JIKO (Jurnal Informatika dan Komputer) Accredited KEMENDIKBUD RISTEK, No.105/E/KPT/2022
Vol. 6, No. 2, Agustus 2023, hlm. 144-149 p-ISSN: 2614-8897
DOI: 10.33387/jiko.v6i2.6394 e-ISSN: 2656-1948

MULTICLASS EMAIL CLASSIFICATION BY USING ENSEMBLE BAGGING AND ENSEMBLE VOTING

Ali Helmut[1], Danang T Murdiansyah[2]

[1,2]School of Computing, Telkom University
Email: [1]alihelmut@students.telkomuniversity.ac.id, [2]danangtri@telkomuniversity.ac.id

(Received: 24 July 2023, Revised: 29 July 2023, Accepted: 3 August 2023)

Abstract

Email is a common communication technology in modern life. The more emails we receive, the more difficult and time consuming it is to sort them out. One solution to overcome this problem is to create a system using machine learning to sort emails. Each method of machine learning and data sampling result in different performance. Ensemble learning is a method of combining several learning models into one model to get better performance. In this study we tried to create a multiclass email classification system by combining learning models, data sampling, and several data classes to obtain the effect of Ensemble Bagging and Ensemble Voting methods on the performance of the macro average f1 score, and compare it with non-ensemble models. The results of this study show that the sensitivity of Naïve Bayes to imbalance data is helped by the Ensemble Bagging and Ensemble Voting method with ΔP (delta performance) of range 0.0001 – 0.0018. Logistic Regression has performance with Ensemble Bagging and Ensemble Voting by ΔP of range 0.0001-0.00015. Decision Tree has lowest performance compared to others with ΔP of -0.01.

Keywords: *Email Classification, Ensemble Bagging, Ensemble Voting.*

**Corresponding Author: Danang Triantoro Murdiansyah*

1. INTRODUCTION

The rapidly evolving communication technology has transformed the way people exchange information in today's world, where we can now exchange information quickly, easily, and affordably. One of the most commonly used communication technologies is email. In this modern life, nearly everyone who uses the internet has an email account. Email is extremely important for communication and sharing information for the majority of the population [1]. It is commonly used to share information and often employed for product promotions. Email is also used for formal communication, such as sending assignments to teachers, reporting work results to supervisor, and can even serve as a person's online identity.

The use of email continues to increase each year. The total number of personal and business emails sent daily reached 281 billion in 2018, and by the end of 2022, it is estimated to reach 333 billion [2]. As previously mentioned, email is often used as an online identity. Every time someone registers to use an internet service or application, such as social media or entertainment apps, they are usually required to provide an email address. Service providers often offer people the option to subscribe to the latest information sent via email, and people often unknowingly agree to it. The more someone uses email for personal communication or as a requirement for using an application, the more emails they receive every day.

Some email providers offer features that allow users to manually move emails to specific folders. However, as the number of received emails increases, it becomes more challenging for users to differentiate between important and less important emails. The time required to sort important emails from the pile of unorganized emails becomes longer. Around 46% of employees who receive over a hundred emails per day spend an hour or even more sorting through important emails within the cluttered collection [3]. Manually sorting emails is not a practical solution to this problem. Therefore, a machine is needed to automatically categorize emails into various folders. Email service providers separate emails into specific categories. Some providers categorize emails based on senders, such as friends, family, colleagues, and so on.

Additionally, some email providers separate emails based on context, as done by Google Mail, for instance, which categorizes emails into six categories: personal, social, updates, promotions, forums, and spam. By dividing emails into different categories, users find it easier to locate the emails they desire. This capability can be developed byusing machine learning techniques. Machine learning is a mathematical method that allows machines to exhibit intelligence without being explicitly programmed by a programmer. Generally, machine learning can be divided into three categories: supervised learning, unsupervised learning, and reinforcement learning.

