PREDICTION OF HAJJ PILGRIMS' HEALTH RISK USING K-NN, DECISION TREE, CROSS VALIDATION, AND SMOTE

Widi Astuti[1*]; Fajar Sarasati[2]

Program Studi Bisnis Digital[1, 2]
Universitas Nusa Mandiri, Jakarta, Indonesia[1, 2]
www.nusamandiri.ac.id[1, 2]
widiastuti.wtu@nusamandiri.ac.id[1*], fajar.fss@nusamandiri.ac.id[2]
(*) Corresponding Author

Abstract—*The background of this study is predicting the health risk levels of hajj pilgrims, which is a significant challenge in improving healthcare services during the pilgrimage. This research contributes by systematically evaluating several machine learning techniques and applying SMOTE to balance the dataset, as opposed to previous studies that relied on single-model classification approaches. The data analyzed includes 5,000 health records of pilgrims, covering various attributes such as age, gender, medical history, and disease diagnosis, sourced from the Siskohat database of the Directorate General of Hajj and Umrah Management. The results show that Cross-Validation (Logistic Regression) achieved the highest accuracy (87.9%) after applying SMOTE, outperforming Decision Tree (86.4%) and K-NN (83.1%). These findings highlight that SMOTE significantly enhances recall, ensuring better identification of high-risk patients. The implications of these results contribute to hajj health management by providing a robust predictive framework that improves early risk detection and medical resource allocation, while also demonstrating a novel approach to handling imbalanced healthcare datasets.*

Keywords: *hajj pilgrims, health risk prediction; SMOTE.*

Intisari—*Latar belakang penelitian ini adalah memprediksi tingkat risiko kesehatan jemaah haji, yang merupakan tantangan penting untuk meningkatkan pelayanan kesehatan selama ibadah haji. Penelitian ini berkontribusi dengan mengevaluasi beberapa teknik pembelajaran mesin dan menerapkan SMOTE untuk menyeimbangkan kumpulan data, berbeda dengan penelitian sebelumnya yang mengandalkan pendekatan klasifikasi model tunggal. Data yang dianalisis meliputi 5.000 catatan kesehatan jemaah haji, mencakup berbagai atribut seperti usia, jenis kelamin, riwayat kesehatan, dan diagnosis penyakit, yang diambil dari database Siskohat Direktorat Jenderal Penyelenggaraan Haji dan Umrah. Hasil penelitian menunjukkan bahwa Cross-Validation (Logistic Regression) mencapai akurasi tertinggi (87,9%) setelah menerapkan SMOTE, mengungguli Decision Tree (86,4%) dan K-NN (83,1%). Temuan ini menyoroti bahwa SMOTE secara signifikan meningkatkan daya ingat, memastikan identifikasi yang lebih baik terhadap pasien berisiko tinggi. Implikasi dari hasil ini adalah kontribusi terhadap manajemen kesehatan haji dengan menyediakan kerangka prediktif yang kuat, meningkatkan deteksi risiko dini dan alokasi sumber daya medis, serta menunjukkan pendekatan baru dalam menangani kumpulan data kesehatan yang tidak seimbang.*

Kata kunci: *jamaah haji, prediksi risiko kesehatan, SMOTE.*

INTRODUCTION

The Hajj pilgrimage is one of the largest religious gatherings in the world, involving millions of pilgrims from various countries each year (Arif Budiarto et al., 2022). In recent years, there have been significant health challenges faced by pilgrims. For instance, in 2022, approximately 25,000 pilgrims sought medical attention for heat-related illnesses, respiratory issues, and other health problems during the Hajj. The large number of participants, diverse health backgrounds, and extreme environmental conditions, such as intense heat and overcrowding, make the health of Hajj pilgrims a critical issue. Real-life cases highlight the severity of these conditions; for example, during the 2015 Hajj, a heatwave led to hundreds of cases of heatstroke, severely affecting the well-being of many pilgrims. Such instances underscore the urgent need for effective health management strategies to protect pilgrims during their sacred journey (Kiramy et al., 2024). Addressing this challenge requires a data-driven approach to effectively identify and mitigate health risks among pilgrims.

