JOIV : Int. J. Inform. Visualization, 7(4) - December 2023 2154-2160

INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

Introversion-Extraversion Prediction Using Machine Learning

Brillian Fieri [a], Joshua La'la [a], Derwin Suhartono [b,*]

[a] Computer Science Department, BINUS Graduate Program, Bina Nusantara University, Palmerah, Jakarta, 11480, Indonesia
[b] Computer Science Department, School of Computer Science, Bina Nusantara University, Palmerah, Jakarta, 11480, Indonesia
Corresponding author: *dsuhartono@binus.edu

Abstract*—Introversion and extroversion are personality traits that assess the type of interaction between people and others. Introversion and extraversion have their advantages and disadvantages. Knowing their personality, people can utilize these advantages and disadvantages for their benefit. This study compares and evaluates several machine learning models and dataset balancing methods to predict the introversion-extraversion personality based on the survey result conducted by Open-Source Psychometrics Project. The dataset was balanced using three balancing methods, and fifteen questions were chosen as the features based on their correlations with the personality self-identification result. The dataset was used to train several supervised machine-learning models. The best model for the Synthetic Minority Oversampling (SMOTE), Adaptive Synthesis Sampling (ADASYN), and Synthetic Minority Oversampling-Edited Nearest Neighbor (SMOTE-ENN) datasets was the Random Forest with the 10-fold cross-validation accuracy of 95.5%, 95.3%, and 71.0%. On the original dataset, the best model was Support Vector Machine, with a 10-fold cross-validation accuracy of 73.5%. Based on the results, the best balancing methods to increase the models' performance were oversampling. Conversely, the hybrid method of oversampling-undersampling did not significantly increase performance. Furthermore, the tree-like models, like Random Forest and Decision Tree, improved performance substantially from the data balancing. In contrast, the other models, excluding the SVM, did not show a significant rise in performance. This research implies that further study is needed on the hybrid balancing method and another classification model to improve personality classification performance.

***Keywords*— Imbalanced dataset; introversion-extraversion; machine learning; personality prediction.**

Manuscript received 7 Jul. 2022; revised 26 Dec. 2022; accepted 30 Jan. 2023. Date of publication 31 Dec. 2023.

I. INTRODUCTION

Personality denotes the characters of a person and their differences from the rest of society [1]. Each person's personality is different and comprises various personality traits [2]. These different personality characteristics, such as behavior, can be caused by environmental factors [3]. Moreover, the personality can change based on the events in the surrounding environment or an individual's eagerness to change [4].

Extroverts and Introverts are personality traits that exist today. The extrovert and introvert personality traits are present in everyone under the influence of certain situations and the surrounding environment [5]. Extroverts tend to be sociable and are likely to look outside themselves for relief, while introverts, on the other hand, are more reserved and typically quiet. Extroverts tend to focus on the outside world or external activities, for instance, meeting new people. Extroverts enjoy interactions and are delighted when there is support from those around them. Dominantly extroverted people have high social life and good teamwork skills. Conversely, introverts tend to do activities that are likely directed inward themselves. It appears that introverts prefer to work independently as it is way more convenient and practical. Generally, introverts enjoy solitude or prefer no interactions with other people. However, most introverted people are capable of solid reflection and are blissful to working independently.

A study stated that finding people who are purely extroverted or introverted is extremely rare [5]. Introversion-extraversion personality can be measured by whether the person is sociable or adaptable to any situation and the surroundings and how they acquire their energy. The external factors, such as interaction with other people, are the ones that give extroverts energy, while introverts obtain it through internal factors, like when they are in solitude [6].

Knowing their personality can help people to acknowledge their strengths and weaknesses. For example, extroverts can easily be affected by emotion when facing a personal dilemma, while introverts usually can keep their sensible judgment [7]. In terms of being a leader in a company or organization, extroverts have a massive opportunity in terms of being leaders [8], because extroverts are more comfortable working in teams, and it is easy to break the ice in a team, whether in a tense or relaxed state [9]. Introverts are not a good advantage of being a leader because introverts prefer to work alone and prefer to interact less. The advantages and disadvantages can be used for their gain from knowing their personality and traits. Consequently, this study aimed to find the best method for predicting the introversion-extraversion personality.

The extraversion-introversion detection can help in a few aspects of life. For example, this can be used to recommend specific treatments for the employees based on their personality to boost their work performance. Another example of the application is to determine the best learning method for students based on their extraversion-introversion personality.

