Comparative Study of Predictive Classification Models on Data with Severely Imbalanced Predictors

Embay Rohaeti[1*], Ani Andriyati[2]

[1,2]*Department of Mathematics, Pakuan University, Bogor, Indonesia*

*corr_author: embay.rohaeti@unpak.ac.id

Abstract – Analysing pre-COVID-19 unemployment in West Java is vital for comprehending and tackling Indonesia's economic challenges. This significance arises not only due to the region's high unemployment rate, but also from the need to understand unemployment patterns before COVID-19, which has become more relevant now during the country's post-pandemic recovery phase. This study evaluates four machine learning models (Random Forest, Linear SVM, RBF SVM, and Polynomial SVM) to classify employment status using demographic and job-related variables. The objective is to find the most suitable model, particularly considering the imbalanced nature of the study-case data. Data from the National Labor Force Survey (SAKERNAS) in August 2019 is utilized, comprising 54,429 respondents across districts in West Java. The four models are evaluated using holdout validation with a 70:30 stratified proportion, repeated for 100 times. Results indicate that the random forest model outperforms others in balanced accuracy, F1-score, and computational time. The random forest model also underscores the importance of gender and age in classifying employment status in West Java, suggesting a need for targeted intervention, especially for female citizens and individuals in productive age groups.

Keywords: Unemployment, Random Forest, Linear SVM, RBF SVM, Polynomial SVM

I. INTRODUCTION

Unemployment remains a critical challenge before, during, and after COVID-19. Now that most countries are slowly recovering from the shock COVID-19 caused to unemployment rates, it has become increasingly important to study unemployment patterns prior to extraordinary disasters such as COVID-19. Accurately identifying patterns of unemployment, particularly pre-COVID-19, will help establishing a baseline for assessing the pandemic's impact, aiding researchers and policymakers in understanding and developing strategies for economic recovery and job creation [1]. While finding patterns is typically a challenging task, the use of machine learning techniques has emerged as powerful tools in recent years.

Previous studies have successfully applied various machine learning techniques to predict unemployment. Ref. [2] utilized tree-based classification models, such as classification tree and random forest, to predict unemployment status. [3] employed logistic regression, SVM, KNN, and decision tree to predict unemployment status. Similarly, [4] utilized naïve bayes, logistic regression, SVM, random forest, and decision tree to predict employment status.

While these existing studies illustrated the high accuracy of machine learning models in predicting unemployment status in different parts of the world, there are still limited works that specifically focus on the Indonesian context, particularly the highly populated province of West Java. West Java is the second province in Indonesia with the highest percentage of open unemployment, with a 7.44% unemployment rate in August 2023, even higher than the national employment rate [5]. Therefore, understanding unemployment patterns in West Java is crucial due to its significant contribution to the total unemployment rate in Indonesia. However, a significant challenge lies in the inherent imbalance present in unemployment data, especially in identified factors such as education and training levels, as well as job search duration and access.

This study aims to identify patterns and accurately classifying unemployment status in West Java, Indonesia. We utilized data from the National Labor Force Survey (*Survei Angkatan Kerja Nasional/SAKERNAS*) conducted in August 2019 [6], focusing on the period preceding the major economic disruptions caused by the COVID-19 pandemic. This allows us to establish a baseline understanding of unemployment pre-pandemic, which is especially relevant now during the recovery stage post-pandemic [7].

While the target variables, employment status, is only slightly imbalanced with around a 65:35 proportion, some predictors exhibit severely imbalanced proportions with more than 90% in the majority class. Such data tends to be biased to the majority class, resulting in

misclassification and poorly performed classification model [8]-[9].

Our primary objective is to compare the performance of four machine learning models: Random Forest [10], Linear SVM [11], RBF SVM [12], and Polynomial SVM [13] in predicting employment status. The evaluation will be based on their classification accuracy and computing time, considering the imbalanced nature of the data.

II. METHOD

The data utilized in this study comprises the unemployment rates in West Java, obtained from the National Labor Force Survey (*Survei Angkatan Kerja Nasional / SAKERNAS*) conducted in August 2019. The data covers various districts in West Java with a total of 54,429 respondents, consisting of 8 variables as detailed in TABLE I.

The predictor variables were selected based on indications from previous studies that these variables impact employment status. Ref. [14]-[16] concluded that education and gender affect the employment status of the educated population in Indonesia. Other studies by [17]-[19] highlighted age as factor of unemployment. Ref. [20]-[21] also found a correlation between education or training levels and unemployment duration, indicating that professional degree holders experience longer unemployment durations, especially among first-time job seekers. Lastly, [22]-[23] concluded that the probability of finding a job is also influenced by the intensity of job search.

TABLE I
LIST OF VARIABLES

Symbols	Variable	Description
Y	Employment status	Binary (0 = Employed, 1 = Unemployed)
X_1	Gender	1 = Male, 0 = Female
X_2	Age	Years
X_3	Education level	Length of school (in years)
X_4	Attending course	1 = Yes, 0 = No
X_5	Active looking for	1 = Yes, 0 = No
X_6	Job search duration	Months
X_7	District	Names of Districts in West Java

The following stages outline the analysis in this study. The overall flowchart of these stages is illustrated in Fig. 1.