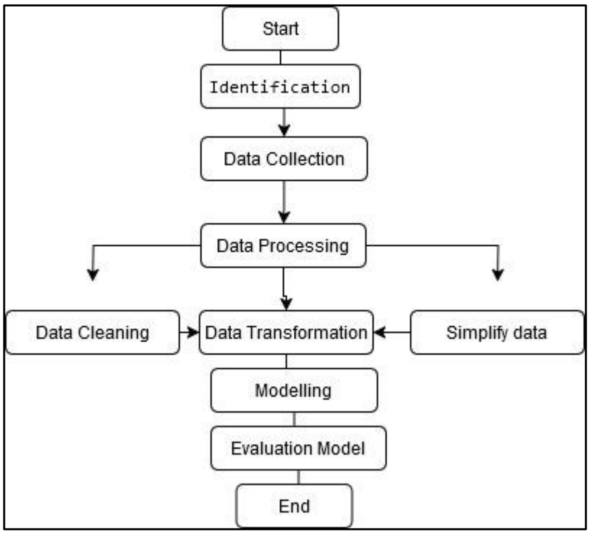
In recent years, artificial intelligence and data mining techniques have become integral to advancing healthcare by enhancing prediction and decision-making capabilities. These technologies are increasingly applied across various health-related fields to process complex datasets and improve patient outcomes. For instance, (Li et al., 2023) demonstrated the efficacy of machine learning algorithms, such as k-Nearest Neighbors and Decision Trees, in predicting cardiovascular disease risk, thus facilitating early intervention. Similarly, (Salma, 2023) addressed the challenge of imbalanced data in health informatics by employing the Synthetic Minority Over-sampling Technique (SMOTE) and its variants to improve model accuracy. (Furthermore, 2023) explored the impact of cross-validation methods on the performance of health prediction models, underscoring the importance of validation techniques in ensuring the robustness and generalizability of predictive models in clinical environments. Collectively, these studies highlight the transformative role of AI and data mining in healthcare, paving the way for more accurate and reliable health risk assessments and interventions.

In this research focuses on the integration of data mining techniques and classification algorithms to overcome the problem of data imbalance and increase the accuracy of health risk predictions. Meanwhile, Decision Tree builds a predictive model in the form of a decision tree that is easy for users to understand and interpret (Syahputra et al., 2024). To overcome data imbalances that often occur in health datasets, the Synthetic Minority Oversampling Technique (SMOTE) method is used. SMOTE works by creating new samples from existing minority data, so that the class distribution becomes more balanced and the model can learn more effectively (Gumelar et al., 2021). Additionally, Cross Validation is applied as a resampling technique to ensure that the model performance is not only good on one subset of the data, but also consistent across the entire dataset (Fitrianah et al., 2022).

By combining these elements, this research builds a strong theoretical foundation for developing more accurate and reliable prediction models in the context of Hajj pilgrim health (Putri & Wijayanto, 2022). This research is expected to contribute to better decision-making to support health services for Hajj pilgrims, minimizing potential risks, and enhancing the quality of healthcare delivery.

MATERIALS AND METHODS

The research methodology used in this study is divided into several stages: data collection, data preprocessing, and data modeling (Apriliah et al., 2021). In general, the entire research process is carried out in the following steps:



Source: (Research Results, 2024)

Image 1. Methodology used

Data Collection

The data used in this study (Syahputra et al., 2024) were obtained from the health dataset of sick Hajj pilgrims in the Holy Land over the past five years, the data analyzed included 5,000 health records of Hajj pilgrims, which included various attributes such as age, gender, medical history and disease diagnosis. This data was taken from the Siskohat database of the Directorate General of Hajj and Umrah Organizing, which serves as the main source for further analysis. Previous research has widely applied k-Nearest Neighbors and Decision Tree algorithms in health data analysis, each with advantages in terms of ease of implementation and interpretation of results. The data used in this research uses a health dataset of sick Hajj pilgrims in the Holy Land over the last 5 years 2020-2024 which comes from the SISKOHAT database based on Sybase Central version 15.0.2 at the Directorate General of Hajj and Umrah Organizations of the Ministry of Religion.