Previous studies on extrovert and introvert prediction have been done multiple times. A previous study compared several prediction models for the introversion-extraversion personality, with the best model gaining 73.81% accuracy [10]. The dataset used in the previous work was obtained from the Multidimensional Introversion-Extroversion Scales assessment result by the Open-Source Psychometrics Project and augmented using oversampling methods, which will also be used in this study. However, what differs from this research is that this study used an additional hybrid augmentation method between oversampling and undersampling, SMOTE-ENN. It also used various supervised machine learning models such as Decision Tree, Logistic Regression, k-NN, Linear Discriminant Analysis, Gaussian Naïve Bayes, Random Forest, SVM Linear, SVM Polynomial, and SVM Gaussian.

An experiment to predict extraversion also had been conducted by a study [11]. Electrocardiographic and NEO-FFI data were used to train the Random Forest model, which resulted in 60.6% accuracy in predicting the extroverts and introverts. Another study used unmentioned campus' data to predict their students' introversion or extraversion, which obtained an accuracy of 72% by using linear SVM [12]. In addition, another study experimented with predicting introversion and extraversion based on the interaction of the subjects with a robot [13]. The result was 70% accuracy on extraversion prediction.

This study compares and evaluates machine learning models and dataset balancing methods to improve introversion-extraversion personality classification performance. Several machine learning models such as Decision Tree, K-Nearest Neighbor, Logistic Regression, Linear Discriminant Analysis, Gaussian Naïve Bayes, Random Forest, and Support Vector Machine were trained and evaluated for each original, oversampling methods such as SMOTE and ADASYN, and a hybrid between oversampling and undersampling dataset, SMOTE-ENN.

II. Material and Method

The method used by the researchers is to conduct an experiment in which to identify one's personality. The Open-Source Psychometrics Project provided the dataset in 2019 [14]. Fig. 1 shows that the dataset was executed in several stages to facilitate the experiment to identify one's personality.

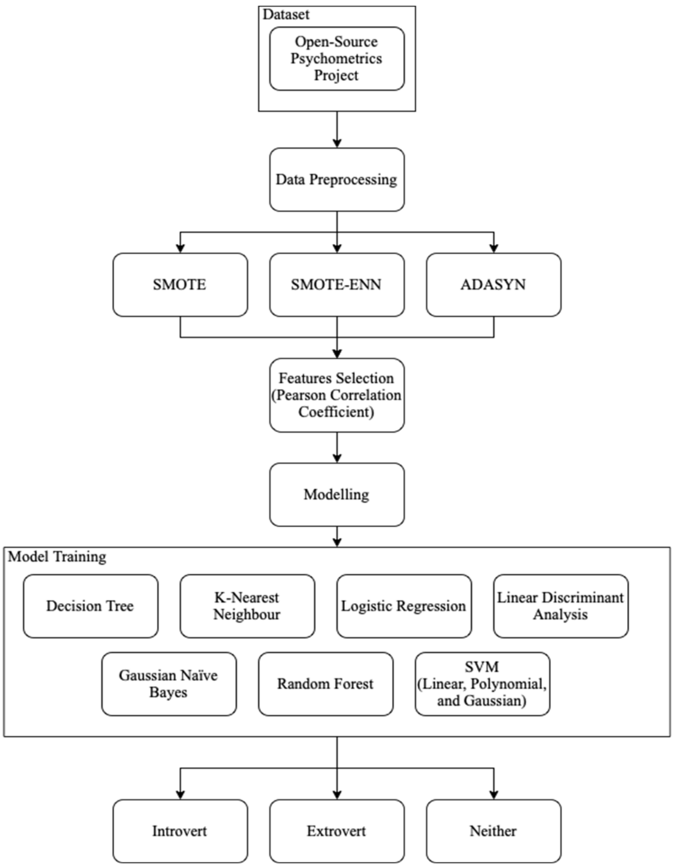

Fig. 1 Workflow of study

A. Dataset

The dataset used in this experiment was provided by Open-Source Psychometrics Project [14]. The data was taken from an online survey called the Multidimensional Introversion-Extroversion Scales on their website. The published result of the survey was last updated on August 19, 2019, with a total of 7,188 responses.