1) Data cleaning and exploratory data analysis: Data cleaning is a preprocessing stage which aimed at fixing or removing corrupted, missing, or duplicated data, transforming incorrectly formatted data, and handling outliers. Exploratory data analysis (EDA) is another preprocessing stage which aimed at summarizing and describing the data's characteristics, including measures of central tendency, dispersion, and graphical representations. The EDA also filters noise that may persist after the data cleaning stage. The final output of this step is a set of clean data that will be used for the next stages.

2) Model training: We employ holdout validation during the evaluation stage, with a split of 70% training and 30% testing data. Thus, the data needs to be stratified before training the predictive models to ensure similar proportions of employed and unemployed respondents in both training and testing data sets. The same training data are then used to fit the four models: Random Forest, Linear SVM, RBF SVM, and Polynomial SVM. Each model produces a predictive model for use in the subsequent stage.

3) Predicting test data: The predictive models from the previous stage are used to predict the test data. Predictions from each model are stored and the training stage is repeated. Steps 2 and 3 are repeated for 100 times. To ensure distinct train-test datasets for each repetition, steps 2 and 3 utilized randomized holdout validation. For reproducibility, a variable "i = *seed*" is used as a random number generator, where i represents the current repetition number.

4) Model evaluation and interpretation: The evaluation process involves comparing each model's predictions from each repetition with the actual employment status in the testing data. We utilized two metrics: balanced accuracy and F1-Score. Computation time is also measured and compared. The best model exhibits the highest balanced accuracy and F1-Score but with relatively lower computation time. This best model will then be used to determine the most important predictors for employment status.

III. RESULT AND DISCUSSION

A. Data Cleaning and Exploratory Data Analysis

The raw data used in this study was relatively clean, with no corrupted, missing, or duplicated entries. However, outliers were observed in the "job search duration" and "age" variables, where some respondents reported searching for jobs for up to 84 months. These outliers were managed by categorizing the variable, as can be seen in Fig. 2 and Table II. The primary challenge

identified was the imbalanced proportions in several variables, namely, school duration, attending course, job search duration, and actively searching for a job.

