JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 10, No. 4, December 2025, Pp. 3772-3786

CLASSIFICATION OF CARDIOVASCULAR AND CHRONIC RESPIRATORY DISEASES UTILIZING ENSEMBLE MODELS WITH DATA EXPLORATION TECHNIQUES

I GNS Putra*[1)], Amri L Najih[2)], Unique DA Resiloy[3)], Rachmat B Yudhianto[4)], Erfiani[5)], Anwar Fitrianto[6)]

1. Department Statistics and Data Science, SSMI, IPB University, Indonesia
2. Department Statistics and Data Science, SSMI, IPB University, Indonesia
3. Department Statistics and Data Science, SSMI, IPB University, Indonesia
4. Department Statistics and Data Science, SSMI, IPB University, Indonesia
5. Department Statistics and Data Science, SSMI, IPB University, Indonesia
6. Department Statistics and Data Science, SSMI, IPB University, Indonesia

Article Info
Keywords: BPJS health insurance; data exploration; ensemble learning; preprocessing; Random Forest.

Article history:
Received 14 October 2025
Revised 30 October 2025
Accepted 3 November 2025
Available online 1 December 2025

DOI :
https://doi.org/10.29100/jipi.v10i4.9311

* Corresponding author.
Corresponding Author
E-mail address:
ngurahsentana@apps.ipb.ac.id

ABSTRACT
Non-communicable diseases, especially cardiovascular and chronic respiratory conditions, contribute significantly to Indonesia's healthcare burden and BPJS expenditure. Health claim data often suffer from class imbalance, multicollinearity, and outliers that impair model accuracy. This study evaluates the impact of essential data exploration techniques such as winsorizing, correlation and VIF analysis, variable selection, and SMOTE on the performance of ensemble classifiers. The dataset comprises 497,439 BPJS health insurance claims from 2022, including 27 predictors (14 numerical and 13 categorical). Two data pipelines were compared: one without preprocessing and another incorporating systematic data exploration. Five ensemble models were tested, namely Decision Tree, Extra Trees, Random Forest, XGBoost, and LightGBM. Model performance was assessed using F1-score, balanced accuracy, and G-mean across 20 stratified cross-validations. The results show that preprocessing substantially improves classification fairness and accuracy. Bagging models, particularly Random Forest, achieved the highest improvement, with balanced accuracy and G-mean increasing from around 0.93 to 0.99. Boosting models showed modest gains. These findings highlight that rigorous data exploration enhances ensemble classifier performance, enabling more reliable disease classification and supporting fairer, data-driven decision-making in BPJS health management.

I. INTRODUCTION

Non-communicable illnesses have become a predominant global health issue, with cardiovascular and chronic respiratory diseases identified as major contributors to disability and mortality. The World Health Organization (WHO) reports that cardiovascular diseases cause over 18 million deaths annually, while chronic respiratory disorders, including asthma, affect hundreds of millions, imposing a significant economic burden. In Indonesia, both conditions directly influence the national healthcare financing system, particularly the Social Security Agency for Health (BPJS Kesehatan), which faces an increasing volume of claims from these disease categories. The growing prevalence and overlapping risk factors, such as hypertension, diabetes, and obesity, demand innovative strategies that enable early detection, improve risk prediction accuracy, and effectively distinguish between conditions to support targeted interventions and ensure the sustainability of the health insurance system. Naser et al. (2024) highlight machine learning offers a promising solution to overcome the limitations of con-ventional cardiovascular diagnostics by improving early detection and prediction accuracy [1].

Nonetheless, predictive endeavours are impeded by the quality of accessible healthcare data. Real-world datasets, such as those administered by BPJS, frequently exhibit unstructured characteristics, suffer from missing values, contain redundant variables, and include outliers. A significant concern is class imbalance, characterized by a predominance of healthy patients in the records, while individuals with serious diseases constitute minority classes of paramount clinical significance. Inadequate management of machine learning (ML) models sometimes results in a bias towards the majority class, yielding statistically elevated accuracy while neglecting the

identification of high-risk patients. Prior research indicates that methodologies such as the Synthetic Minority Oversampling Technique (SMOTE) are crucial for rectifying the imbalance [2], whereas outlier detection approaches like Isolation Forest have demonstrated the capacity to enhance model stability without eliminating clinically significant information [3].

Moreover, health-related factors such as age, blood pressure, cholesterol, and body weight often exhibit strong interrelationships that cause multicollinearity. Yoo et al. (2014) demonstrated that multicollinearity can substantially distort coefficient estimates [4], and Yıldırım (2024) further confirmed that even advanced algorithms, such as XGBoost and SVM, experience performance degradation under these conditions [5]. Outliers also present a serious challenge, whether caused by recording errors or atypical clinical cases, as they may skew decision boundaries and reduce model generalizability. Consequently, meticulous data exploration encompassing class balance, feature selection, and outlier management constitutes an essential foundation before the commencement of modeling.

After establishing a robust data foundation, the selection of the algorithm becomes the subsequent factor influencing success. In this situation, ensemble learning has proven to be superior for addressing intricate health classification challenges. Bagging approaches, exemplified by Random Forest, construct several decision trees from varied sub-samples and consolidate their outputs, therefore enhancing resilience against overfitting and excessive variance. Conversely, boosting methodologies like AdaBoost, Gradient Boosting, and XGBoost function iteratively by allocating increased weight to misclassified examples from prior iterations, therefore progressively refining the model's emphasis on challenging cases, especially minority classes. A multitude of research corroborates the efficacy of these methodologies.

Previous studies emphasize the essential importance of data exploration. Li et al. (2020) demonstrated that the integration of Mutual Information-based feature selection with SVM enhanced both accuracy and efficiency in identifying heart disease utilizing the Cleveland dataset [6]. Tiwari et al. (2022) established an ensemble framework integrating Extra Trees, Random Forest, and XGBoost, attaining an accuracy of 92.34% and surpassing the performance of individual models [7]. These findings confirm that the integration of extensive data exploration with sophisticated ensemble modeling can produce more precise and dependable predictions.

The classification of cardiovascular and respiratory diseases is particularly relevant because these conditions share risk factors but differ in pathophysiological mechanisms and treatment strategies. A model’s ability to predict risk and effectively differentiate between the two is essential for clinicians and healthcare policymakers. In the BPJS context, enhanced data exploration enables the development of fairer and more resilient predictive models, improving claim mapping accuracy and supporting data-driven preventive actions.

This study aims to evaluate how integrating extensive data exploration with ensemble learning can produce robust, equitable, and generalizable classification models. The research contributes to the growing literature on machine learning applications in healthcare by emphasizing the pivotal role of data exploration. It supports BPJS Kesehatan in managing the increasing burden of non-communicable disease claims through improved predictive accuracy, optimized healthcare planning, and strengthened precision medicine initiatives to enhance population health outcomes.

II. METHOD

This research employs healthcare data from BPJS Indonesia, concentrating on cardiovascular and respiratory ailments. The methodological framework aims to evaluate the influence of data exploration on the classification accuracy of several machine learning models.