According to the survey, 91 statements require the participants to submit an answer using a five-point scale starting from 1, which represents disagreement, to 5, which represents an agreement, also known as the Likert scale. Moreover, the participant was asked to determine to which category they belonged. The dataset contained the responses, times needed, each statement's position, introversion-extraversion self-identification, and personal data, e.g., country, gender, and age. The data taken for the experiment are the responses of each statement, the introversion-extraversion identification (IE), country, gender, and age.

B. Pre-processing

The first step of pre-processing the dataset was to remove the invalid data. This data includes null and out-of-scope data values. Afterward, the IE data was cleansed from the dataset, excluding the 1 (Introvert), 2 (Extrovert), and 3 (Neither). Furthermore, the dataset was balanced using several re-sampling methods. This step was necessary because the original dataset was imbalanced, as seen in Fig. 2. The class data comprised 4,404 introverts, 989 extroverts, and 1,768 neither. The imbalanced dataset can make the machine learning model more biased toward the majority class [15]. This can result in overfitting and a lousy performance in classifying the minority class. Re-sampling can be applied to the imbalanced dataset to achieve a balanced dataset. This process consists of oversampling and undersampling. In oversampling, new data is produced to increase the number of minority classes, while in undersampling, the majority class data is decreased to match with the minority class [16]. In this study, three re-sampling methods were used to balance the dataset.

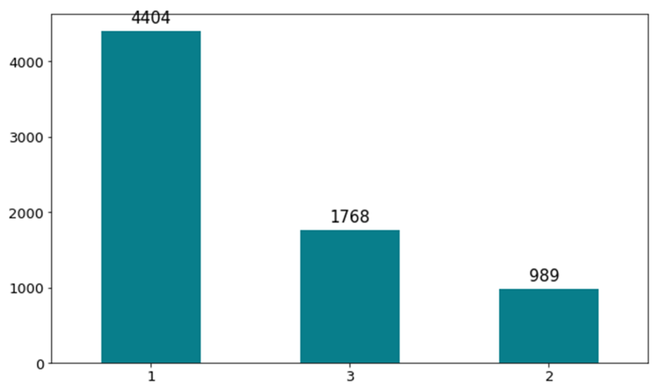

Fig. 2 Imbalanced Original Dataset [14]

The first method was the oversampling method, Synthetic Minority Oversampling (SMOTE). SMOTE is one of the most popular re-sampling methods [17]. This method refers to the synthesis of the minority class. Therefore, the sum of the minority class will be equal to the number of the majority class by synthesizing new data for the minority class [18]. The result of SMOTE in Fig. 3 showed that new data were added for all the minority classes to match the amount of the majority class data. The result was that each of the classes had 4,404 data.

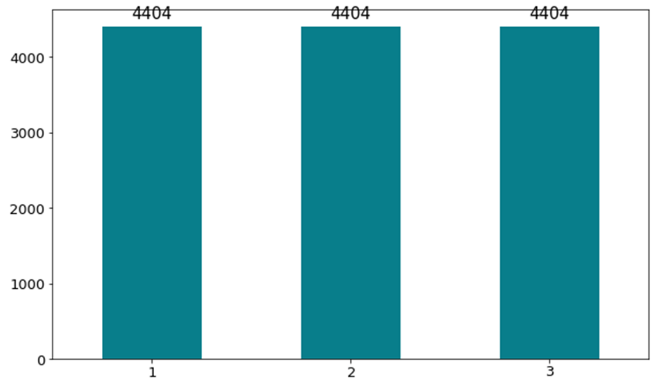

Fig. 3 SMOTE Dataset

The second method that can balance a dataset is Adaptive Synthesis Sampling (ADASYN). This method helps the classification process by generating more data for the unbalanced minority classes [19]. The amount of generated data for each class is based on the weighted distribution depending on the level of learning difficulties [19]. The result of the ADASYN dataset can be seen in Fig. 4.

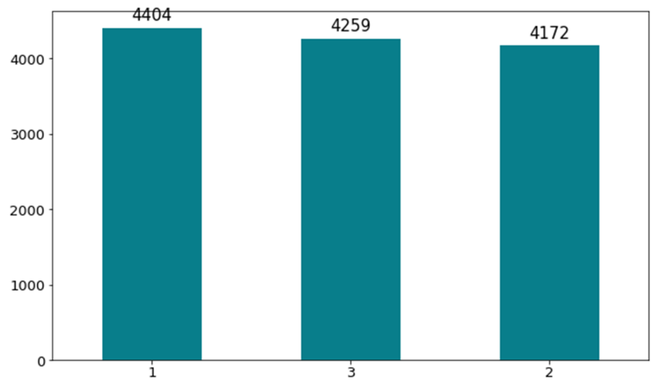

Fig. 4 ADASYN Dataset

There is also the SMOTE-ENN method, which combines SMOTE and Edited Nearest Neighbor (ENN). The ENN method works by finding the K-nearest neighbor, starting by taking samples based on the category of its nearest neighbor, then using the k-NN rules to the rest of the data [20]. If there is a minority class data with two or more majority class data as its neighbor, the majority class data will be deleted. Thus, the distance between the majority and the minority data will be reduced. The result of SMOTE-ENN can be seen in Fig. 5.

