 2 shows that the first step is to input data. The datasets will be preprocessed (cleaned and merged, splitting of datasets into 70,30). Then, training of datasets will take place to create a model. One model was created, that model will be used in the system. The user can log-in to the system GUI by providing a password and log-in successfully. Only those who have a user account can upload the datasets and predict if the list of students is employable or less employable.

2.5.1 Constraints

The system will predict student's employability if it was merged in the On-the-Job Training CSV file and Mock Job CSV file at a time. The system will use acquired machine learning techniques using Python programming language to process the data that will show the result in the GUI and save it as CSV file directly on the documents folder of your computer. PYQT5 and QT designer will also be used for the GUI design

2.5.2 Functional View

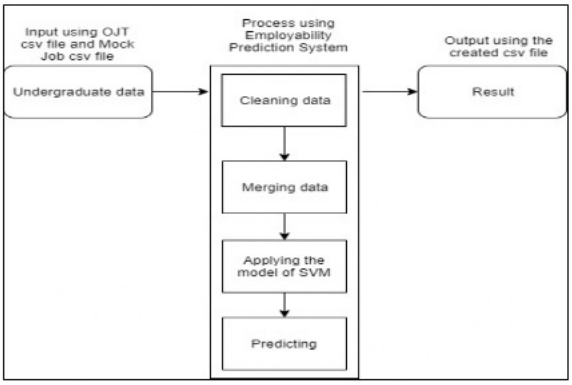

Figure 3: The Proposed System Functional View

Figure 3 shows the proposed system input, process, and output views. Mock-job interview .csv and OJT assessment tool Ratings will be accepted for input. Then, once cleaned and merged, the SVM model will be applied to predict the employability of the students. There are different studies applied prediction such as in career management [13], also in shortlisting of job [14]and some are for modelling purposes [15] – [17].

2.5.3 System Flowchart

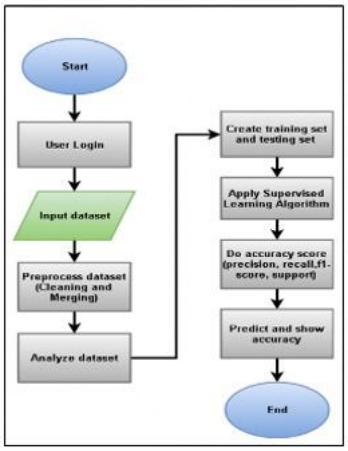

Figure 4: Students' employability prediction system flow chart

The figure above shows how the data flows and the decision was made to control the events. The process will apply the algorithm that was chosen. After the pre-processing, the analyzed dataset will split into three categories which are testing, training, and validation. After the training and using the model, the system will now predict and show the accuracy result based on the datasets.

3 TESTING AND VALIDATION

3.1 Performance Measures

Table 1: Students' Employability Datasets

MODEL	ACCURACY	RECALL	F1-SCORE	PRECISION
SVM	.9122	.9110	.9100	.9100
KNN	.7990	.7884	.7900	.7900
RF	.6475	.8730	.8730	.6950
NB	.6181	.6363	.6150	.6150
LR	.6545	.4090	.6150	.6400
DT	.5636	.4545	.5450	.5450

Among the learning algorithms in table 2, SVM obtained the highest accuracy which is 91.22%, 91.10% for recall, and both 91% for f1-score and precision.

3.2 Learning curve

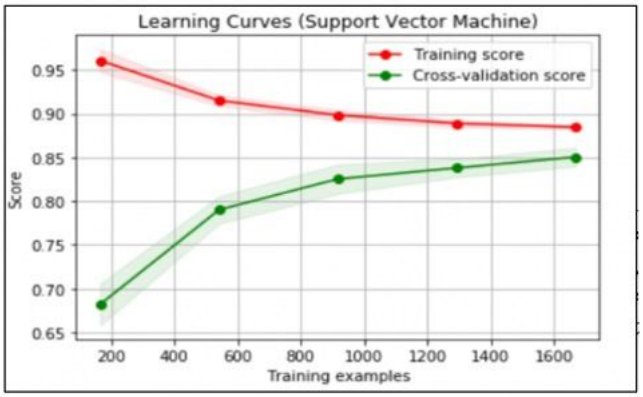

Figure 5: Support Vector Machine learning curves in gamma 5

Figure 5 shows the learning curve of SVM where the values in maximum training score mean is 0.960843, the maximum cross-validation score means is 0.850478, maximum training score is 0.9819277, maximum cross-validation score is 0.72966. The learning curve for the training error results was above the one for the validation error. The accuracy measure described how good the model is and the MSE on the other side described how bad the model is. The irreducible error gives an upper bound.