Research on email classification by using machine learning has been conducted by other researchers in the past, including the work by Miodrag Zivkovic et al., with paper titled "Training Logistic Regression Model by Hybridized Multi-verse Optimizer for Spam Email Classification" [4], published in 2023. They detected spam by building a model that combined Logistic Regression and Swarm Intelligence. Another study was conducted by Doaa Mohammed Ablel-Rheem et al., with the paper titled "Hybrid Feature Selection and Ensemble Learning Method for Spam Email Classification" [5], published in 2020. They classified emails by using Naïve Bayes, Decision Tree, and Ensemble Boosting. Pradeep Kumar conducted research with the paper titled "Predictive Snalytics for Spam Email Classification using Machine Learning Techniques" [6], published in 2020. Pradeep Kumar classified emails by using Logistic Regression, k-Nearest Neighbors (k-NN), Naive Bayes, Decision Trees, AdaBoost, ANNs, and SVM. Aakanksha Sharaff et al. conducted research with the paper titled "Towards Classification of Email Through Selection of Informative Features" [7], published in 2020. They classified emails by using Decision Tree, Multinomial Naive Bayes, Random Forest classifiers, Linear Support Vector Machine, and N-Gram feature extraction. Lastly, Ahmed Alghoul et al. conducted the research, with the paper titled "Email Classification Using Artificial Neural Network" [8], published in 2018. Ahmed Alghoul et al. performed spam filtering using ANN algorithms.

The supervised learning method generally produces good intelligence if the training data distribution for each label has an equal amount. However, in the real world, this is rarely the case, as with the number of different types of emails we receive each day. The number of promotional and social emails we receive may be greater than the number of personal emails we receive. Some methods to address this problem are oversampling and undersampling. Each sampling method has its advantages and disadvantages, as do machine learning methods. Each method performs better than others in specific cases. We can combine multiple different machine learning models into a unified model using ensemble techniques. Each model undergoes separate learning processes, and after the separate learning processes are completed, a merging process takes place. This method is known as the bagging ensemble technique (bootstrap aggregating). Generally, ensemble methods improve performance. Therefore, we attempted to create several models by combining various sampling techniques and estimators, and then observed the performance differences between ensemble and non-ensemble approaches using the macro-average F1 score performance parameter.

2. RESEARCH METHOD

This study was conducted with the architecture shown in Figure 1. The training data was transformed into three sets, that are non-sampling data, undersampling data, and oversampling data. Each training data set was used with three estimators, that are Logistic Regression, Naïve Bayes, and Decision Tree, resulting in a total of nine models. Each model was evaluated for performance, including non-ensemble performance, Ensemble Voting performance, and Ensemble Bagging performance. The final step involved calculating the delta performance using the formula shown in equation (1).

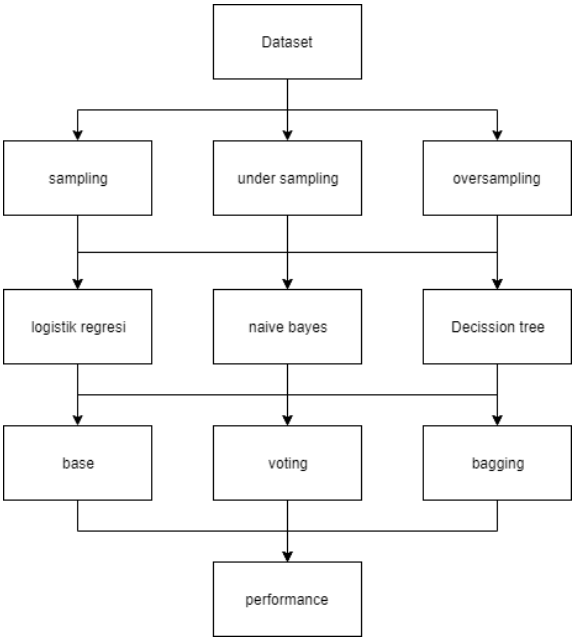

Figure 1. General Research Model

The dataset for this study was obtained from PT Proofn Indonesia, an email company in Bandung, Indonesia, where the data was collected from the employees' Gmail accounts. The dataset consists of five classes: personal, update, promotion, social, and forum. The personal class represents emails exchanged between individuals. The update class includes personal auto-generated update emails such as confirmations, bills, recipes, and statements. The social class comprises emails from social media platforms. The forum class consists of emails from