A. Data

This study utilized data derived from the 2022 sample of BPJS Kesehatan claims and medical records. The dataset includes various medical examination variables, such as the frequency of visits to cardiology or pulmonology clinics, the number of outpatient and inpatient visits, and administrative data encompassing membership status, class of care, healthcare facility location, INA-CBGs tariffs, and total charges. The dataset principally concentrates on three principal categories: cardiovascular diseases (I. Cardiovascular System Groups), respiratory diseases (J. Respiratory System Groups), and general examinations. The dataset comprises 497,439 observations (patients or visits) and includes 27 variables, which contain numerical, categorical, and administrative data. The variables included in this dataset are jenis_kelamin (Participant’s gender), status_kepesertaan (BPJS membership status), hub_kel (Relationship with the head of household), stat_kawin (Marital status), kelas_rawat (Participant’s entitlement class of care), seg_pserte (Membership segmentation), bobot_peserta (Risk weight or participant weighting factor), kepemilikan_fas (Type of healthcare facility ownership), seg_tl (Segmentation of service destination facility), kelas_tl (Class of hospital service destination), tarif_INACBGs (Claim tariff according to INA-

CBGs), tagihan_verif (Verified claim amount approved by BPJS), layanan (Type of service), jen_poli (Type of target polyclinic), tagihan_tl (Total claim amount submitted to BPJS), umur (Participant's age), jumlah_anggota (Number of family members registered under the BPJS card), lama_kunjungan (Length of stay [days]), lokasi_faskes (Location of healthcare facility), freq_IGD (Frequency of visits to the emergency department), freq_PAR (Frequency of visits for pulmonary/respiratory examinations), freq_ANA (Frequency of visits for laboratory/analysis examinations), freq_JAN (Frequency of visits to cardiology clinics), freq_INT (Frequency of visits to internal medicine services), freq_Rinap (Number of inpatient visits), freq_Rjalan (Number of outpatient visits), casemix_INACBGs (INA-CBGs category)

B. Phases of Research

This study utilizes two categories of datasets for comparison: untreated data and data that has been adjusted for outliers, multicollinearity, inter-variable correlations, and class imbalance.

1. Data Preparation: Following the elimination of missing values, the dataset is replicated into two subsets. One group undergoes particular therapies, whereas the other remains untreated.
2. Data splitting: The dataset is partitioned into 70% for training and 30% for testing. The partitioning is executed utilizing Stratified K-Fold to preserve class proportions.
3. Modeling: The modeling technique employs stratified cross-validation to maintain constant class distribution across folds. Multiple ensemble learning algorithms are utilized, including Decision Tree, Extra Trees, Random Forest, XGBoost, and LightGBM. These algorithms were chosen for their capacity to manage intricate data, identify non-linear patterns, and mitigate bias and variance concerns. Bagging approaches, such as Random Forest, Decision Tree, and Extra Trees, provide stability by aggregating many decision trees, whereas boosting techniques, such as XGBoost and LightGBM, increasingly focus on challenging instances through iterative weighting.
4. Assessment: Models are assessed with the F1-score and Balanced Accuracy measures. The F1-score is selected to equilibrate the trade-off between precision and recall, offering a more accurate assessment of the model's capacity to detect patients with critical illnesses. Balanced accuracy evaluates imbalanced datasets by considering sensitivity across all classes.
5. Repetition: The stability of the model is assessed by conducting the process 20 times for each treatment, followed by the calculation of the mean and standard deviation of the assessment metrics.
6. Results: The concluding phase evaluates model efficacy across datasets with and without specialized interventions, determining the most effective and stable model.

C. Winzorizing

Winsorizing was initially proposed by Charles P. Winsor (1895–1951) and subsequently popularized by Tukey (1962), who characterized it as a method for substituting extreme values with designated values from a range. Tukey designated the method in tribute to Charles P. Winsor, who significantly advocated for its use in statistical data analysis [8].Winsorizing is a data preprocessing method employed to mitigate the impact of outliers in statistical research. The process is substituting extreme values in a dataset, including both the minimum and maximum, with values in lower or upper percentiles. In 5% winsorizing, values below the 5th percentile are substituted with the 5th percentile value, while those above the 95th percentile are replaced with the 95th percentile value. This method guarantees that the data stays within a plausible range, while reducing the influence of outliers that could otherwise skew the analytical results [9].

Recent studies highlight that Winsorizing continues to be relevant across modern statistical and computational contexts. Han, Kim, and Jung (2025) found that it enhances robustness in *Principal Component Analysis* (PCA) [10], while Lafit et al. (2022) applied *multivariate Winsorization* within the *graphical lasso* framework to stabilize covariance estimation [11]. Adaptive variants have also been proposed, such as Orenstein's (2018) *adaptive Winsorization* for balancing bias and variance. Beyond traditional statistics, Winsorizing has been utilized to correct data non-normality [12] and to reduce false positives in RNA-seq differential expression analyses [13]. Overall, it remains an effective and practical approach for preserving data integrity while minimizing the influence of extreme values.

D. SMOTE

SMOTE (Synthetic Minority Over-sampling Technique) is an oversampling approach employed to rectify class imbalance in datasets. Class imbalance can lead to bias in machine learning models favoring the majority class.

SMOTE produces new synthetic samples by interpolating between existing minority data points within a specified feature space [14]. SMOTE is well acknowledged for its capacity to improve model performance, especially in addressing underrepresented classes. Recent studies have continued to refine SMOTE to address its limitations, such as the potential generation of noisy or overlapping samples. Li et al. (2025) introduced an improved variant to mitigate overfitting and improve generalization [15], while Kosolwattana et al. (2023) developed SASMOTE, which adaptively filters ambiguous synthetic data for higher quality in medical datasets [16]. Another enhancement, *MeanRadius-SMOTE* [17], focuses on creating boundary samples near decision regions to better cacaptuture minority characteristic. Empirical evidence further shows shows that integrating SMOTE with modern algorithms, such as XGBoost, improves sensitivity and F1-score despite minor accuracy reduction, ensuring more balanced and clinicaly meaningful predictions [18].

E. Spearman Rank Correlation

Spearman's rank correlation coefficient, also known as Spearman's rho (ρ), is a non-parametric measure of association established by Spearman in 1904. This correlation evaluates the strength of a monotonic relationship, whether linear or non-linear, between two variables assessed on an ordinal or interval/ratio scale. The method exhibits greater resilience to outliers and does not need a normal distribution. Spearman correlation is defined as:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{1}$$

where d_i is the difference in ranks between paired observations X_i and Y_i, and n indicates the total number of paired observations. In instances of tied ranks, the tied values receive the mean of their ranks. Alternatively, the calculation can be articulated by covariance:

$$\rho = \frac{Cov(R_X, R_Y)}{\sigma_{R_X} . \sigma_{Y_X}} \tag{2}$$

where R_X and R_Y denote the ranks of the corresponding variables. The value of ρ varies from -1 and +1 [15].

F. Cramer's V

Cramer's V quantifies the relationship between two categorical variables, originating from the chi-square (χ^2) test. Cramer's V values vary from 0 to 1. Cramer's V is mathematically defined as follows [16]:

$$V = \sqrt{\frac{\chi^2}{n \times (k-1)}} \tag{3}$$

where $V = \sqrt{\frac{\chi^2}{n \times (k-1)}}$ is the chi-square statistic derived from the test of independence, $\chi^2$is the total sample size, and n is the lesser value between the number of rows and columns.