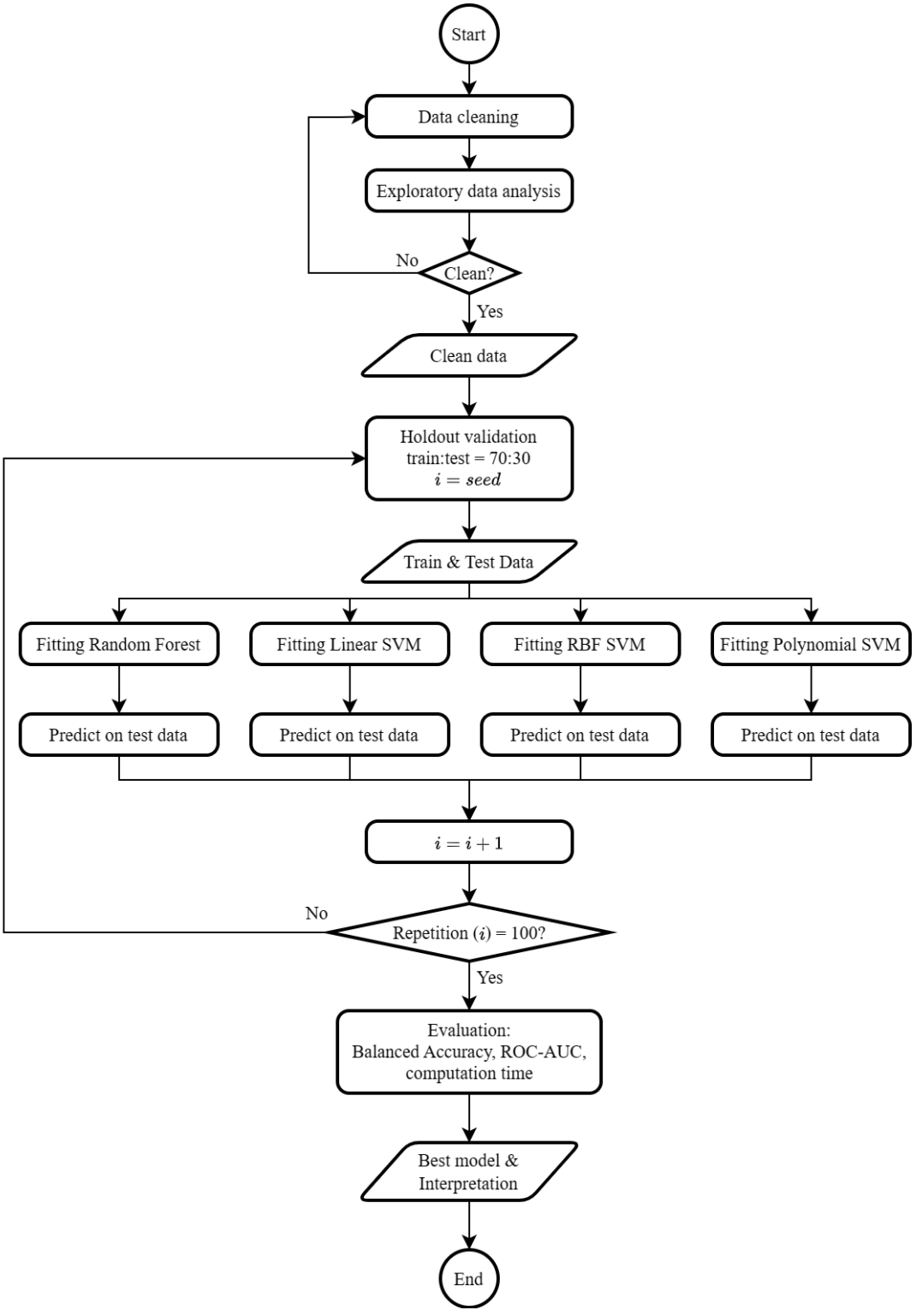

Fig. 1 Study flowchart

Fig. 2 illustrates the number of respondents based on variables. Approximately one-third of all respondents are unemployed. Gender distribution is relatively balanced, with each gender accounting for around 50% of respondents. Age groups are also relatively balanced, ranging from 5% to 12% of the total respondents. While district is not perfectly balanced, it aligns with the population distribution of each district.

However, some variables exhibit unbalanced proportions across their classes. Around 90% of respondents have an education level of high school or below, and a similar proportion did not attend any courses or training. On a positive side, approximately 90% of the respondents are the first-time job seekers and/or actively searching for a job.

These four variables display similar proportions, with the majority class or classes representing about 90% of all respondents. This is due to their strong association as depicted in Fig. 3. Other variables also show strong associations with each other, except for District.

One notable example of variable association is between school duration and training/course status. Fig. 4 shows that respondents with a high school diploma or below, i.e. respondents with 12 years or less school duration, are mostly not attending any courses.