online groups. Finally, the promotion class includes promotional emails, discounts, and other marketing emails. The email data contains various information obtained from the email header, such as class, sending time, email subject, snippet of email content, and email content type. In this study, the dataset's classes were divided into four test cases. The first case used the original classes (labeled as "_Label"), which include personal, forum, social, update, and promotion. The second case used two classes (labeled as "_Label_a"), that are personal and the others. The third case also used two classes (labeled as "_Label_b"), with personal and update as the first class and the remaining classes as the second class. The last case used three classes (labeled as "_Label_c"), with personal and update as the first class, social and forum as the second class, and promotion as the third class. The dataset consists of several attributes, which are label (data class), date (email sending time), sender (sender's email), subject (email subject), snippet (snippet of email content), unsubscribe (unsubscribe status), mime (email mime type (text, pdf, rar, etc.)).

The data are transformed into a matrix form (bag of words) and have a very large dimension. To reduce this high dimensionality, the next step is feature selection using Chi Square. First, feature selection is performed on the original class (_Label). The results of the Chi Square calculation can be seen in Figure 2. From Figure 2, it is concluded that the use of 12,000 features is too many because there are many features with Chi Square values near to 0. To reduce the features, the top 1000 features based on the Chi Square calculation will be selected. The Chi Square calculation for the 1000 features can be seen in Figure 3. The Chi Square calculation with 1000 features for the _Label_a, _Label_b, and _Label_c classes can be seen in Figures 4, Figures 5, and Figures 6, respectively. From the Chi Square calculation with 1000 features from each class, it can be concluded that many features are suitable for use in this email classification case, for example, by using 250 features.

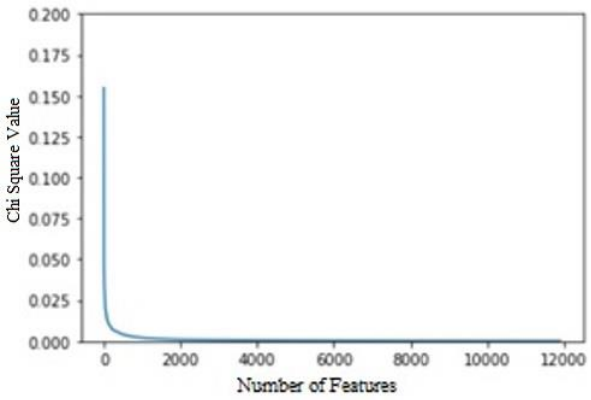

Figure 2. Chi Square Calculation Results for _Label Class Data (12000 Features)

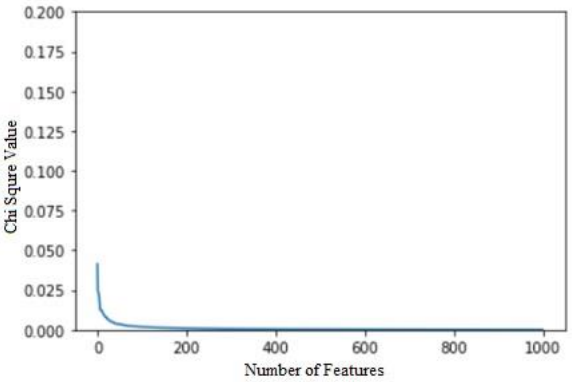

Figure 3. Chi Square Calculation Results for _Label Class Data (1000 Features)

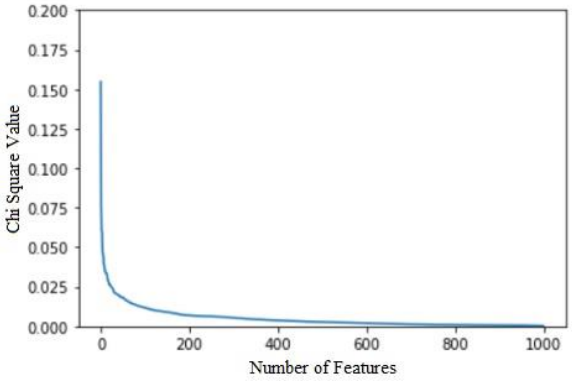

Figure 4. Chi Square Calculation Results for _Label_a Class Data (1000 Features)