G. Decision Tree

The Decision Tree algorithm adheres to the Classification and Regression Trees (CART) approach, functioning recursively by dividing data into progressively homogeneous subsets about the target variable through the identification of appropriate features and split points [19]. This study utilizes the *DecisionTreeClassifier* from the Python package scikit-learn, which is fundamentally derived from the CART algorithm. This software offers multiple parameters to regulate tree complexity, including criterion (Gini or Entropy), max_depth, min_samples_split, and ccp_alpha for cost-complexity pruning. Moreover, scikit-learn enables the incorporation of decision trees into workflows that amalgamate preprocessing, cross-validation, and grid search for parameter optimization [20].

H. Random Forest

Random Forest is an ensemble learning technique based on the notion of bagging (bootstrap aggregating). It generates several decision trees (CART) from various subsets of the training data selected randomly with replacement (bootstrap sampling) [21]. The ultimate forecast in Random Forest is established through majority vote (for classification) or averaging (for regression). This study uses the *RandomForestClassifier* from the Python package scikit-learn as the foundational learner, incorporating both bagging and random feature selection [20]. Key parameters in *RandomForestClassifier* comprise: n_estimators, which denotes the quantity of trees in the forest (default = 100); criterion, the function employed to assess the quality of a split (default = "gini"; alternatives include "entropy" or "log_loss"); max_features, the number of features evaluated at each split (default = "sqrt" for classification); and max_depth, min_samples_split, min_samples_leaf, which regulate tree complexity, alongside bootstrap and class_weight.

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 10, No. 4, December 2025, Pp. 3772-3786

I. Extra Trees

Extremely Randomized Trees (Extra Trees) is an enhancement of Random Forest, principally designed to diminish variance and enhance computing efficiency. The primary distinction between Extra Trees and Random Forest pertains to the method of determining node splits. In Random Forest, a random subset of features is picked, and the optimal split point is determined based on impurity metrics (such as Gini or Entropy). Conversely, Extra Trees not only randomly select features but also randomly determine split thresholds. The degree of randomization exceeds that of Random Forest, resulting in less variance among trees [22]. This study utilizes the ExtraTreesClassifier from the Python package scikit-learn, which is founded on CART as the base learner and incorporates random threshold selection at each split node [20].

J. Xgboost

Extreme Gradient Boosting (XGBoost), created by Chen and Guestrin in 2016, is a tree-based boosting technique that enhances the Gradient Boosting approach proposed by Friedman in 2001. XGBoost generates models additively as follows:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{4}$$

where f_t indicates the new tree added at the t^{th} iteration. Each new tree is constructed to minimize the following objective function:

$$Obj = \sum_i l(y_i, \hat{y}_i^{(t)} + \sum_t \Omega(f_t) \tag{5}$$

where l indicates the loss function and Ω is the regularization term that penalizes tree complexity. Optimization is executed utilizing a second-order Taylor expansion method, facilitating enhanced speed and precision in training. This study utilizes the xgboost package, notably the XGBClassifier, which employs CART as the foundational learner within a boosting architecture [23].

K. LightGBM

Light Gradient Boosting Machine (LightGBM) is a tree-based gradient boosting system created by Microsoft Research as an enhancement of XGBoost. LightGBM was introduced to address the shortcomings of XGBoost regarding computational speed and memory efficiency, especially in the context of large datasets [24]. This study utilizes the LGBMClassifier from the lightgbm package, which employs CART as the foundational learner within a boosting framework. The principal distinctions between LightGBM and XGBoost are as follows: 1. The tree growth strategy of LightGBM involves a leaf-wise approach, picking the node with the largest gain among all leaves, in contrast to the level-wise method employed by XGBoost. Histogram-based binning: LightGBM transforms continuous feature values into discrete intervals utilizing histograms, hence enhancing training efficiency; 3. LightGBM natively supports categorical features, eliminating the need for one-hot encoding with the categorical_feature argument. LightGBM facilitates parallel learning and GPU acceleration, enhancing computational efficiency and scalability.

III. RESULT AND DISCUSSION

A. Data Exploration

An exploratory data analysis was performed prior to modeling to comprehend the essential aspects of the 2022 BPJS claims dataset. The dataset has 497,439 observations and 27 variables, including 13 category and 14 numerical variables. The distribution of variables is depicted in Figures 1 and 2.

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 10, No. 4, December 2025, Pp. 3772-3786

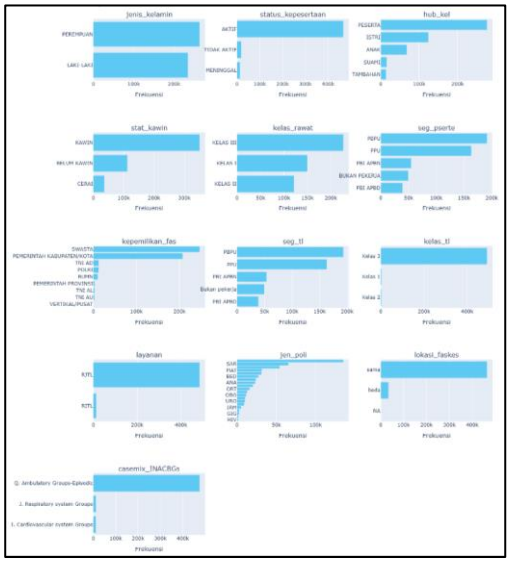


Fig. 1. Distribution of frequencies for categorical variables.

Figure 1 illustrates that the categorical factors are predominantly influenced by active membership status, class III inpatient care, and outpatient service type. The variable with the highest number of categories is Jenis Poli (kind of clinic), of which only the 20 most prevalent categories are presented. Moreover, the predominant number of participants are registered under the PBPU and PPU categories, with the bulk of healthcare facilities being held by private organizations and local government entities. The distribution of the casemix variable reveals a predominance of Ambulatory patients, while this work concentrates on classification modeling for respiratory and cardiovascular illnesses. This disparity will necessitate particular focus during data preparation.

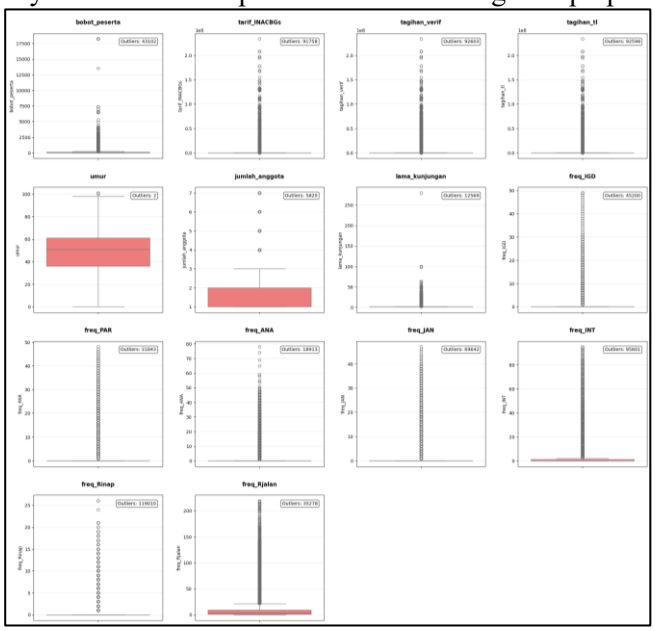


Fig. 2. Boxplot representation of numerical variables' distribution

Figure 2 indicates that almost all numerical variables exhibit a significant presence of outliers. For example, tarif_INACBGs, tagihan_verif, and tagihan_tl display extreme value distributions, containing over 90,000 outlier observations, which signify the existence of significantly high claims relative to the bulk of the data. Likewise, bobot_peserta and freq_IGD exhibit significantly skewed distributions characterized by numerous extreme occurrences, potentially distorting the mean and biasing subsequent analysis.