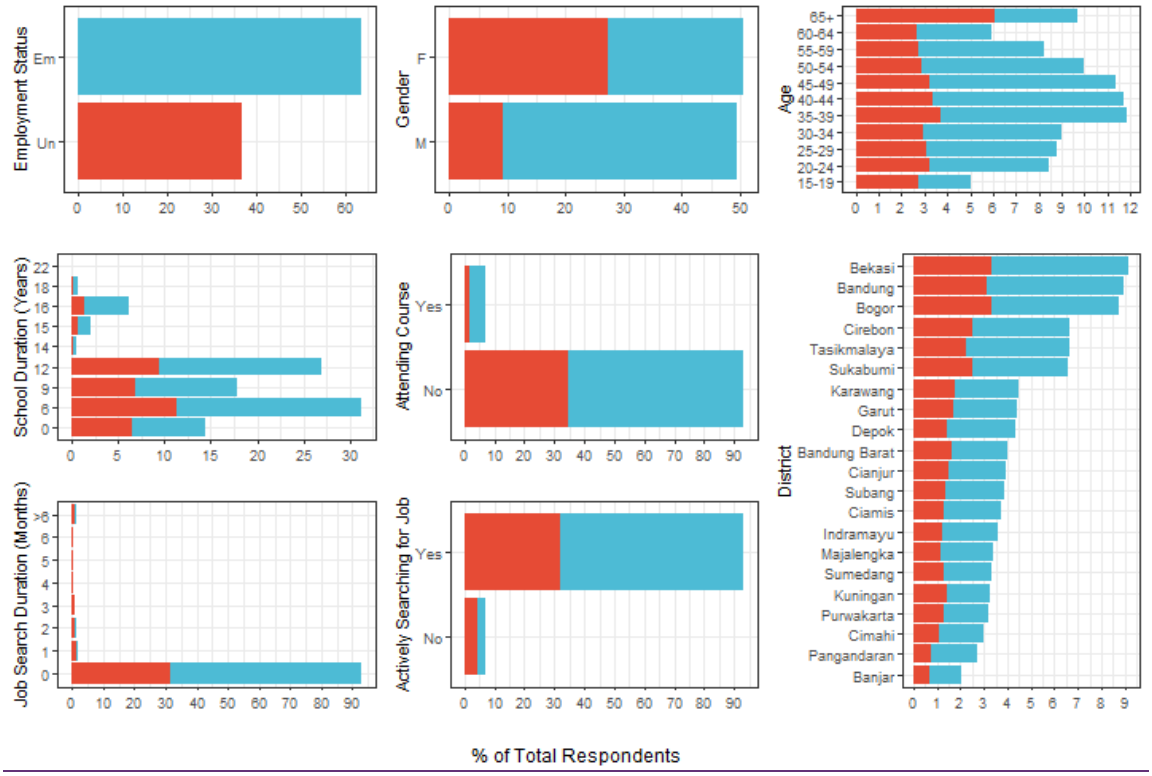

Fig. 2 Number of respondents based on variable

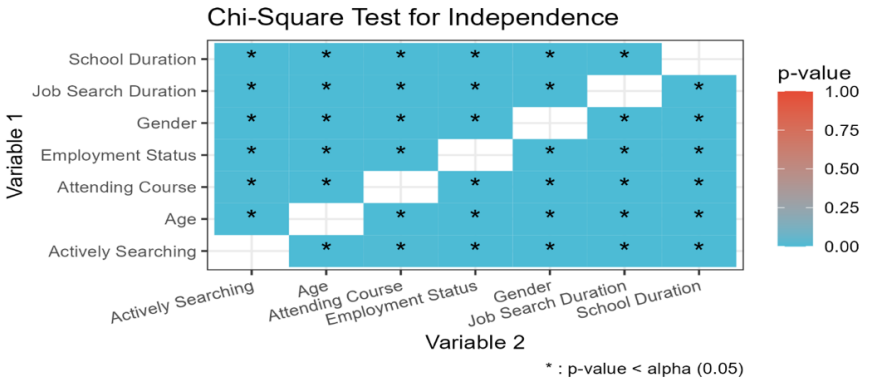

Fig. 3 Chi-Square test for independence

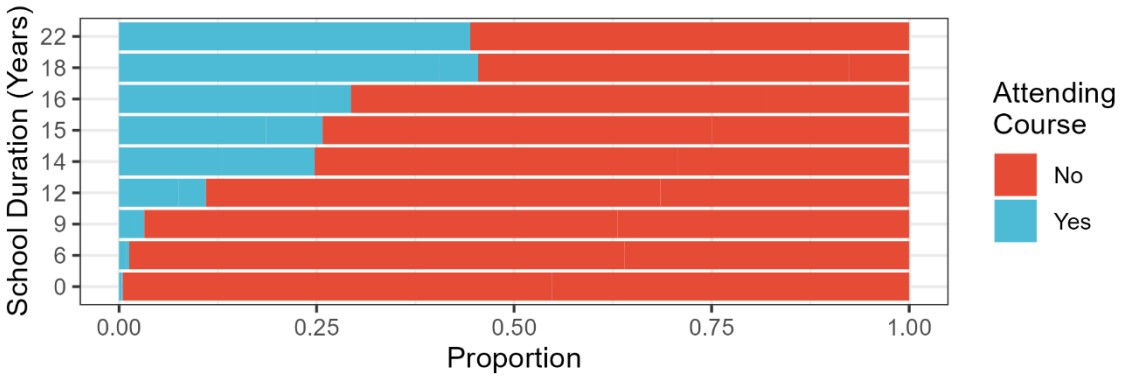

Fig. 4 Proportion of respondents based on school duration and training

Despite strong associations among variables, the data is well-diversified across demographic with no dominant group in the data. Ignoring the district location, the three most frequent groups each represent less than 2% of the entire dataset, as detailed in Table II.

As outlined in the methods section, the data were split into train-test sets with 100 repetitions. Each model was trained to fit a predictive model using the same train data across repetitions. Fig. 5 displays the distribution of balanced accuracy for each model.

B. Model Training, Predicting, and Evaluation

TABLE II
MOST FREQUENT GROUP IN THE DATA

Y	X_1	X_2	X_3	X_4	X_5	X_6	Number of Respondents	Proportion to the Whole Data
Employed	1 (Male)	45-49	6	0 (No)	1 (Yes)	0	962	1.77%
Unemployed	0 (Female)	65+	0	0 (No)	1 (Yes)	0	961	1.77%
Employed	1 (Male)	40-44	6	0 (No)	1 (Yes)	0	946	1.74%

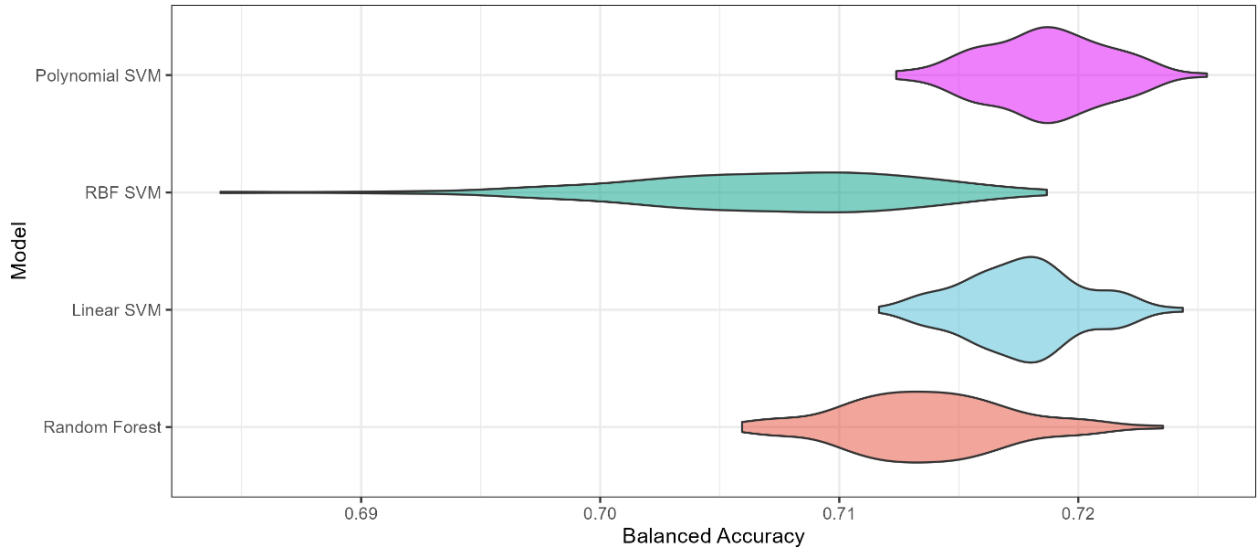

Fig. 5 Distribution of balanced accuracy for each model of the 100 repetitions

Fig. 5 indicates that the RBF SVM model has the lowest balanced accuracy among the four models. Random forest demonstrates better balanced accuracy, but falls short of the linear and polynomial SVMs. The linear and polynomial SVMs exhibit very similar balanced accuracy, with the polynomial SVM slightly edging ahead. Fig. 6 further confirms the slight superiority of the polynomial SVM over the linear SVM if based on balanced accuracy.