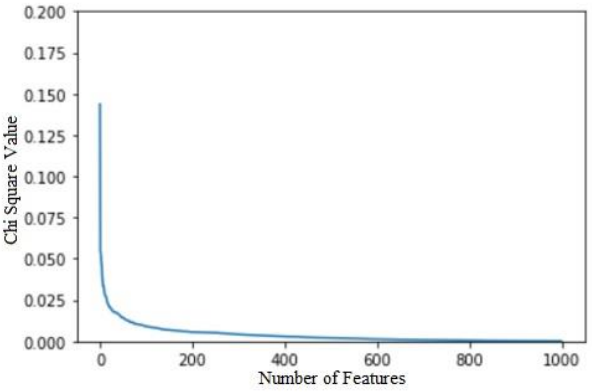

Figure 5. Chi Square Calculation Results for _Label_b Class Data (1000 Features)

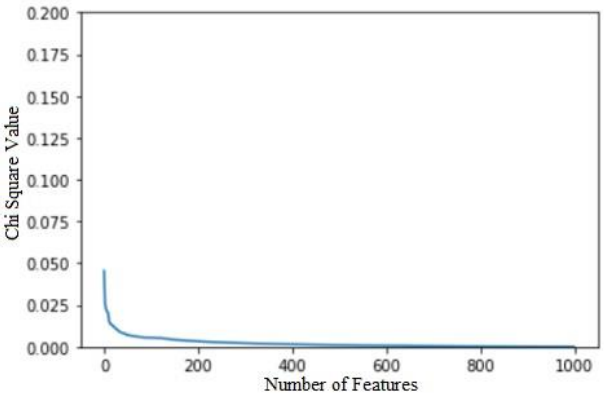

Figure 6. Chi Square Calculation Results for _Label_c Class Data (1000 Features)

Undersampling is one technique that works by removing some majority class data, balancing the number of majority class data with the minority class data [9][10]. In this study, we used random undersampling. Reducing the majority class data to balance the number of data between classes can eliminate a significant amount of information and cause the model to experience underfitting. One solution offered to address the shortcomings of undersampling techniques is by applying the opposite approach, known as oversampling. Oversampling is a technique to balance the number of data between classes by duplicating some minority class data [9]. In this study, we also employed the k-means-smote oversampling technique [10][11]. In the training data, oversampling was performed using the k-means-smote method, while undersampling was done randomly. Then, each data was divided into two parts: the training data and the test data with an 80:20 ratio. Subsequently, the training data was further divided into training and validation data by using k-fold cross-validation with a value of k = 12.

Ensemble Learning is a technique used to combine multiple trained models to solve a problem. One of the objectives of Ensemble Learning is to improve performance and avoid overfitting [12]. One commonly used ensemble method is the Ensemble Bagging method. The Ensemble Bagging method consists of two main processes: bootstrap and aggregating [13]. The first process is bootstrap, where multiple base models are trained separately with different data for each model, resulting in each model having different intelligences. The next process is aggregating or combining. In this stage, a finisher model learns the outputs from each base model on the same training data. The general scheme of Ensemble Bagging can be seen in Figure 7.

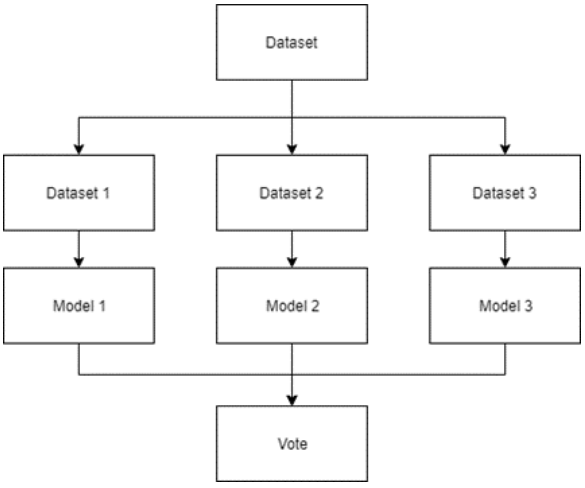

Figure 7. General Scheme of Ensemble Bagging

In general, machine learning can also be classified into three types based on the form of the formula, which are statistical, geometric, and discrete models. In this study, we selected three types of estimators, each representing these three forms of machine learning. Naïve Bayes [14] represents the statistical model as it utilizes probability calculations. Random Forest [15] represents the discrete model as its final form is a logic branching function. Logistic Regression [16] represents the geometric model as it uses a hyperplane as the boundary between classes. All estimators: Logistic Regression, Decision Tree, and Naïve Bayes, used the default hyperparameters of Sklearn 0.21.3. The implemented scheme involved a combination of data sampling, non-ensemble, Ensemble Bagging (with the same estimator), and Ensemble Voting (using three different estimators).