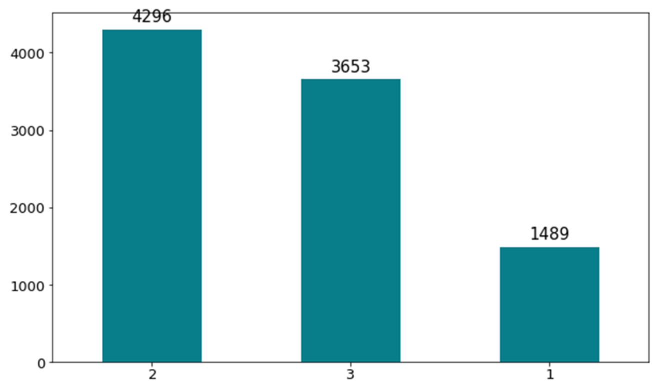

Fig. 5 SMOTE-ENN Dataset

C. Feature Selection

The feature selection was made by calculating the correlation between the features and the output label on the original dataset using the Pearson correlation coefficient. The Pearson correlation coefficient is one of the popular methods to calculate the correlation between two variables. This method resulted in a value with a range between -1 and 1, where the closer the value to 1, the higher the correlation between values, and the same thing applied to -1. However, the correlation is in the opposite direction [21].

For this study, the results of the Pearson correlation coefficient were turned into absolute values. The correlation list was sorted from the highest to the lowest correlation values, as seen in Table I. Thereafter, the top 15 features were selected and stored for the next section of the experiment.

TABLE I
RESULT OF SORTED CORRELATION FEATURES

Question [14]	Correlation Score (Absolute)
Q83A: "I keep in the background."	0.412
Q91A: "I talk to a lot of different people at parties."	0.396
Q82A: "I don't talk a lot."	0.394
Q90A: "I start conversations."	0.366
Q80A: "I love large parties."	0.347
Q89A: "I don't mind being the center of attention."	0.340
Q81A: "I am quiet around strangers."	0.340
Q84A: "I don't like to draw attention to myself."	0.324
Q14A: "I want a huge social circle."	0.309
Q13A: "I can keep a conversation going with anyone about anything."	0.295
Q5A: "I mostly listen to people in conversations."	0.293
Q44A: "I mostly listen to people in conversations."	0.288
Q16A: "I act wild and crazy."	0.269
Q15A: "I talk to people when waiting in lines."	0.267
Q85A: "I have little to say."	0.266

D. Modeling

The features would be separated from its label for the selected 15 features on the original, SMOTE, SMOTE-ENN, and ADASYN datasets. The dataset will be used to train supervised machine learning models, e.g., Decision Tree, Logistic Regression, k-NN, Linear Discriminant Analysis, Gaussian Naïve Bayes, Random Forest, SVM Linear, SVM Polynomial, and SVM Gaussian. The models were evaluated using 10-fold cross-validation.

III. RESULTS AND DISCUSSION

A. Results

This section will display the accuracy, precision, recall, and F1 score of all models for original, SMOTE, SMOTE-ENN, and ADASYN datasets. There would also be an overall performance summary for each dataset and comparisons between the best-performing model on one dataset and the others. Moreover, the confusion matrix for the best model on every dataset will be displayed.

1) Evaluation Result of Original Dataset: Based on the original dataset result, as seen in Table II, it can be observed that the best result was the SVM Linear, resulting in a mean accuracy of 0.735. The second-best result was SVM Gaussian, resulting in a mean accuracy of 0.734. The results' difference was 0.001, which was insignificant.