Different results emerge when evaluating models using F1-score, as shown in Fig. 7. Random forest and RBF SVM models achieve the highest F1-scores, contrasting with their lower balanced accuracy. Although, it is important to note that random forest's balanced accuracy is closer to the linear and polynomial SVMs' than it is to the RBF SVM's.

Since the F1-score focuses on positive class, i.e. the employed class, the high balanced accuracy but low F1-score in linear and polynomial SVMs can be interpreted as the models performing well in terms of accuracy across classes, but performing poorly in terms of identifying the employed class. It means that, despite the class imbalance, these two models are performing relatively well including in minority class. On the other hand, the RBF SVM has low balanced accuracy but high F1-score, indicating that the RBF SVM is performing well in majority (employed) class, but not performing well in minority class. Of the four compared models, the random forest is the most well-rounded model since the balanced accuracy is only slightly lower than the linear and polynomial SVM, but has the highest F1-score among the models, indicating that the random forest model is performing well on both cases, either focusing across classes, or focusing on the majority (employed) class.

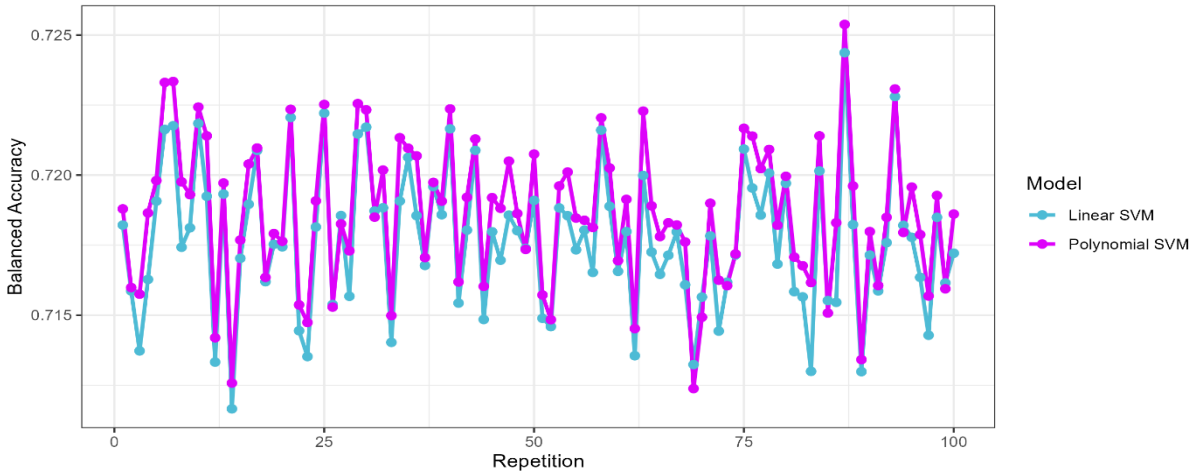

Fig. 6 Comparison of linear and polynomial SVMs' balanced accuracy for each repetition