In this study, the performance metric of the models was evaluated based on the macro-average F1 score. To observe the impact of ensemble performance, the delta performance was calculated according to equation (1). The performance of the training data was computed using cross-validation, which provided the average macro-average F1 score across segments.

$$\Delta P = P_e - P_i \quad (1)$$

ΔP represents delta performance, P_e represents the performance obtained from the ensemble scheme, and P_i represents the performance obtained from the non-ensemble scheme.

3. RESULT AND DISCUSSION

The results of this study are presented in the form of boxplots, which can be seen in Figure 8 to Figure 13. From Figure 8 and Figure 11, the average performance of the validation and testing stages for the three estimators can be observed. Decision Tree has the best average performance, followed by Logistic Regression and then Naïve Bayes. Naïve Bayes exhibits the performance shown in Figure 8 and Figure 11 because Naïve Bayes is sensitive to imbalanced data. In Naïve Bayes, the use of non-sampling and sampling data yields significantly different performances. The reason why Naïve Bayes is sensitive to imbalanced data is due to its characteristic that is assumption that data between classes are independent.

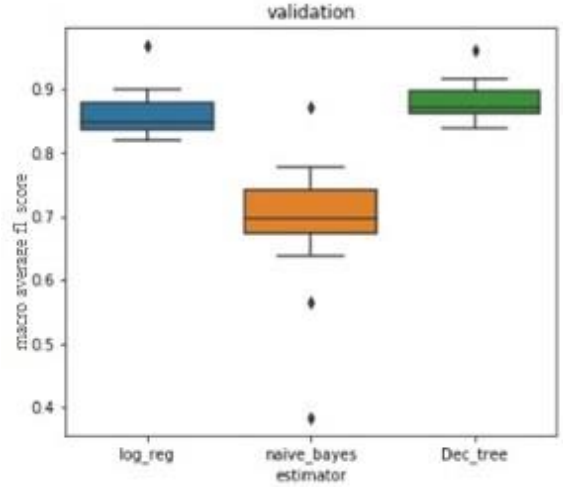

Figure 8. Average Performance of the Validation Stage in Boxplot

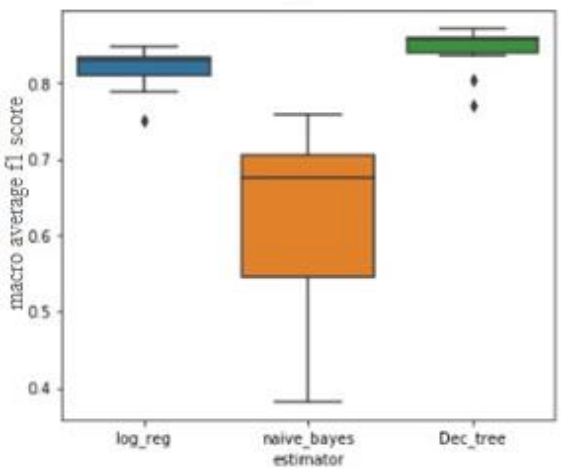

Figure 11. Average Performance of the Testing Stage in Boxplot

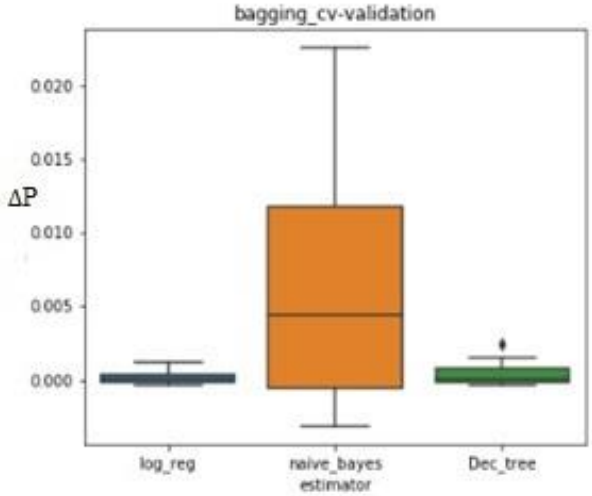

Figure 9. ΔP (Delta Performance) of Validation Stage of Ensemble Bagging Method in Boxplot