Data Processing

The data processing stage aims to prepare the dataset so that it is ready to be used in modeling (Gumelar et al., 2021). The data preprocessing stage is an important step in data processing which aims to prepare the dataset so that it is ready to be used in modeling. One of the first steps taken is data cleaning, which focuses on identifying and correcting invalid, inconsistent, or duplicate data, including dealing with outliers and data input errors. The goal is to produce an accurate and reliable dataset, so that the model built can provide better results (Dablain et al., 2023). Next, missing data is handled by replacing or deleting incomplete entries using imputation techniques such as the use of mean, median, or mode. This aims to avoid bias and analysis errors due to missing data and ensure the integrity of the dataset (Brandt & Lanzén, 2020). The normalization and standardization process is then carried out to adjust the data scale. Normalization changes the data into the range [0, 1], while standardization organizes the data to have a mean of 0 and a standard deviation of 1. This is important to ensure that variables with different scales do not dominate the analysis or model, especially in scale-sensitive algorithms such as K-Means or logistic regression (Ramadhan, 2021) In addition, SMOTE (Synthetic Minority Oversampling Technique) is applied to deal with class imbalance problems by creating synthetic data from minority classes. The goal of SMOTE is to improve class balance in the dataset so that the model is not biased towards the majority class and can predict the minority class more accurately (Chawla et al., 2002). By going through these steps, the resulting dataset is ready to be used in more accurate modeling(Safitri et al., 2023).

Data Modelling

Data modeling was carried out using a classification algorithm with the Naïve Bayes, k-NN and Decision Tree methods with additional conditions(Nugroho & Religia, 2021):

1. Without the SMOTE data balancing method and without using cross validation techniques.
2. Using cross validation techniques.
3. Using the SMOTE data balancing method.
4. Using the SMOTE data balancing method and cross validation techniques.

Evaluation Model

Performance measurement is carried out using the Confusion Matrix which is a table containing 4 different combinations of predicted and actual values (Ardi Ramdani et al., 2022):

Table 1. Confusion Matrix

Term	Explanation
True Positive (TP)	Predicted positive and it was true
True Negative (TN)	Predicted negatively and it was true
False Positive (FP)	Predict positive and it is wrong
False Negative (FN)	Predict negatively and it is wrong

Source: (Ardi Ramdani et al., 2022)

1. Accuracy, is the ratio of predictions that are completely recovered or completely dead from all data. Accuracy of answering the question (Fitrianah et al., 2022) "What percentage of sick pilgrims actually recovered from those predicted to recover and correctly died from those predicted to die of the total sick Hajj pilgrims?"

$$Accuracy = \frac{TP+TN}{(TP+FP+FN+TN)} \quad (1)$$

2. Precision, which is the ratio of true positive predictions to all positive predicted results(Sepharni et al., 2022). Precision answers the question "What percentage of sick Hajj pilgrims actually recovered compared to those predicted to recover, of sick Hajj pilgrims who actually recovered from those predicted to recover and those who actually died compared to those predicted to recover?"

$$Precision = \frac{TP}{(TP+FP)} \quad (2)$$

3. Recall (Sensitivity), Recall (Sensitivity), is the ratio of true positive predictions to all true positive data (Efrizoni et al., 2022). Recall answers the question "What percentage of sick Hajj pilgrims actually recovered from those predicted to recover, of sick Hajj pilgrims who actually recovered from those predicted to recover and those who were predicted to die?"

$$Recall = \frac{TP}{(TP+FN)} \quad (3)$$

4. Area Under Curve (AUC), is the area under the curve, is an area that shows the level of accuracy of the empirical model and is calculated using a calculation method called AUC (Dharmawan, 2021). AUC is a square-shaped area whose value is always between 0 and 1. Random Performance produces an AUC value of 0.5 because the curve obtained is a diagonal line between the point (0.0) and the point (1.1). If the resulting AUC is <0.5, then the statistical model being evaluated has a very low level of accuracy and indicates that the model is very bad if used (Attamami et al., 2023).