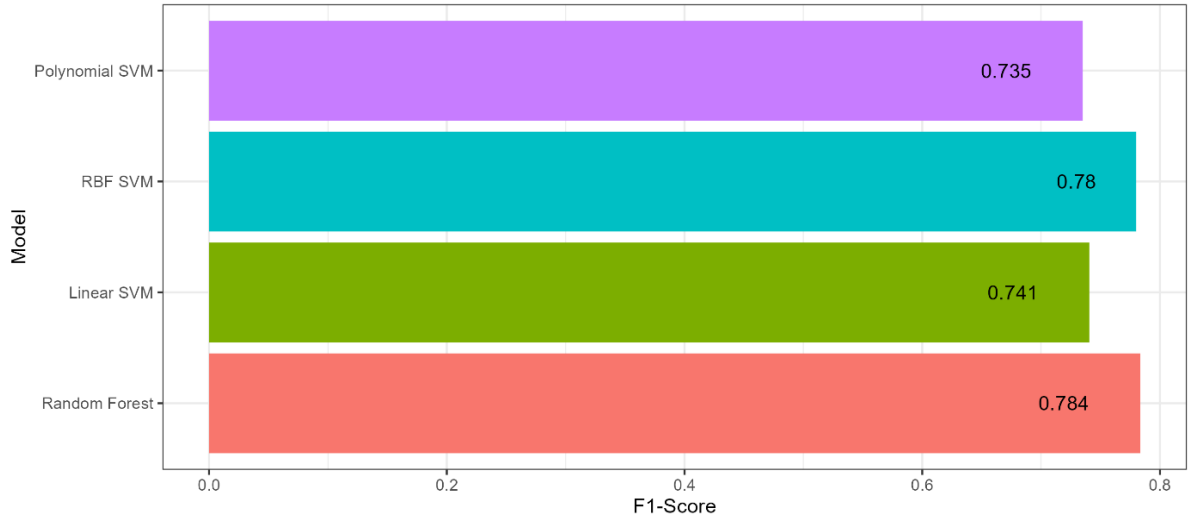

Fig. 7 Average F1-Score for each model

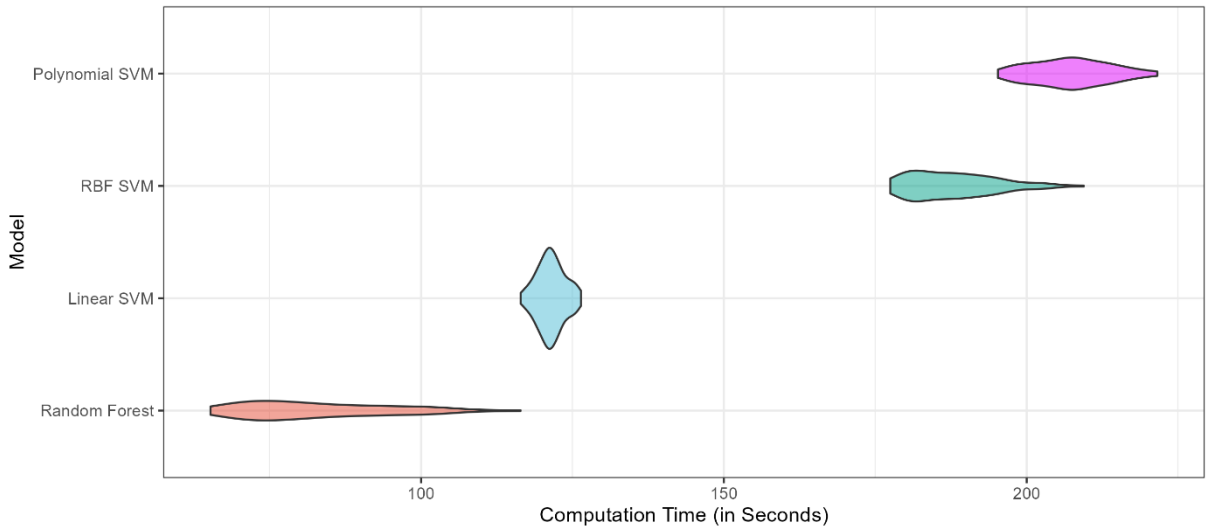

Fig. 8 Distribution of computation time in seconds

Fig. 8 highlights the computation time for the four models across the 100 repetitions. Random forest emerges as the fastest model, followed by the linear SVM with much more consistent computation time. Meanwhile, the RBF and polynomial SVM requires roughly twice the computation time of random forest.

Considering its well-rounded performance in balanced accuracy, F1-score, and faster computation time, random forest emerges as the more suitable model for classifying unemployment data in West Java.

C. Model Interpretation

Based on the evaluation results, random forest model is used to interpret the overall unemployment data in West Java. Fig. 9 showcases the variable importance in the resulting random forest model.

Fig. 9 shows that the gender and age are the two most important predictors in predicting employment status in West Java. Fig. 10 below shows these more clearly.

Fig. 10 shows that male respondents are more likely to be employed indicated by the 70% employed male respondents. Meanwhile, female respondents are more likely to be unemployed indicated by the 65% unemployed female respondents.