Conversely, variables such as age exhibit relative stability with minimal outliers, leading to more concentrated distributions. The jumlah_anggota variable exhibits irregularities in homes with abnormally large family sizes, despite a significant concentration of records among lesser numbers. The lama_kunjungan (length of stay) variable reveals that the majority of patients received outpatient care characterized by brief visits, whilst a minority experienced significantly prolonged stays.

Furthermore, the frequency-of-visit variables (freq_PAR, freq_ANA, freq_JAN, freq_INT, freq_Rinap, and freq_Rjalan) display extreme distributions: the majority of individuals had minimal visitations, while a tiny

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 10, No. 4, December 2025, Pp. 3772-3786

minority reported significantly elevated visit frequencies, classified as outliers. The findings indicate that the BPJS dataset exhibits uneven distributions and is significantly affected by extreme values, highlighting the necessity for specialized treatments like winsorizing to prevent subsequent analytical models from being skewed by outliers.

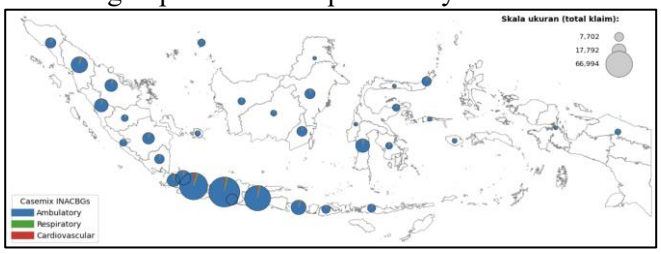


Fig. 3. Distribution of claims among three disease groups (Ambulatory Groups–Episodic, Respiratory System Groups, and Cardiovascular System Groups) over all provinces in Indonesia.

Figure 3 illustrates that the size of the circle denotes the aggregate number of claims, whereas the colors of the pie chart reflect the distribution of each disease group. The Ambulatory Groups–Episodic (blue) predominantly account for claims in all provinces. The proportions of Respiratory System Groups (green) and Cardiovascular System Groups (red) are comparatively minor, constituting only a small fraction of the overall claims. This observation aligns with the initial dataset condition, which indicates a substantial class imbalance, hence warranting the application of data balancing strategies like SMOTE.

Geographically, provinces on Java Island have the highest volume of claims relative to other locations in Indonesia. Central Java is distinguished as the province with the highest number of claims, as evidenced by the greatest circle size. This can be ascribed to its elevated population density, enhanced access to healthcare services, and a greater number of healthcare institutions relative to other provinces.

Conversely, areas beyond Java, especially in Eastern Indonesia, exhibit comparatively smaller circles, signifying fewer claims. Nevertheless, the prevalence of Ambulatory Groups persists uniformly across all regions, underscoring that the disparity in claims is not confined to specific locales but is a national concern.

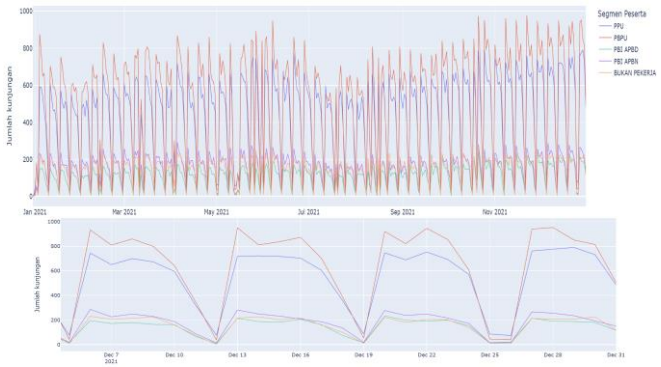
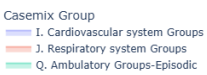


Fig. 4. Daily data of BPJS claims.

Figure 4 illustrates that daily data analysis indicates that BPJS claims are mostly produced by the PBPU segment (non-wage workers), followed by the PPU segment (wage workers). The percentage of claims from PBPU regularly surpasses that of PPU on almost all observation dates, indicating that PBPU represents the largest membership component in BPJS nationally.

An analysis of daily time patterns reveals a continuous trend: a reduction in claims on weekends (Saturday–Sunday) and a significant rise each Monday. This decrease is largely attributable to the restricted availability of healthcare services on weekends, especially outpatient specialty clinics, which typically do not function at full capacity. The increase in claims on Mondays may come from the aggregation of healthcare services rendered over the weekend, thereafter, recorded and filed to the BPJS system on the initial working day.

This daily seasonal trend is significant as it demonstrates that variations in the number of claims are impacted not only by changes in medical need but also by operational and administrative factors inside healthcare institutions and the BPJS claims reporting system.



JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 10, No. 4, December 2025, Pp. 3772-3786

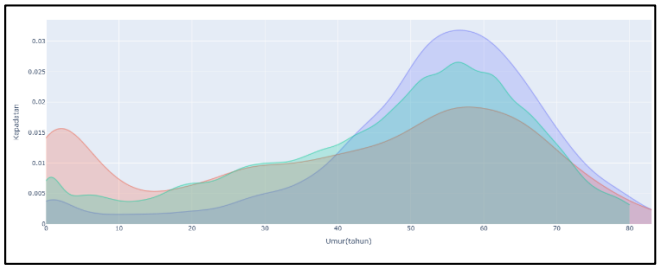


Fig. 5. Density distribution of BPJS claims segmented by age group.

Figure 5 illustrates the density distribution of BPJS claims segmented by age group across three primary disease categories: Cardiovascular System Groups, Ambulatory Groups–Episodic, and Respiratory System Groups.

BPJS claims are primarily concentrated among those aged 40 to 70 years, peaking between the ages of 55 and 60. In this age group, cardiovascular disorders represent the predominant category of claims, followed by ambulatory cases and respiratory diseases. This pattern aligns with medical research indicating that the risk of cardiovascular disease escalates with age, especially in middle-aged and elderly demographics.

Conversely, a distinct trend arises within the 0–10 age demographic, when respiratory illnesses predominate the claims. At age 0 (infants under one year), respiratory claims are the most prevalent, primarily linked to neonatal respiratory diseases. In older children and adolescents, respiratory claims are substantial, albeit less pronounced than in the neonatal period.

The findings indicate that the distribution of disease claims is significantly affected by age: respiratory disorders predominate in children, particularly newborns, whereas cardiovascular diseases prevail in adults and the elderly. Consequently, health intervention methods and BPJS claim management must account for the age-specific distribution of disease risks.