TABLE II
ORIGINAL DATASET RESULT

Method	Accuracy	Precision	Recall	F1
Decision Tree	0.624	0.630	0.623	0.626
Logistic Regression	0.732	0.709	0.731	0.715
k-NN	0.698	0.665	0.698	0.672
Linear Discriminant Analysis	0.730	0.708	0.729	0.714
Gaussian Naïve Bayes	0.699	0.711	0.698	0.701
Random Forest	0.726	0.704	0.725	0.710
SVM Linear	**0.735**	**0.711**	**0.734**	**0.716**
SVM Polynomial	0.725	0.696	0.725	0.704
SVM Gaussian	0.734	0.709	0.733	0.714

Despite being the best model for the original dataset, the result for the correct prediction of each class was relatively poor. As seen in the confusion matrix in Fig. 6, the recall score for class 3 or "Neither" was 35.0%, the score for class 2 or "Extrovert" was 62.9%, and class 1 or "Introvert" was 91%, resulting in the weighted average recall score of 73.4%.

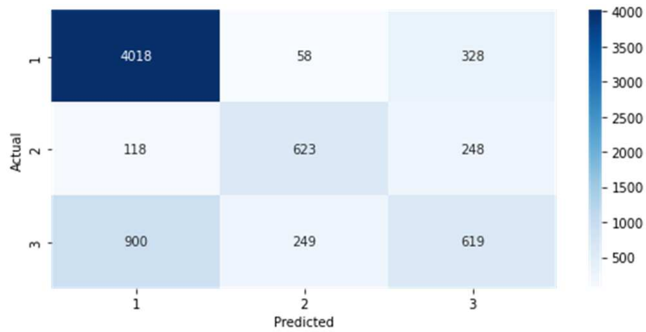

Fig. 6 SVM Linear (Original Dataset) Confusion Matrix

2) Evaluation Result of SMOTE Dataset: Table III shows that the best result for SMOTE dataset was the Random Forest, resulting in a mean accuracy of 0.955. The second-best result was Decision Tree, with a mean accuracy of 0.940. The results' difference was 0.015 or 1.5%.

The best result from the original dataset was the SVM Linear method with a mean accuracy of 0.735. In contrast, the best result for the SMOTE dataset was the Random Forest method, with a mean accuracy of 0.955. Comparing the results of the original dataset with the SMOTE dataset, the score difference between the datasets was quite significant; it stood at 0.220 or 22.0% in percentage.

The best result obtained from the original dataset was the SVM Linear method with a mean accuracy of 0.735, and applying the same method for the selected features to the SMOTE dataset obtained a mean accuracy of 0.690, whereas resulted in a 0.045 score range or 4.5% in percentage. Conversely, the best method used on the SMOTE dataset was the Random Forest method, with a mean accuracy of 0.955.

By applying the same method for the selected features to the original dataset, the mean accuracy obtained was 0.726, which resulted in a significant score range between the mean accuracy scores of both methods that stood at 0.229 or 22.9%.

TABLE III
SMOTE DATASET RESULT

Method	Accuracy	Precision	Recall	F1
Decision Tree	0.940	0.940	0.939	0.940
Logistic Regression	0.690	0.705	0.689	0.695
k-NN	0.782	0.797	0.781	0.785
Linear Discriminant Analysis	0.683	0.699	0.683	0.688
Gaussian Naïve Bayes	0.686	0.712	0.685	0.694
Random Forest	**0.955**	**0.955**	**0.955**	**0.955**
SVM Linear	0.690	0.706	0.689	0.695
SVM Polynomial	0.730	0.742	0.730	0.734
SVM Gaussian	0.736	0.744	0.736	0.739

Along with the accuracy growth for the Random Forest with SMOTE dataset, the recall score on each class also increased compared to the best model on the original dataset. The recall score for Random Forest classes 1, 2 and, 3 were 96.7%, 96.6%, 91.6%, with a weighted average score of 95.5%. Compared to the best model from the original dataset, the recall score for SMOTE was higher by 22.1%. The confusion matrix for this model can be found in Fig. 7.

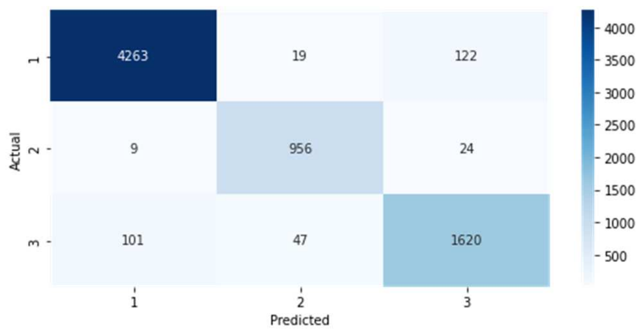

Fig. 7 Random Forest (SMOTE Dataset) Confusion Matrix

3) Evaluation Result of SMOTE-ENN Dataset: Table IV shows that the best result for the SMOTE-ENN dataset was the Random Forest, resulting in a mean accuracy of 0.710. The second-best result was Decision Tree, resulting in a mean accuracy of 0.696. The results' difference was 0.014 or 1.4%.