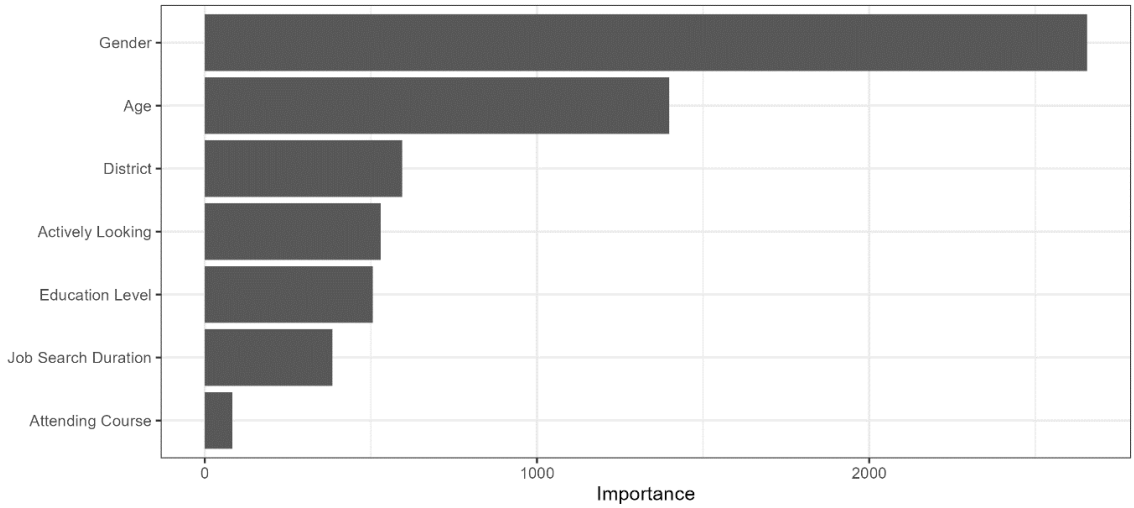

Fig. 9 Variable Importance of unemployment data in West Java

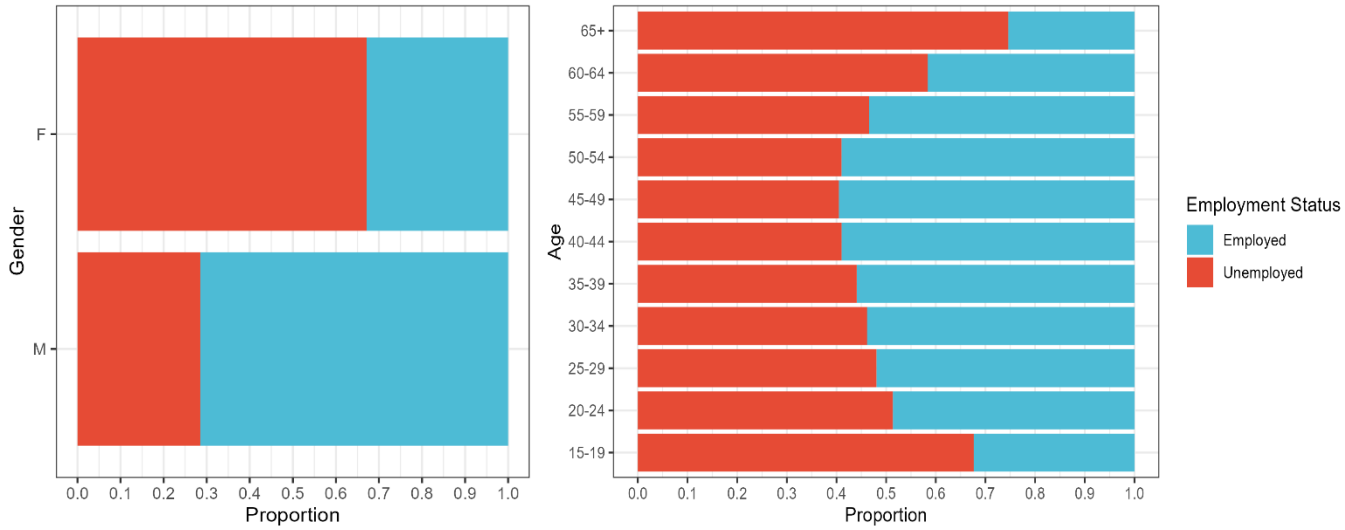

Fig. 10 Proportion of gender (left) and age group (right) based on employment status

Fig. 10 also shows that adolescent (15-19 years old) and retiree (65+ years old) are less likely to be employed, indicated by around 60%-70% unemployed respondents in these age groups. The younger retiree at the age group of 60-64 also has high unemployment rate, with around half of the respondents are unemployed. Meanwhile, the more productive age groups are generally more likely to be employed, especially in the older productive age groups. Respondents at the age between 40-54 years old shows the highest employment rate since around 60% of respondents in this age group is employed.

IV. CONCLUSION

Among the four machine learning models evaluated in this study, Random Forest is the most suitable model for classifying unemployment data in West Java due to its well-rounded performance in balanced accuracy and F1-score, as well as its much faster computation time compared to Linear, RBF, and Polynomial SVMs. The random forest model not only demonstrates high accuracy across classes, indicated by its 71%-72% balanced accuracy, but also efficiently identifies the majority (employed) class, indicated by the 78.4% F1-score. Moreover, the faster computation time makes random forest more well-suited for large-scale data processing. Evaluation also reveals the most important variables in classifying employment status in West Java. Gender emerges as pivotal predictors with male respondents showing higher likelihood of employment. Age also plays a pivotal role in determining employment status with respondents in adolescent (15-19 years old) and retirees (60+ years old) age groups are less likely to be employed. Integrating anomaly detection algorithms such as Isolation Forest in future studies may further enhance the prediction of unemployment status in West Java.

REFERENCES

[1] R. Layard and J.-E. De Neve, "Unemployment," *Wellbeing*, pp. 166–177, Mar. 2023, doi: 10.1017/9781009298957.015.

[2] M. G. Celbiş, "Unemployment in Rural Europe: A Machine Learning Perspective," *Appl Spat Anal Policy*, vol. 16, no. 3, 2023, doi: 10.1007/s12061-022-09464-0.

[3] M. Sen, Shreya Basu, Arijit Chatterjee, Anwesha Banerjee, Saheli Pal, Pritam Kumar Mukhopadhyay, Stobak Dutta, Arunabha Tarafdar, "Prediction of Unemployment using Machine Learning Approach," in *Proceedings - 2022 OITS International Conference on Information Technology, OCIT 2022*, 2022. doi: 10.1109/OCIT56763.2022.00072.

[4] O. Awujoola, Philip O Odion, Martins E Irhebhude, and Halima Aminu, "Performance Evaluation of Machine Learning Predictive Analytical Model for Determining the Job Applicants Employment Status," *Malaysian Journal of Applied Sciences*, vol. 6, no. 1, 2021, doi: 10.37231/myjas.2021.6.1.276.