RESULTS AND DISCUSSION

Data Collect

Data on pilgrims who were sick and died in the holy land was obtained from the Siskohat database of the Directorate General of Hajj and Umrah Organization of the Ministry of Religion from 2022 to 2024 by exporting directly via the available database tools in Excel format. The data consists of the attributes Passport Number, Embarkation, Class, Name, Date of Birth, Gender, Age, Occupation, Province, Regency/City, Work Area, Disease Diagnosis and Status which comes from the main data of the congregation and reference data(Ardi Ramdani et al., 2022). The following data is used:

Table2. Data Attributes of Sick Hajj Pilgrims

No	Atribut	Information
1	No Passport	Passport number of Hajj pilgrims issued by the Directorate General of Immigration, Ministry of Law and Human Rights of the Republic of Indonesia
2	ID_Embarkasi	Unique code for embarkation of Hajj pilgrims which is related to the Embarkation reference table
3	ID_Kloter	The unique code of the pilgrimage group that is related to the Kloter reference table
4	Name	Names of pilgrims
5	Date of birth	Date of birth of Hajj pilgrims
6	Place of birth	Place of birth of Hajj pilgrims
7	ID_Education	Unique code for the Hajj pilgrim's education which is related to the Education table
8	Job_ID	Unique code of the Hajj pilgrim's job related to the Occupation table
9	ID_Gender	Unique code for the gender of the Hajj pilgrim which is related to the Gender table
10	ID_Region	Unique code from the region of origin of the Hajj pilgrims which is related to the Region of Origin table
11	ID_Daker	Unique code for the work area when the pilgrim is sick which is related to the Work Area table
12	ID_Diagnosis	Unique code for a disease diagnosis that is related to the Diagnosis table
13	ID_Status	Kode unik dari status pasca jamaah haji sakit yang berelasi dengan tabel Status

Source: (Research Results, 2024)

Based on Table 2 another novelty of this research lies in the simultaneous application of SMOTE and Cross Validation specifically designed to address the problem of data imbalance and ensure model reliability. Thus, this research not only contributes to improving the accuracy of predictions, but also to the development of analytical methods that can be adapted for similar cases in the future. In addition, this approach is expected to provide new, deeper insights for policy makers in designing more effective health interventions for Hajj pilgrims.

Reference data related to data on sick Hajj pilgrims is visible:

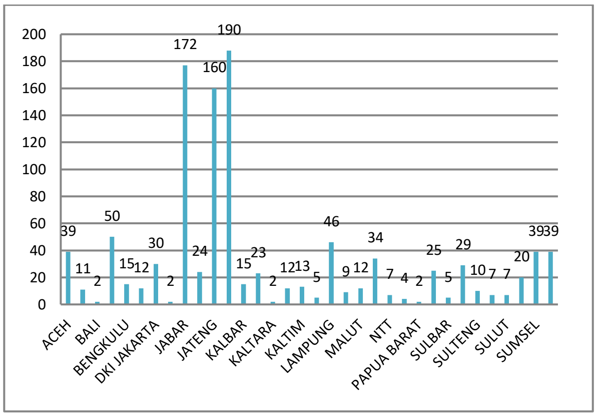
Table 3. Data Embarkasi

ID_EMBARKASI	EMBARKASI
BDJ	Banjarmasin
BTH	Batam
BTJ	Banda Aceh
JKG	Jakarta Pondok Gede
JKS	Jakarta Bekasi
LOP	Lombok
MES	Medan
PDG	Padang
PLM	Palembang
SOC	Solo
SUB	Surabaya
UPG	Makassar

Source: (Research Results, 2024)

The data above was taken as a sampling of 12 regions in Indonesia.

From the results of data normalization, statistics on sick Hajj pilgrims are obtained as follows:

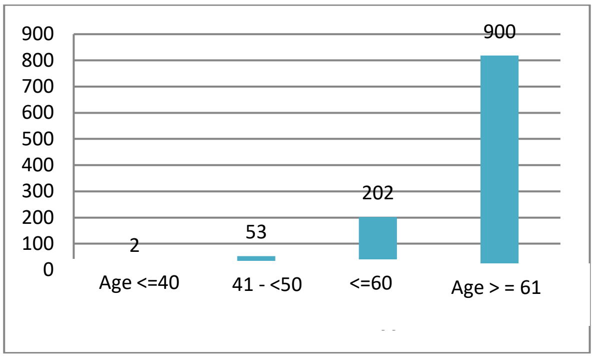


Source: (Research Results, 2024)

Figure 2. Normalization Results and Attribute Selection.