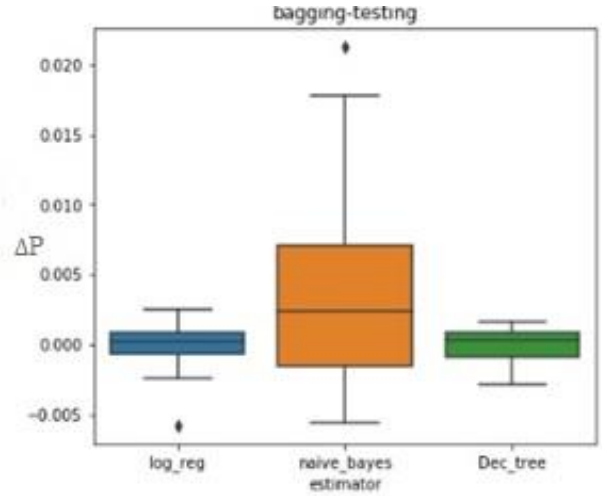

Figure 12. ΔP (Delta Performance) of Testing Stage of Ensemble Bagging Method in Boxplot

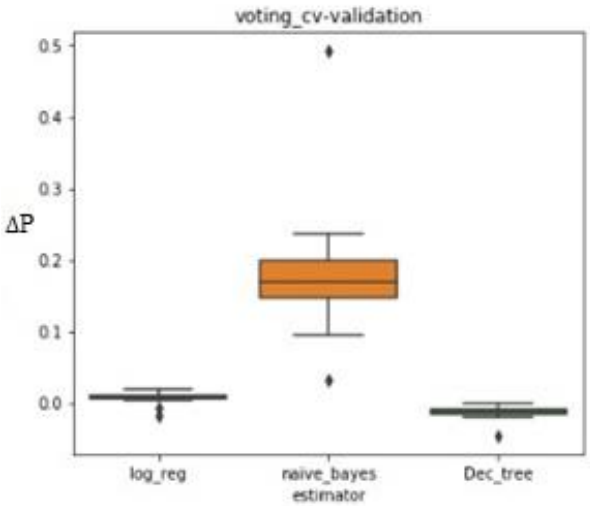

Figure 10. ΔP (Delta Performance) of Validation Stage of Ensemble Voting Method in Boxplot

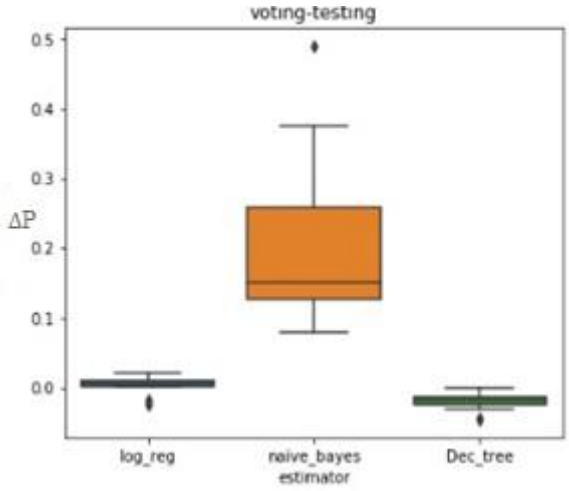

Figure 13. ΔP (Delta Performance) of Testing Stage of Ensemble Voting Method in Boxplot

Figure 9 and Figure 12 provide conclusions regarding the performance (in terms of ΔP (delta performance)) generated by the Ensemble Bagging method. By using Ensemble Bagging, Naïve Bayes benefits the most as it has the largest ΔP value. The performance of the system using the Ensemble Voting method can be seen in Figure 10 and Figure 13. In Ensemble Voting, the algorithm that clearly shows a decrease in performance is Decision Tree. This is likely due to the influence of Naïve Bayes, which is sensitive to imbalanced data.