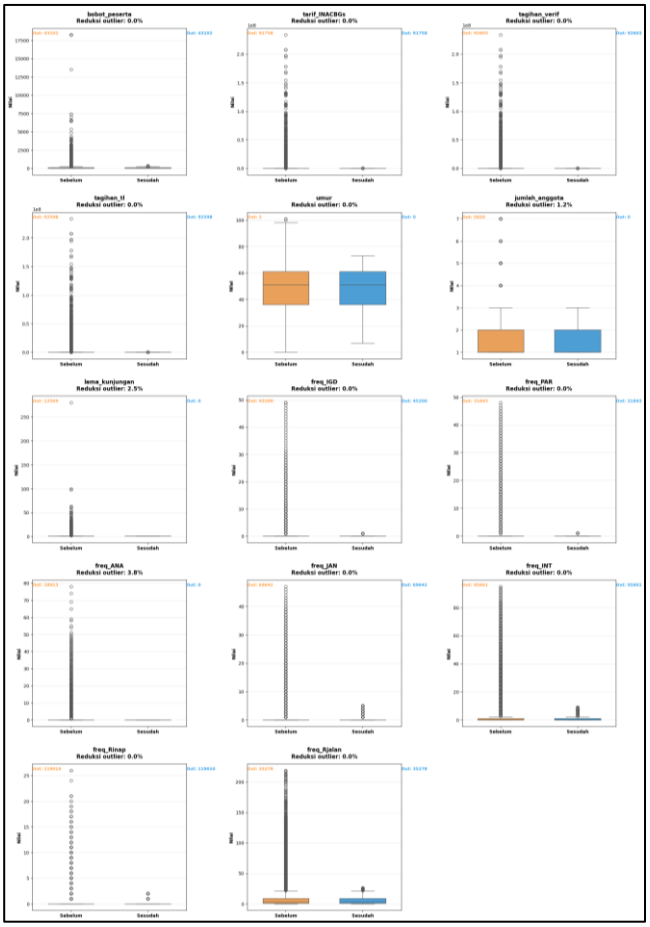


Fig. 6. Boxplot distribution of numerical variables, comparing pre- and post-winsorization.

Figure 6 illustrates a comparative plot pre- and post-winsorization, indicating that the majority of numerical variables in the BPJS dataset continue to display distributions characterized by a significant presence of outliers.

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 10, No. 4, December 2025, Pp. 3772-3786

Nevertheless, numerous variables underwent a decrease in outliers. For example, jumlah_anggota diminished by 1.2%, lama_kunjungan declined by 2.5%, and freq_ANA reduced by 3.8%. This suggests that winsorizing was effective in mitigating extreme values in these variables, leading to more balanced distributions.

Conversely, the primary claim-related variables, including tarif_INACBGs, tagihan_verif, and tagihan_tl, exhibited no significant alterations post-winsorization, with an outlier reduction documented at 0%. This criterion indicates that the dispersion of claim costs is intrinsically broad and cannot be adequately constrained solely by implementing winsorizing at the 5% threshold. Likewise, specific visit frequency variables, including freq_IGD, freq_PAR, freq_JAN, freq_INT, freq_Rinap, and freq_Rjalan, continued to exhibit severe outlier patterns without any discernible decrease.

The findings indicate that winsorizing has a constrained impact when utilized on data characterized by significant outlier ranges, especially for financial and service frequency variables. Nonetheless, for demographic factors and duration of stay, winsorizing effectively mitigated extreme outliers and yielded more representative distributions. This finding underscores that the integration of winsorizing with additional techniques (such as logarithmic transformation or cutting) may be essential for more effectively managing severely skewed data before to advancing to the modeling phase.

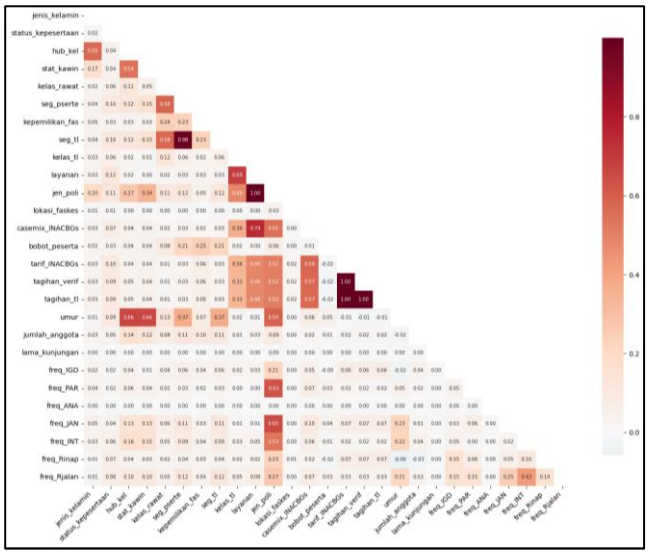


Fig. 7. Correlation heatmap of variables prior to preprocessing.

Figure 7 presents a mixed correlation heatmap before preprocessing, indicating that the majority of variables demonstrate weak to moderate relationships, while many pairs reveal very strong correlations. For instance, seg_pserte and seg_tl exhibit a correlation of 0.977, signifying that participant segmentation and service facility segmentation provide substantially comparable information. Likewise, layanan and jen_poli exhibit a perfect correlation (1.000), indicating that both variables convey identical information. A robust correlation is noted between layanan and casemix_INACBGs (0.743), indicating a significant association between service type and INA-CBGs classification.

The financial variables tarif_INACBGs, tagihan_verif, and tagihan_tl demonstrate exceptionally strong correlations, surpassing 0.99 and, in many instances, achieving perfect correlation (1.000). This is justifiable as these variables all denote cost aspects derived from roughly identical computation frameworks, thus exhibiting considerable redundancy. Such correlations may induce multicollinearity and bias in modeling, especially when utilizing techniques that are sensitive to linearity, such as logistic regression.

To resolve this issue, superfluous variables were eliminated as illustrated in Figure 5. The eliminated variables were layanan, seg_tl, tagihan_tl, and tagihan_verif, as they demonstrated significant associations with other, more representative variables. This reduction is anticipated to assist the model in circumventing multicollinearity issues and yielding more stable and interpretable outcomes.

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 10, No. 4, December 2025, Pp. 3772-3786

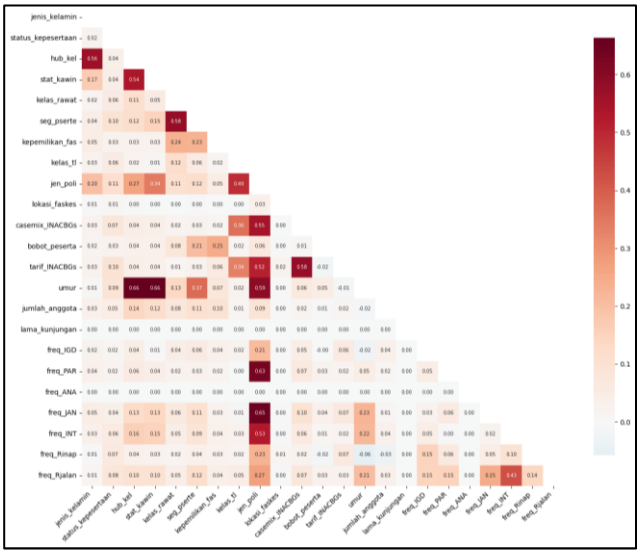


Fig. 8. Heatmap korelasi antara variabel setelah intervensi.

Figure 8 illustrates that a correlation criterion of $|r| \geq 0.7$ is frequently employed as a preliminary signal that collinearity may significantly impact model estimate and forecast accuracy. The 0.7 threshold is a commonly utilized criterion across several fields, serving as the cut-off point beyond which multicollinearity may compromise analytical outcomes. Dormann et al. observed that while 0.7 is the predominant value utilized, alternative thresholds—more stringent (0.4) or more lenient (0.85)—are employed based on the research environment. Consequently, implementing a criterion below 0.7 is deemed prudent when researchers seek to use a more cautious approach to mitigate multicollinearity concerns [25].