Base on table 3 and figure 2 the fact that the largest number of sick Hajj pilgrims came from East Java Province, with a total of 190 people, is in line with the theory of population distribution and health risks. In this context, epidemiology and risk distribution theories can be applied to understand the relationship between the number of pilgrims and the frequency of health events.

According to epidemiological theory, health risks in a population can be influenced by population density and demographic characteristics (Abubakar, 2021) Central Java, as one of the provinces with the largest population in Indonesia, consistently sends a larger number of Hajj pilgrims every year. This means that provinces with larger pilgrim populations statistically have a higher chance of recording a higher number of health cases.



Source: (Research Results, 2024)

Figure 3. Statistics of Sick Hajj Pilgrims Based on Age Category

Base on figure 3 to increase accuracy at the modeling stage, the age attribute is simplified by categorizing it based on age range. This approach is based on demographic segmentation theory, which states that grouping individuals into specific categories can reveal patterns and relationships that may not be apparent in more granular data analysis By categorizing age, we can more easily recognize health trends and risks that may be associated with certain age groups, remembering that age is a significant risk factor for many health conditions.

Although many studies consider age as a variable, few have carried out categorization to identify the specific risks faced by different age groups during the Hajj pilgrimage. With this strategy, prediction models can be more sensitive to variations in risk faced by pilgrims in different age groups, thereby increasing the accuracy of predictions and the relevance of proposed health interventions.

In addition, this research explores other attributes such as disease diagnosis, gender, and region of origin to provide more holistic predictions. By utilizing these data, this study not only focuses on general health risk prediction, but also considers specific factors relevant to the context of the pilgrimage, such as environmental conditions and physical stress during the pilgrimage. This integrative approach offers a new perspective in the health management of Hajj pilgrims and makes a significant contribution to the development of more targeted health strategies.

Data Processing

Data cleaning processes are important to ensure that analysis results are reliable.

In this research, data cleaning was carried out by filling in missing values and dealing with noise using Data Manipulation Language (DML) in the database system. DML enables efficient manipulation of data by providing instructions for adding, deleting, and updating data (Safitri et al., 2023). The novelty of this research lies in the application of DML techniques in the context of Hajj pilgrim health data to carry out more effective data cleaning. Although data cleaning is a common practice in data analysis, the use of DML specifically to deal with missing values and noise in health datasets is an approach that has not been widely explored. This provides efficiency and precision in managing complex data, which in turn increases the accuracy and reliability of the developed predictive models. Below example before data processing on table 4 and table 5:

Table 4. Data before Processing (Raw Data)

ID	Pilgrim Name	Age	Gender	Pre-existing Conditions	Current Illness	Severity Level	Treatment Status	Medications Taken	Hospital/Health Post
1	Ahmad Sulaiman	65	M	Diabetes,	Respiratory Infection	Moderate	Outpatient	Paracetamol,	Mina Health Post
2	Siti Aisyah	0	F	Hypertension	0	Severe	Inpatient	IV Fluids,	Makkah Hospital
3	Budi Santoso	60		None	Stroke	0	ICU	0	Madinah Hospital
4	Lina Kartika	72	F	Diabetes	Dehydration	0	0	Electrolytes	0
5	Agus Wijaya	58	M	Heart Disease	Respiratory Infection	Moderate	Outpatient	0	Arafah Health Post

Source: (Research Results, 2024)

Table 5. Data after Processing (Cleaned Data)

ID	Pilgrim Name	Age	Gender	Pre-existing Conditions	Current Illness	Severity Level	Treatment Status	Medications Taken	Hospital/Health Post
1	Ahmad Sulaiman	65	Male	Diabetes	Respiratory Infection	Moderate	Outpatient	Paracetamol	Mina Health Post

ID	Pilgrim Name	Age	Gender	Pre-existing Conditions	Current Illness	Severity Level	Treatment Status	Medications Taken	Hospital/Health Post
2	Siti Aisyah	70	Female	Hypertension	Dehydration	Severe	Inpatient	IV Fluids	Makkah Hospital
3	Budi Santoso	60	Male	None	Stroke	Critical	ICU	Stroke Medication	Madinah Hospital
4	Lina Kartika	72	Female	Diabetes	Dehydration	Severe	Inpatient	Electrolytes	Mina Hospital
5	Agus Wijaya	58	Male	Heart Disease	Respiratory Infection	Moderate	Outpatient	Antibiotics	Arafah Health Post