TABLE IV
SMOTE-ENN DATASET RESULT

Method	Accuracy	Precision	Recall	F1
Decision Tree	0.696	0.783	0.696	0.706
Logistic Regression	0.582	0.717	0.581	0.606
k-NN	0.640	0.750	0.639	0.653
Linear Discriminant Analysis	0.599	0.707	0.598	0.622
Gaussian Naïve Bayes	0.619	0.718	0.618	0.641
Random Forest	**0.710**	**0.801**	**0.710**	**0.721**
SVM Linear	0.586	0.714	0.585	0.610
SVM Polynomial	0.634	0.729	0.634	0.653
SVM Gaussian	0.630	0.735	0.629	0.649

The SMOTE and the SMOTE-ENN datasets scored the highest mean accuracy for the same method, i.e., Random Forest Method. The best result was obtained from the SMOTE dataset scoring a mean accuracy of 0.955. Furthermore, the SMOTE-ENN dataset scored 0.710 for the mean accuracy. By comparing the results of the SMOTE dataset with the SMOTE-ENN dataset, the score range of the datasets was considered significant; it stood around 0.245 or 24.5% in percentage.

The recall score for the best SMOTE-ENN dataset model was below the result of the best original dataset model with a 2.4% decrease. The best model, Random Forest, obtained 61.3%, 92.4%, and 81.2% for classes 1, 2, and 3, respectively. The model scored 71% on weighted average recall. The confusion matrix for this model can be found in Fig. 8.

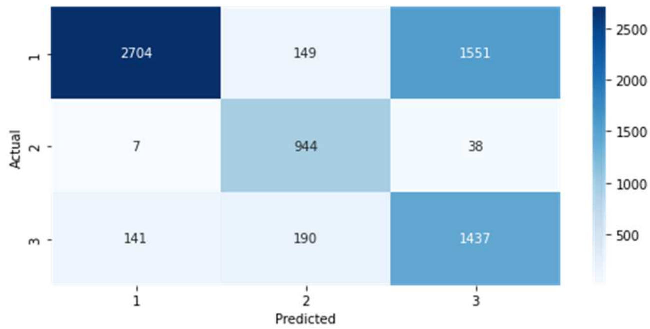

Fig. 8 Random Forest (SMOTE-ENN Dataset) Confusion Matrix

4) Evaluation Result of ADASYN Dataset: Based on the ADASYN dataset results in Table V, it can be observed that the best result was the Random Forest, resulting in a mean accuracy of 0.953. The second-best result was Decision Tree, resulting in a mean accuracy of 0.938. The results' difference was 0.015 or 1.5%.

The SMOTE-ENN and ADASYN datasets also scored the highest mean accuracy for the same method, i.e., Random Forest Method. The best result was obtained from the SMOTE-ENN dataset scoring a mean accuracy of 0.710. Moreover, the ADASYN dataset scored 0.953 for the mean accuracy. By comparing the results of the SMOTE-ENN and ADASYN datasets, the score range of the mean accuracy on both methods stood around 0.243 or 24.3% in percentage.

TABLE V
ADASYN DATASET RESULT

Method	Accuracy	Precision	Recall	F1
Decision Tree	0.938	0.939	0.938	0.938
Logistic Regression	0.684	0.696	0.683	0.685
k-NN	0.780	0.799	0.780	0.782
Linear Discriminant Analysis	0.681	0.695	0.680	0.684
Gaussian Naïve Bayes	0.676	0.706	0.675	0.683
Random Forest	**0.953**	**0.953**	**0.953**	**0.953**
SVM Linear	0.681	0.694	0.680	0.682
SVM Polynomial	0.720	0.731	0.720	0.722
SVM Gaussian	0.734	0.740	0.733	0.734

The weighted average recall score on the Random Forest on the ADASYN dataset was 95.3%, which was 0.2% lower than the best SMOTE model and 21.9% higher in contrast to the best original dataset model. The individual recall score for class 1, 2, and 3 was 96.7%, 96.6%, and 91.0%. Fig. 9 shows the confusion matrix for ADASYN Random Forest.