TABLE I
VARIANCE INFLATION FACTOR (VIF) VALUES

Variable	VIF
freq_Rjalan	1.408
freq_INT	1.317
umur	1.117
freq_JAN	1.111
freq_Rinap	1.077
freq_IGD	1.045
freq_PAR	1.027
tarif_INACBGs	1.014
jumlah_anggota	1.008
bobot_peserta	1.004

Table I presents the Variance Inflation Factor (VIF) analysis for all numerical variables reveals that no variable exceeds a VIF value of 10. All VIF values are notably low, ranging from 1.004 to 1.408. This indicates the absence of multicollinearity problems among the numerical variables in the BPJS dataset. Consequently, each numerical variable offers reasonably separate and non-redundant information. Contrary to the previous correlation study, multicollinearity issues were identified among category and financial variables, including seg_pserte–seg_tl ($r = 0.977$), layanan–jen_poli ($r = 1.000$), and tarif_INACBGs–tagihan_verif–tagihan_tl ($r > 0.99$). Consequently, the elimination of superfluous categorical variables (layanan, seg_tl, tagihan_tl, and tagihan_verif) was a judicious decision, given that no significant collinearity indicators were seen among the numerical variables. The amalgamation of correlation and VIF outcomes indicates that the dataset, post-preprocessing, exhibits enhanced stability: elevated correlations among categorical and financial variables were alleviated through variable selection, whereas numerical variables were preserved as they were devoid of multicollinearity. It is essential to guarantee that the subsequently constructed prediction models remain untainted by superfluous information.

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 10, No. 4, December 2025, Pp. 3772-3786

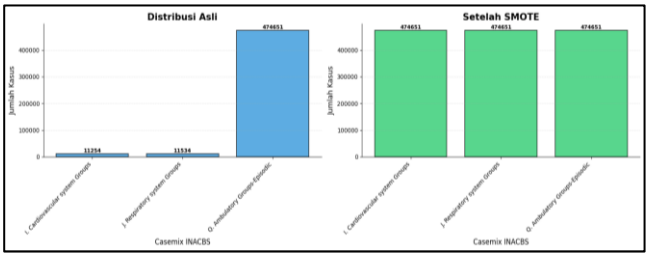


Fig. 9. Comparison of class distribution before and after SMOTE over-sampling.

Figure 9 depicts the comparison of target class distributions for Casemix INACBGs prior to and after to balancing with SMOTE (Synthetic Minority Oversampling Technique). The original distribution exhibited a significant imbalance among the classes. The Q. Ambulatory Groups — The episodic class comprised 474,651 patients, whereas the I. The Cardiovascular System Groups comprised merely 11,254 instances, together with the J. The Respiratory System Groups comprised 11,534 instances. Such inequalities threaten to skew the predictive model, as algorithms typically "learn" more from the dominant class while disregarding the minority classes.

Following the use of SMOTE (depicted on the right-hand graph), the distribution achieved equilibrium, with each of the three classes comprising 474,651 instances. This was accomplished by creating synthetic samples for the minority classes, thus equalizing class proportions. A balanced dataset enables the classification model to more effectively identify patterns in the minority classes (cardiovascular and respiratory) without unduly favoring the majority class. The implementation of SMOTE effectively resolved the class imbalance problem by equalizing the case distribution among classes. This ultimately enhances the probability of attaining equitable and balanced model performance, especially in metrics like F1 Score and Balanced Accuracy.

B. Modeling Outcomes

In the dataset devoid of data exploration treatments, bagging-based ensemble models (Random Forest, Extra Trees) exhibited superior performance relative to boosting algorithms (XGBoost, LightGBM) concerning Balanced Accuracy (BA) and G-mean, as illustrated in Figures 7 and 8. The Random Forest model attained a balanced accuracy (BA) of 0.9270 and a geometric mean (G-mean) of 0.9256, surpassing the performance of LightGBM (0.9136, 0.9116) and XGBoost (0.8809, 0.8772). The results indicate that bagging approaches exhibit more stability in preserving balanced detection across classes in the presence of extremely imbalanced data distributions, whereas boosting methods provide increased sensitivity to the majority class.

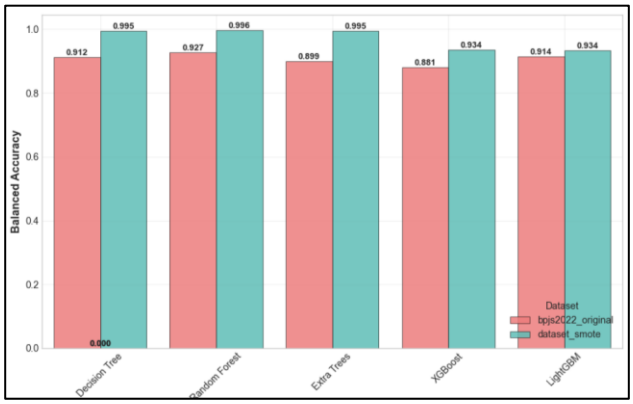


Fig 10. Comparison of modeling outcomes between unprocessed data and preprocessed data utilizing Balanced Accuracy (BA) as the evaluative metric.

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 10, No. 4, December 2025, Pp. 3772-3786

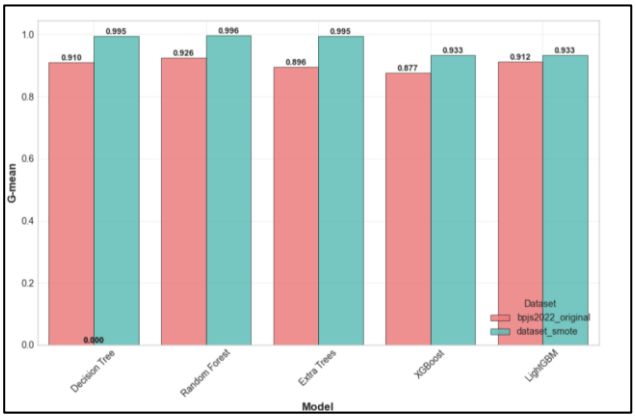


Fig 11. Comparison of modeling outcomes between unprocessed data and preprocessed data utilizing G-mean as the evaluation metric.

Figures 10 and 11 illustrate that, following the management of outliers, multicollinearity, and class imbalance, there was a significant enhancement in the efficacy of all bagging techniques. The Random Forest model exhibited an increase in BA from 0.9270 to 0.9959 and in G-mean from 0.9256 to 0.9959. Both Decision Tree and Extra Trees demonstrated analogous patterns of enhancement. Simultaneously, boosting algorithms (XGBoost, LightGBM), albeit demonstrating enhancements, attained comparatively lower metrics (BA ~0.934, G-mean ~0.933).

The significant enhancement noted in bagging approaches can be attributed to the characteristics of these algorithms. Bagging constructs multiple decision trees from bootstrap samples, significantly enhancing performance when minority classes are enriched by SMOTE. With an equitable class distribution, bagging can create more thorough representations of all classes without significant bias towards the majority. In contrast, boosting, which focuses on iterative error correction, is sometimes sensitive to mixed feature distributions (categorical and numerical), leading to suboptimal results relative to bagging.