3.3 Validation Curve

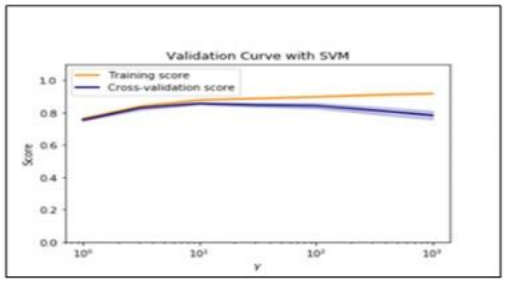

Figure 6: SVM Validation Curve in gamma5

Figure 6 shows the validation curve with SVM in gamma 5 where the maximum training R-squared score was 0.918 and the maximum cross-validation score was 0.857. It shows that the validation curve with SVM in gamma 5, the gamma is best at 10 to 100.

3.4 System Graphical User Interface

The best model created which is the SVM was applied to the employment prediction system.

3.4.1 User's account registration

Figure 7: Account registration

The figure above shows the account registration where the user will create an account first.

Figure 8: Log-in Interface

Figure 8 shows the log-in interface of the system. The user log-in was his/her credentials to be able to use the system.

3.4.2 Uploading and merging of datasets

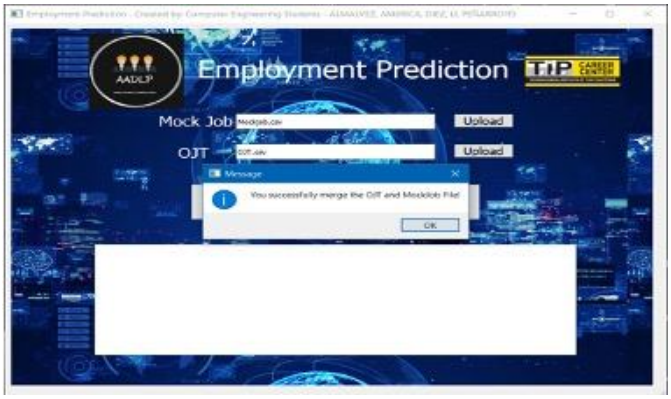

Figure 9: Uploading the Mock job and OJT datasets

3.4.3 Student's Employability Result

Figure 10: Students' employability prediction resultswhere the SVM model was applied

Figure 10 shows the application of the SVM model in predicting student's employability. The system predicts if the student is employable or less employable. Then the system recommends what areas the student needs to improve to be more employable at the time of graduation.

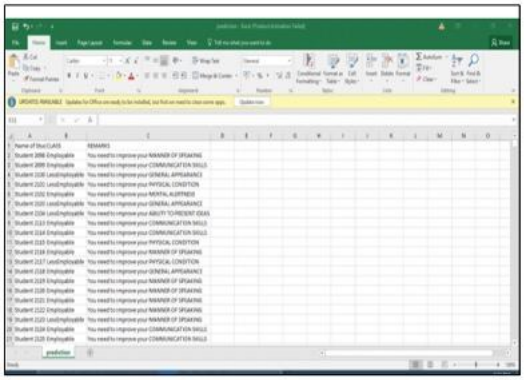

Figure 11: Excel file generated Prediction Results

Figures 7-11 shows the developed student's employability prediction system. The users need to just log-in and upload the datasets needed to predict the employability of the students.

4.CONCLUSION

The Higher Education Institutions (HEIs) becoming more accountable for student's career outcomes and as jobs in the labor market increases its competition, the institution needs to identify students' employability. This study develops a student employability prediction system using an SVM machine learning approach of predicting students' employability where the issues in imbalanced datasets have been addressed using SMOTE. The best algorithm that has the highest accuracy and has the highest performance evaluation compare to the other five learning algorithms that have been trained to create the best model. Therefore, researchers concluded that Support Vector Machine (SVM) produces a predictive model that obtained 91.22% for the accuracy and for recall measures which are .911 or 91.10% and 91% for precision respectively. The researchers realized that gamma is best at 10 to 100 in gamma 5 as shown in figure 6. The researchers concluded that the learning curve and validation curve that it showedwas not overfitted or underfit.