4. CONCLUSION

In this study, email classification was successfully performed by using Ensemble Learning methods, namely Ensemble Bagging and Ensemble Voting. Three estimators were utilized, which are Logistic Regression, Naïve Bayes, and Decision Tree. Naïve Bayes, which is sensitive to imbalanced data, exhibited better performance when using Ensemble Learning compared to when not using Ensemble Learning, with ΔP (delta performance) range of 0.0001 to 0.0018. Logistic Regression has ΔP range of 0.0001 to 0.00015. Decision Tree has the lowest ΔP performance when using Ensemble Learning, with ΔP value of -0.01. From the observations of email classification results using Ensemble Bagging and Ensemble Voting, with three sampling methods, which are non-sampling, random oversampling, and k-means-smote oversampling, it can be concluded that Ensemble Bagging and Ensemble Voting do not always yield better performance compared to not using these methods when considering the macro-average F1 score.

5. REFERENCE

[1] X. L. Wang and I. Cloete, "Learning to classify email: A survey," *2005 Int. Conf. Mach. Learn. Cybern. ICMLC 2005*, pp. 5716–5719, 2005.

[2] The Radicati Group.inc, "Email Statistics Report, 2017-2021", 2018

[3] S. Tsugawa, K. Takahashi, H. Ohsaki, and M. Imase, "Robust estimation of message importance using inferred inter-recipient trust for supporting email triage," *Proc. - 2010 10th Annu. Int. Symp. Appl. Internet, SAINT 2010*, pp. 177–180, 2010.

[4] M. Zivkovic *et al.*, "Training Logistic Regression Model by Hybridized Multi-verse Optimizer for Spam Email Classification," in *Proceedings of International Conference on Data Science and Applications: ICDSA 2022, Volume 2*, 2023, pp. 507–520.

[5] D. M. Ablel-Rheem, A. O. Ibrahim, S. Kasim, A. A. Almazroi, M. A. Ismail, and others, "Hybrid feature selection and ensemble learning method for spam email classification," *Int. J.*, vol. 9, no. 1.4, pp. 217–223, 2020.

[6] P. Kumar, "Predictive analytics for spam email classification using machine learning techniques," *Int. J. Comput. Appl. Technol.*, vol. 64, no. 3, pp. 282–296, 2020.

[7] A. Sharaff and U. Srinivasarao, "Towards classification of email through selection of informative features," in *2020 First International Conference on Power, Control and Computing Technologies (ICPC2T)*, 2020, pp. 316–320.

[8] A. Alghoul, S. Al Ajrami, G. Al Jarousha, G. Harb, and S. S. Abu-Naser, "Email Classification Using Artificial Neural Network," 2018.

[9] V. Babar and R. Ade, "MLP-based undersampling technique for imbalanced learning," in *2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT)*, 2016, pp. 142–147.

[10] A. Indrawati, "Penerapan Teknik Kombinasi Oversampling dan Undersampling untuk Mengatasi Permasalahan Imbalanced Dataset," *JIKO (Jurnal Inform. dan Komputer)*, vol. 4, no. 1, pp. 38–43, Apr. 2021.

[11] 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, pp. 321–357, 2002.

[12] B. Singh, N. Kushwaha, and O. P. Vyas, "A Scalable Hybrid Ensemble model for text classification," *IEEE Reg. 10 Annu. Int. Conf. Proceedings/TENCON*, pp. 3148–3152, Feb. 2017.

[13] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, Aug. 1996.

[14] V. Metsis, I. Androutsopoulos, and G. Paliouras, "Spam filtering with naive bayes-which naive bayes?," in *CEAS*, 2006, vol. 17, pp. 28–69.

[15] M. Dumont, R. Marée, L. Wehenkel, and P. Geurts, "Fast multi-class image annotation with random subwindows and multiple output randomized trees," in *International Conference on Computer Vision Theory and Applications (VISAPP)*, 2009.

[16] H.-F. Yu, F.-L. Huang, and C.-J. Lin, "Dual coordinate descent methods for logistic regression and maximum entropy models," *Mach. Learn.*, vol. 85, pp. 41–75, 2011.