TABLE II
COMPARISON OF BALANCED ACCURACY AND G-MEAN BETWEEN UNTREATED AND PREPROCESSED DATA

Variable	Without Preprocessing		With Preprocessing	
	BA	BA	G-mean	G-mean
Decision Tree	0.9117	0.9955	0.9096	0.9955
Random Forest	0.927	0.9959	0.9256	0.9959
Extra Trees	0.899	0.9949	0.8963	0.9949
XGBoost	0.8809	0.9344	0.8772	0.9334
LightGBM	0.9136	0.9338	0.9116	0.9325

Table II indicates that following the rectification of outliers, multicollinearity, and class imbalance, there was a significant enhancement in the efficacy of all bagging techniques. The Random Forest model exhibited an enhancement in BA from 0.9270 to 0.9959 and in G-mean from 0.9256 to 0.9959. Decision Tree and Extra Trees demonstrated analogous enhancement trends. Concurrently, boosting algorithms (XGBoost, LightGBM), despite enhancements, attained comparatively lower metrics (BA ~0.934, G-mean ~0.933).

The performance improvement noted in bagging can be attributed to the algorithm's intrinsic characteristics. Bagging generates several decision trees from bootstrap samples and notably benefits from the augmentation of minority classes through SMOTE. With an equitable class distribution, bagging can construct more thorough representations of all classes without significant bias towards the majority. Conversely, boosting, which focuses on iterative error correction, is generally more susceptible to mixed feature distributions (categorical and numerical), leading to suboptimal enhancements relative to bagging.

The findings imply that dataset exploration and preprocessing are essential processes that directly improve model quality. In untreated datasets, despite Random Forest yielding satisfactory Balanced Accuracy and G-mean, the model demonstrated a bias towards the majority class—an issue that is especially concerning in healthcare sectors where high precision is critical for decision-making. Following comprehensive data exploration and preprocessing, notable enhancements in Balanced Accuracy and G-mean were attained across almost all models, particularly in bagging. This illustrates that dataset exploration processes, including outlier detection, multicollinearity assessment, and class imbalance management, are essential for generating more equitable and accurate models, hence minimizing prediction mistakes in healthcare applications.

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 10, No. 4, December 2025, Pp. 3772-3786

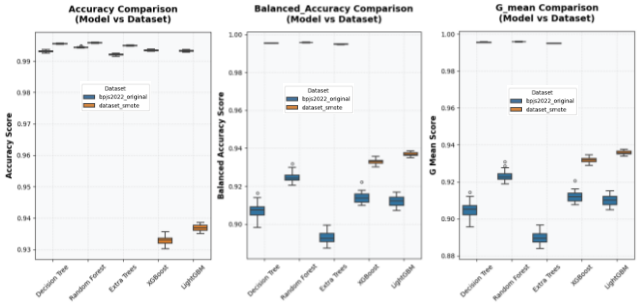


Fig. 12. Boxplot illustrating model performance measured by Accuracy, Balanced Accuracy, and G-mean over 20 iterations.

Figure 12 indicates that the overall accuracy of all models for the untreated dataset was exceedingly high (>0.99). Nonetheless, the more pertinent metrics for unbalanced data—Balanced Accuracy (BA) and G-mean—indicated a decline in performance. The Random Forest model, identified as the most effective, attained a balanced accuracy of 0.9247, but other models, including Extra Trees, had inferior performance with a score of 0.8928. This signifies a predisposition towards the majority class and an insufficient equilibrium in acknowledging minority classes.

TABLE III
MODELING RESULTS FROM 20 ITERATIONS

Dataset	Model	Accuracy (±SD)	Balanced Accuracy (±SD)	G-mean (±SD)
Without Preprocessing	Decision Tree	0.9931 ± 0.0003	0.9070 ± 0.0042	0.9047 ± 0.0044
Without Preprocessing	Extra Trees	0.9920 ± 0.0002	0.8928 ± 0.0031	0.8897 ± 0.0033
Without Preprocessing	LightGBM	0.9933 ± 0.0002	0.9120 ± 0.0027	0.9099 ± 0.0028
Without Preprocessing	Random Forest	0.9944 ± 0.0002	0.9247 ± 0.0026	0.9232 ± 0.0027
Without Preprocessing	XGBoost	0.9934 ± 0.0002	0.9139 ± 0.0029	0.9119 ± 0.0030
With Preprocessing	Decision Tree	0.9955 ± 0.0001	0.9955 ± 0.0001	0.9955 ± 0.0001
With Preprocessing	Extra Trees	0.9949 ± 0.0001	0.9949 ± 0.0001	0.9949 ± 0.0001
With Preprocessing	LightGBM	0.9369 ± 0.0009	0.9369 ± 0.0009	0.9358 ± 0.0010
With Preprocessing	Random Forest	0.9958 ± 0.0001	0.9958 ± 0.0001	0.9958 ± 0.0001
With Preprocessing	XGBoost	0.9329 ± 0.0014	0.9329 ± 0.0014	0.9318 ± 0.0014

Table 3 presents the post-preprocessing, a notable enhancement was evident, especially in bagging techniques (Decision Tree, Random Forest, Extra Trees). Both Balanced Accuracy and G-mean approached near-perfect levels (≥0.9949), with Random Forest once more identified as the top performer (0.9958). These findings affirm that data exploration and preprocessing are essential processes that directly enhance model performance in the equitable classification of all categories. Conversely, boosting approaches (XGBoost and LightGBM) did not attain comparable enhancements. Their performance significantly trailed behind bagging, with BA values of approximately 0.93–0.94. This indicates that in mixed categorical–numerical datasets, such as BPJS claims data, bagging is more effective at utilizing the balanced class distribution, whereas boosting is more responsive to feature distributions and necessitates additional tweaking. The results confirm that bagging, namely Random Forest, is the most reliable and precise option for this dataset, particularly when Balanced Accuracy and G-mean are emphasized as primary assessment measures. An exploratory data analysis was performed before modeling to comprehend the essential aspects of the 2022 BPJS claims dataset. The dataset has 497,439 observations and 27 variables, including 13 categorical and 14 numerical variables. The distribution of variables is depicted in Figures 1 and 2.

IV. CONCLUSION

This study illustrates that data exploration and preprocessing are indispensable for improving the efficacy of ensemble models in diagnosing cardiovascular and chronic respiratory disorders using BPJS health claims data. The preliminary analysis revealed that the dataset possessed complex characteristics, including highly skewed class distributions, numerous outliers in financial and service frequency variables, and strong correlations among

certain predictors. These characteristics can create estimation bias during model training and reduce generalizability. The systematic implementation of data exploration techniques winsorizing, correlation, and VIF analysis, variable selection, and oversampling with SMOTE proved effective in enhancing data quality. Winsorizing mitigated extreme values in demographic and length-of-stay variables, while SMOTE corrected the previously imbalanced target distribution. Correlation and VIF analysis aided in identifying and removing redundant variables contributing to multicollinearity. Collectively, these processes produced a more representative and stable dataset, optimally prepared for modeling.

Comparative analyses demonstrated that bagging-based ensemble models, namely Decision Tree, Random Forest, and Extra Trees, achieved the most substantial improvements after data preprocessing. Their Balanced Accuracy and G-mean metrics approached perfection (≥0.9949), with Random Forest consistently exhibiting the highest accuracy and stability across multiple trials. In contrast, boosting approaches such as XGBoost and LightGBM showed moderate yet noticeable enhancements. These findings suggest that in mixed numerical–categorical datasets characterized by severe class imbalance, bagging ensembles demonstrate superior robustness to rectified data distributions derived through systematic exploration. The findings emphasize that rigorous data exploration is a critical determinant of predictive model performance. Within the BPJS healthcare system, these practices have direct practical implications for achieving fairer, more accurate, and more dependable predictions. Such models can assist policymakers and healthcare professionals in early disease detection, targeted intervention planning, and maintaining the sustainability of the national health insurance system.