ACKNOWLEDGEMENT

The proponents would like to thank the Career Center of TIP-Manila especially to the SDP Officer and Career Adviser for their unwavering support to the proponents and the MR. SUAVE Laboratory of Technological Institute of the Philippines for all computing facilities that researches have been used to make this study possible.

REFERENCES

1. **CMO 46 s. 2012 - CHED**, *CHED*, 2020. [Online]. Available: https://ched.gov.ph/cmo-46-s-2012/.
2. **Implementing Rules and Regulations of the Enhanced Basic Education Act of 2013 | GOVPH**, *Official Gazette of the Republic of the Philippines*, 2020. [Online]. Available: https://www.officialgazette.gov.ph/2013/09/04/irr-republic-act-no-10533/.
3. W. Teng, C. Ma, S. Pahlevansharif and J. Turner, **Graduate readiness for the employment market of the 4th industrial revolution**, *Education + Training*, vol. 61, no. 5, pp. 590-604, 2019. doi: 10.1108/et-07-2018-0154
4. M. Alias, G. Sidhu and C. Fook, **Unemployed Graduates' Perceptions on their General Communication Skills at Job Interviews**, *Procedia - Social and Behavioral Sciences*, vol. 90, pp. 324-333, 2013. doi: 10.1016/j.sbspro.2013.07.098
5. B. Tapado, G. Acedo and T. Palaoag, **Evaluating information technology graduates employability using decision tree algorithm**, *Proceedings of the 9th International Conference on E-Education, E-Business, E-Management and E-Learning - IC4E '18*, 2018. doi: 10.1145/3183586.3183603
6. R. Bridgstock and D. Jackson, **Strategic institutional approaches to graduate employability: navigating meanings, measurements and what really matters**, *Journal of Higher Education Policy and Management*, vol. 41, no. 5, pp. 468-484, 2019. doi: 10.1080/1360080x.2019.1646378
7. R. Bringula, A. Balcoba and R. Basa, **Employable Skills of Information Technology Graduates in the Philippines**, *Proceedings of the 21st Western Canadian Conference on Computing Education - WCCCE '16*, 2016. doi: 10.1145/2910925.2910928
8. L. Almendarez, **Human Capital Theory: Implications for Educational Development in Belize and the Caribbean**, *Caribbean Quarterly*, vol. 59, no. 3-4, pp. 21-33, 2013. doi: 10.1080/00086495.2013.11672495
9. W. Fok et al., **Prediction model for students' future development by deep learning and tensorflow artificial intelligence engine**, *2018 4th International Conference on Information Management (ICIM)*, 2018. doi: 10.1109/infoman.2018.8392818
10. Y. Bharambe, N. Mored, M. Mulchandani, R. Shankarmani and S. Shinde, **Assessing employability of**

students using data mining techniques, *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2017. doi: 10.1109/icacci.2017.8126157

11. A. Farahat, A. Elgohary, A. Ghodsi and M. Kamel, **Greedy column subset selection for large-scale data sets**, *Knowledge and Information Systems*, vol. 45, no. 1, pp. 1-34, 2014. doi: 10.1007/s10115-014-0801-8
12. N. Chawla, K. Bowyer, L. Hall and W. Kegelmeyer, **SMOTE: Synthetic Minority Over-sampling Technique**, *Journal of Artificial Intelligence Research*, vol. 16, pp. 321-357, 2002. doi: 10.1613/jair.953
13. S. J, **Career Prediction through Cognitive Models using Sudoku Game – The Assessment of Applicability**, *International Journal of Emerging Trends in Engineering Research*, vol. 7, no. 11, pp. 473-480, 2019. doi: 10.30534/ijeter/2019/127112019
14. R. Gustilo, **An Analytic Hierarchy Process Approach in the Shortlisting of Job Candidates in Recruitment**, *International Journal of Emerging Trends in Engineering Research*, pp. 333-339, 2019. doi: 10.30534/ijeter/2019/17792019.
15. A. Alon, **A Machine Vision Detection of Unauthorized On-Street Roadside Parking in Restricted Zone: An Experimental Simulated Barangay-Environment**, *International Journal of Emerging Trends in Engineering Research*, vol. 8, no. 4, pp. 1056-1061, 2020. doi: 10.30534/ijeter/2020/17842020
16. A. Alon, **Machine Vision Recognition System for Iceberg Lettuce Health Condition on Raspberry Pi 4b: A Mobile Net SSD v2 Inference Approach**, *International Journal of Emerging Trends in Engineering Research*, vol. 8, no. 4, pp. 1073-1078, 2020.doi: 10.30534/ijeter/2020/20842020