[5] Badan Pusat Statistik, "Tingkat Pengangguran Terbuka Menurut Provinsi (Persen), 2023," Nov. 2023. Accessed: Mar. 03, 2024. [Online]. Available: https://www.bps.go.id/id/statistics-table/2/NTQzIzI=/tingkat-pengangguran-terbuka--agustus-2023.html

[6] Badan Pusat Statistik, "Booklet Agustus 2019 Survei Angkatan Kerja Nasional," 2019.

[7] "Planning a Sustainable Post-Pandemic Recovery in Latin America and the Caribbean," in *The Socio-Economic Implications of the COVID-19 Pandemic*, 2021. doi: 10.18356/9789210055390c014.

[8] A. J. Mohammed, "Improving Classification Performance for a Novel Imbalanced Medical Dataset using SMOTE Method," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 3, 2020, doi: 10.30534/ijatcse/2020/104932020.

[9] D. Brzezinski, L. L. Minku, T. Pewinski, J. Stefanowski, and A. Szumaczuk, "The impact of data difficulty factors on classification of imbalanced and concept drifting data streams," *Knowl Inf Syst*, vol. 63, no. 6, 2021, doi: 10.1007/s10115-021-01560-w.

[10] M. N. Wright and A. Ziegler, "Ranger: A fast implementation of random forests for high dimensional data in C++ and R," *J Stat Softw*, vol. 77, no. 1, 2017, doi: 10.18637/jss.v077.i01.

[11] M. Azimi-Pour, H. Eskandari-Naddaf, and A. Pakzad, "Linear and non-linear SVM prediction for fresh properties and compressive strength of high volume fly ash self-compacting concrete," *Constr Build Mater*, vol. 230, 2020, doi: 10.1016/j.conbuildmat.2019.117021.

[12] J. Bao, J. Nie, C. Liu, B. Jiang, F. Zhu, and J. He, "Improved blind spectrum sensing by covariance matrix cholesky decomposition and RBF-SVM decision classification at low SNRs," *IEEE Access*, vol. 7, 2019, doi: 10.1109/ACCESS.2019.2929316.

[13] S. K. Lee, J. H. Shin, J. Ahn, J. Y. Lee, and D. E. Jang, "Identifying the risk factors associated with nursing home residents' pressure ulcers using machine learning methods," *Int J Environ Res Public Health*, vol. 18, no. 6, 2021, doi: 10.3390/ijerph18062954.

[14] A. Salim, "Karakteristik Tenaga Kerja dan Pertumbuhan Ekonomi Terhadap Pengangguran Tenaga Kerja Terdidik di Indonesia," *Jurnal Ekonomi-Qu*, vol. 13, no. 1, 2023, doi: 10.35448/jequ.v13i1.20534.

[15] A. Fakih, N. Haimoun, and M. Kassem, "Youth Unemployment, Gender and Institutions During Transition: Evidence from the Arab Spring," *Soc Indic Res*, vol. 150, no. 1, 2020, doi: 10.1007/s11205-020-02300-3.

[16] M. Ryczkowski and M. Zinecker, "Gender unemployment in the Czech and Polish labour market," *Argumenta Oeconomica*, vol. 2020, no. 2, 2020, doi: 10.15611/aoe.2020.2.09.

[17] L. B. Strober and R. M. Callanan, "Unemployment in multiple sclerosis across the ages: How factors of unemployment differ among the decades of life," *J Health Psychol*, vol. 26, no. 9, 2021, doi: 10.1177/1359105319876340.

[18] R. Mulero and A. Garcia-Hiernaux, "Forecasting unemployment with Google Trends: age, gender and digital divide," *Empir Econ*, vol. 65, no. 2, 2023, doi: 10.1007/s00181-022-02347-w.

[19] A. Manzoni and I. Mooi-Reci, "The cumulative disadvantage of unemployment: Longitudinal evidence across gender and age at first unemployment in Germany," *PLoS One*, vol. 15, no. 6, 2020, doi: 10.1371/journal.pone.0234786.

[20] S. K. Jwasshaka and N. Fadila, "Minimizing Unemployment of Graduates through Technical Education and Training: Meta-Analysis Approach in Nigeria," *International Journal of Academic Research in Business and Social Sciences*, vol. 10, no. 2, 2020, doi: 10.6007/ijarbss/v10-i2/6858.

[21] R. D. P. Loka and P. A. P. Purwanti, "THE EFFECT OF UNEMPLOYMENT, EDUCATION AND THE NUMBER OF POPULATION ON THE POVERTY LEVEL OF REGENCY/CITY IN BALI PROVINCE," *International Journal of Economics, Business and Accounting Research (IJEBAR)*, vol. 6, no. 2, 2022, doi: 10.29040/ijebar.v6i2.5357.

[22] E. A. J. van Hooft, J. D. Kammeyer-Mueller, C. R. Wanberg, R. Kanfer, and G. Basbug, "Job search and employment success: A quantitative review and future research agenda.," *Journal of Applied Psychology*, vol. 106, no. 5, 2021, doi: 10.1037/apl0000675.

[23] C. R. Wanberg, A. A. Ali, and B. Csillag, "Job Seeking: The Process and Experience of Looking for a Job," *Annual Review of Organizational Psychology and Organizational Behavior*, vol. 7. 2020. doi: 10.1146/annurev-orgpsych-012119-044939.