Future research should consider alternative outlier-handling methods—such as logarithmic transformation, robust scaling, or model-based techniques like Isolation Forest—to evaluate their comparative effectiveness against winsorizing in stabilizing financial and service frequency variables. Further studies could also explore advanced resampling techniques beyond SMOTE, including ADASYN, Borderline-SMOTE, or hybrid approaches like SMOTE-ENN, to assess their influence on model stability and generalizability. Finally, applying interpretability algorithms such as SHAP and LIME would ensure that predictive models not only deliver high accuracy but also elucidate the key risk factors underlying cardiovascular and respiratory disease claims.

ACKNOWLEDGMENTS

The author wishes to convey profound appreciation to the Department of Statistics and Data Science, School of Graduate Studies, IPB University, for the help and resources extended during this project. Profound gratitude is expressed to Dr. Anwar Fitrianto, M.Sc., and Dr. Erfiani, M.Si., lecturers of the Data Exploration and Visualization course, for their direction, supervision, and invaluable input that significantly facilitated the successful completion of this project.

REFERENCES

[1] M. A. Naser, A. A. Majeed, M. Alsabah, T. R. Al-Shaikhli, dan K. M. Kaky, "A Review of Machine Learning's Role in Cardiovascular Disease Prediction: Recent Advances and Future Challenges," *Algorithms*, vol. 17, no. 2, hlm. 78, Feb 2024, doi: 10.3390/a17020078.

[2] Md. S. Sikder dan Engr. Md. E. Uddin, "Machine Learning-Based Cardiovascular Disease Prediction: Comparative Analysis of SMOTE Impact on Imbalanced Healthcare Data," *PREPRINT (Version 1)* , Agu 2025, doi: 10.21203/rs.3.rs-7428299/v1.

[3] A. Sihabuddin, N. Rokhman, dan E. E. Wahyudi, "A Machine Learning Approach on Outlier Removal for Decision Tree Regression Method," *Ingénierie des systèmes d information*, vol. 29, no. 4, hlm. 1397–1403, Agu 2024, doi: 10.18280/isi.290414.

[4] W. Yoo, R. Mayberry, S. Bae, K. Singh, Q. Peter He, dan J. J. Lillard, "A Study of Effects of MultiCollinearity in the Multivariable Analysis," *Int J Appl Sci Technol*, vol. 4, no. 5, hlm. 9–19, 2014.

[5] H. Yıldırım, "The Multicollinearity Effect on the Performance of Machine Learning Algorithms: Case Examples in Healthcare Modelling," *Academic Platform Journal of Engineering and Smart Systems*, vol. 12, no. 3, hlm. 68–80, Sep 2024, doi: 10.21541/apjess.1371070.

[6] J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan, dan A. Saboor, "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare," *IEEE Access*, vol. 8, hlm. 107562–107582, 2020, doi: 10.1109/ACCESS.2020.3001149.

[7] A. Tiwari, A. Chugh, dan A. Sharma, "Ensemble framework for cardiovascular disease prediction," *Comput Biol Med*, vol. 146, hlm. 105624, Jul 2022, doi: 10.1016/j.compbiomed.2022.105624.

[8] J. W. Tukey, "The Future of Data Analysis," *The Annals of Mathematical Statistics*, vol. 33, no. 1, hlm. 1–67, Mar 1962, doi: 10.1214/aoms/1177704711.

[9] R. R. . Wilcox, *Introduction to robust estimation and hypothesis testing*. Academic Press, 2012.

[10] S. Han, K. Kim, dan S. Jung, "Subspace Recovery in Winsorized PCA: Insights into Accuracy and Robustness," Feb 2025, [Daring]. Tersedia pada: http://arxiv.org/abs/2502.16391

[11] G. Lafit, F. Nogales, M. Ruiz, dan R. Zamar, "Robust graphical lasso based on multivariate Winsorization," Jan 2022, [Daring]. Tersedia pada: http://arxiv.org/abs/2201.03659

[12] P. Orenstein, "Robust Importance Sampling with Adaptive Winsorization," Feb 2021, [Daring]. Tersedia pada: http://arxiv.org/abs/1810.11130

[13] L. Yang, X. Zhang, dan J. Chen, "Winsorization greatly reduces false positives by popular differential expression methods when analyzing human population samples," *Genome Biol*, vol. 25, no. 1, Des 2024, doi: 10.1186/s13059-024-03230-w.

[14] N. V. Chawla, K. W. Bowyer, L. O. Hall, dan W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, hlm. 321–357, Jun 2002, doi: 10.1613/jair.953.

[15] Y. Li, Y. Yang, P. Song, L. Duan, dan R. Ren, "An improved SMOTE algorithm for enhanced imbalanced data classification by expanding sample generation space," *Sci Rep*, vol. 15, no. 1, Des 2025, doi: 10.1038/s41598-025-09506-w.

[16] T. Kosolwattana, C. Liu, R. Hu, S. Han, H. Chen, dan Y. Lin, "A self-inspected adaptive SMOTE algorithm (SASMOTE) for highly imbalanced data classification in healthcare," *BioData Min*, vol. 16, no. 1, Des 2023, doi: 10.1186/s13040-023-00330-4.

[17] F. Duan, S. Zhang, Y. Yan, dan Z. Cai, "An Oversampling Method of Unbalanced Data for Mechanical Fault Diagnosis Based on MeanRadius-SMOTE," *Sensors*, vol. 22, no. 14, Jul 2022, doi: 10.3390/s22145166.

[18] M. Kivrak, U. Avci, H. Uzun, dan C. Ardic, "The Impact of the SMOTE Method on Machine Learning and Ensemble Learning Performance Results in Addressing Class Imbalance in Data Used for Predicting Total Testosterone Deficiency in Type 2 Diabetes Patients," *Diagnostics*, vol. 14, no. 23, hlm. 2634, Nov 2024, doi: 10.3390/diagnostics14232634.
[19] L. Breiman, J. H. Friedman, R. A. Olshen, dan C. J. Stone, *Classification And Regression Trees*. Routledge, 1984. doi: 10.1201/9781315139470.
[20] F. Pedregosa *dkk.*, "Scikit-learn: Machine Learning in Python," Jun 2011.
[21] L. Breiman, "Random Forests," *Mach Learn*, vol. 45, no. 1, hlm. 5–32, 2001, doi: 10.1023/A:1010933404324.
[22] P. Geurts, D. Ernst, dan L. Wehenkel, "Extremely randomized trees," *Mach Learn*, vol. 63, no. 1, hlm. 3–42, Apr 2006, doi: 10.1007/s10994-006-6226-1.
[23] G. Ke *dkk.*, "LightGBM: a highly efficient gradient boosting decision tree," dalam *Proceedings of the 31st International Conference on Neural Information Processing Systems*, dalam NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, hlm. 3149–3157.
[24] G. Ke *dkk.*, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," dalam *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, dan R. Garnett, Ed., Curran Associates, Inc., 2017.
[25] C. F. Dormann *dkk.*, "Collinearity: a review of methods to deal with it and a simulation study evaluating their performance," *Ecography*, vol. 36, no. 1, hlm. 27–46, Jan 2013, doi: 10.1111/j.1600-0587.2012.07348.x.