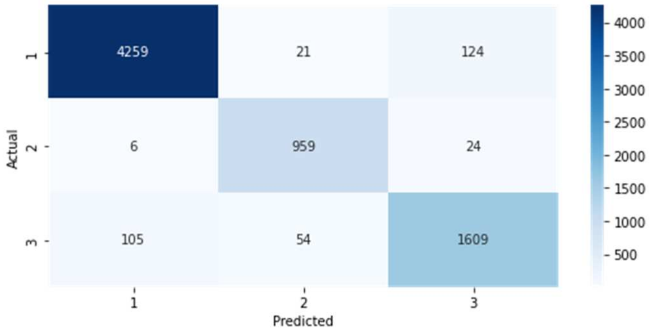

Fig. 9 Random Forest (ADASYN Dataset) Confusion Matrix

B. *Discussion*

This study aims to boost the performance of introversion-extraversion classification by comparing and evaluating various machine learning models with several balanced datasets. The discovery showed that the performance of certain models could accomplish quite a significant increase through dataset balancing operations.

The oversampling method, such as SMOTE and ADASYN, raised the F1 to 24.5% on tree-like machine learning models like Decision Tree and Random Forest. For the other models tested in this study, the increment is less significant than the tree-like models, with the biggest boost of 11.3% by k-Nearest Neighbour. Based on this research, SMOTE is superior to ADASYN on every model, though with a thin margin. The biggest gap between both was the SVM Linear model, with a 1.3% difference.

In contrast to the oversampling method, the hybrid between oversampling and undersampling, SMOTE-ENN, did not perform well in this study. The tree-like models exceeded the original dataset score, with the biggest margin of 8% by Decision Tree. The rest of the models were below the original dataset model, with the worst margin of 10.9% by Logistic Regression.

From the obtained results, this study found that from several machine learning models that were tested, the overall best-performing model for the re-sampling datasets were the tree-like models, such as Decision Tree and Random Forest. The other model, the k-Nearest Neighbor, was able to gain a boost from the balanced datasets but not as well as the tree-like model. The rest of the models were not affected by the balanced datasets. This study also perceived that all the tested sampling methods increased some of the model's performance. Although, with the hybrid method, SMOTE-ENN, only a few models gained the advantage.

IV. Conclusions

This study was conducted using an original dataset to predict the extraversion-introversion personality. Fifteen questions were chosen based on their high correlation score to the output goal using the Pearson correlation coefficient. The imbalanced dataset was balanced with oversampling techniques, like SMOTE and ADASYN, and a hybrid of oversampling and undersampling techniques, SMOTE-ENN. The study used nine classification models to find the best method to predict personality.

The best method for the SMOTE, ADASYN, and SMOTE-ENN datasets was the Random Forest with a mean accuracy of 0.955, 0.953, and 0,710. For the original dataset, the best method was SVM Linear, with a mean accuracy of 0.735.

The balancing method using oversampling, such as SMOTE and ADASYN, increased the accuracy for several models compared to the original dataset, with the biggest jump of 22.9% and 22.7% on the Random Forest model. The results show that these oversampling methods help to boost machine learning models, especially tree-like models. The SMOTE was better overall than ADASYN, although with only minor performance differences.

In contrast, the hybrid method of oversampling and undersampling, such as SMOTE-ENN, showed little growth in accuracy, with only 2 out of 9 surpassing the original dataset. The biggest boost of SMOTE-ENN was the Decision Tree compared to the original dataset, with a margin of 7.2% in mean accuracy. These outcomes concluded that the oversampling methods were better than the hybrid method for this case.

Further study could be done using other balancing methods to balance the dataset, especially with the hybrid method. Also, other classification models can be used to achieve higher accuracy in predicting personality. Additionally, the dataset balancing approach can be used to increase models' classifying performance in other research areas.

References

[1] R. M. Bergner, "What is personality? Two myths and a definition," *New Ideas Psychol.*, vol. 57, 2020, doi: 10.1016/j.newideapsych.2019.100759.

[2] P. G. Zimbardo, R. L. Johnson, and V. McCann, *Psychology : core concepts*, 8th ed. NY: Pearson, 2017.

[3] C. D. Nye and B. W. Roberts, *A neo-socioanalytic model of personality development*. Elsevier Inc., 2019.

[4] A. Baumert et al., "Integrating Personality Structure, Personality Process, and Personality Development," European Journal of Personality, vol. 31, no. 5, pp. 503–528, Sep. 2017, doi: 10.1002/per.2115.

[5] D. Petric, "Introvert , Extrovert and Ambivert," *Knot Theory Mind*, no. September, pp. 1–4, 2019, doi: 10.13140/RG.2.2.28059.41764/2.

[6] M. C. Shehni and T. Khezrab, "Review of Literature on Learners' Personality in Language Learning: Focusing on Extrovert and Introvert Learners," Theory and Practice in Language Studies, vol. 10, no. 11, p. 1478, Nov. 2020, doi: 10.17507/tpls.1011.20.

[7] Y. Tao, Y. Cai, C. Rana, and Y. Zhong, "The impact of the Extraversion-Introversion personality traits and emotions in a moral decision-making task," Personality and Individual Differences, vol. 158, p. 109840, May 2020, doi: 10.1016/j.paid.2020.109840.

[8] A. M. Grant, F. Gino, and D. A. Hofmann, "Reversing the Extraverted Leadership Advantage: The Role of Employee Proactivity," Academy of Management Journal, vol. 54, no. 3, pp. 528–550, Jun. 2011, doi: 10.5465/amj.2011.61968043.

[9] J. E. Bono and T. A. Judge, "Personality and Transformational and Transactional Leadership: A Meta-Analysis.," Journal of Applied Psychology, vol. 89, no. 5, pp. 901–910, 2004, doi: 10.1037/0021-9010.89.5.901.

[10] C. So, "Are You an Introvert or Extrovert? Accurate Classification With Only Ten Predictors," 2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Feb. 2020, doi: 10.1109/icaiic48513.2020.9065069.

[11] H. Baumgartl, S. Bayerlein, and R. Buettner, "Measuring Extraversion Using EEG Data," Lecture Notes in Information Systems and Organisation, pp. 259–265, 2020, doi: 10.1007/978-3-030-60073-0_30.

[12] L. Ge, H. Tang, Q. Zhou, Y. Tang, and J. Lang, "Classification Algorithms to Predict Students' Extraversion-Introversion Traits,"

2016 International Conference on Cyberworlds (CW), Sep. 2016, doi: 10.1109/cw.2016.27.

[13] S. M. Anzalone, G. Varni, S. Ivaldi, and M. Chetouani, "Automated Prediction of Extraversion During Human–Humanoid Interaction," International Journal of Social Robotics, vol. 9, no. 3, pp. 385–399, Feb. 2017, doi: 10.1007/s12369-017-0399-6.

[14] Open-Source Psychometrics Project, "Development of the Multidimensional Introversion-Extraversion Scales." 2019.

[15] J. Tanha, Y. Abdi, N. Samadi, N. Razzaghi, and M. Asadpour, "Boosting methods for multi-class imbalanced data classification: an experimental review," Journal of Big Data, vol. 7, no. 1, Sep. 2020, doi: 10.1186/s40537-020-00349-y.

[16] R. Mohammed, J. Rawashdeh, and M. Abdullah, "Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results," 2020 11th International Conference on Information and Communication Systems (ICICS), Apr. 2020, doi: 10.1109/icics49469.2020.239556.

[17] V. S. Spelmen and R. Porkodi, "A Review on Handling Imbalanced Data," 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT), Mar. 2018, doi:10.1109/icctct.2018.8551020.

[18] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. September 28, pp. 321–357, 2002, [Online]. Available: https://arxiv.org/pdf/1106.1813.pdf%0Ahttp://www.snopes.com/horrors/insects/telamonia.asp.

[19] H. He, Y. Bai, E. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In IEEE International Joint Conference on Neural Networks, 2008," *IJCNN 2008.(IEEE World Congr. Comput. Intell. (pp. 1322– 1328)*, no. 3, pp. 1322– 1328, 2008.

[20] T. Lu, Y. Huang, W. Zhao, and J. Zhang, "The Metering Automation System based Intrusion Detection Using Random Forest Classifier with SMOTE+ENN," 2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT), Oct. 2019, doi: 10.1109/iccsnt47585.2019.8962430.

[21] H. Zhu, X. You, and S. Liu, "Multiple Ant Colony Optimization Based on Pearson Correlation Coefficient," IEEE Access, vol. 7, pp. 61628–61638, 2019, doi: 10.1109/access.2019.2915673.