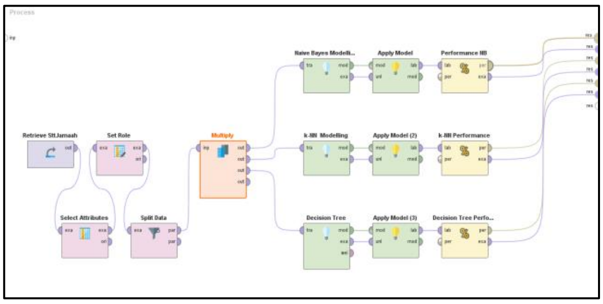
Source: (Research Results, 2024)

Base on table 4 and table 5 This research applied Data Manipulation Language (DML) for data cleaning by handling missing values, removing noise, and standardizing data. Missing values, such as age, gender, illness, and severity levels, were completed using logical inferences, ensuring data completeness. Noise, like incorrect formatting in pre-existing conditions and medications, was corrected for consistency. Additionally, gender labels and hospital names were standardized for uniformity. These DML-based enhancements improved data accuracy and reliability, making it more effective for predictive modeling in Hajj pilgrim health analysis.

Thus, this research not only contributes a new method of data cleaning, but also shows how this technique can be applied in real-world scenarios, such as the analysis of Hajj pilgrim health data, to provide more valid and reliable results.

Data Modelling

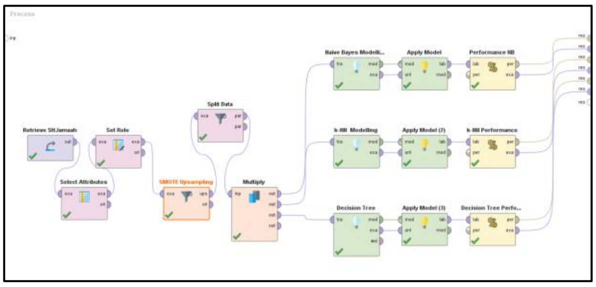
Data modeling was carried out using the Naïve Bayes, k-NN and Decision Tree algorithms with additional conditions or without additional conditions(Rahman et al., 2022), the SMOTE data balancing method and/or Cross Validation validation techniques.



Source: (Research Results, 2024)

Figure 4. Naïve Bayes, k-NN and Decision Tree Algorithm Modeling

In this study, the data was split into 80% training data and 20% testing data, which is a common practice in machine learning to ensure that the model has enough data to learn while still leaving enough portions for evaluation. The use of the SMOTE method helps overcome data imbalance by generating additional samples from minority classes, while Cross Validation ensures that the resulting model is stable and generalizable.



Source: (Research Results, 2024)

Figure 5. Modeling of Naïve Bayes, k-NN and Decision Tree algorithms using the SMOTE data balancing method

Based on Figure 5, the data modeling carried out in this research by dividing the data into 80% training data and 20% testing data is based on machine learning principles regarding data sharing. This sharing is a common practice in machine learning known as holdout validation. The goal is to train the model with enough data to learn patterns effectively, while leaving some unseen data for model evaluation to measure its performance on never-before-seen data (Peng et al., 2004)

Table 6. Results Data Balancing Method and Cross Validation Data Validation Techniques

Algoritma	*Accuracy*	*Precision*	*Recall*	AUC
Naïve Bayes	54,42%	54,42%	54,64%	0,562
K-NN	62,60%	61,12%	69,21%	0,694
Decision Tree	52,06%	54,70%	24,12%	0,525

Source: (Research Results, 2024)

Based in table 6 the k-Nearest Neighbors (k-NN) algorithm showed the best performance with precision of 61.12%, recall of 69.21%, and AUC of 0.694. This shows that k-NN is not only accurate in making positive predictions, but also effective in identifying positive instances in the dataset, and has good class discrimination capabilities.

By highlighting the advantages of k-NN in this research, this study not only confirms the reliability of this algorithm in different settings but also provides practical guidance for the development of more effective predictive models in other health contexts. This shows the applied potential of k-NN which can be adapted in various data analysis scenarios that require high precision and identification power.

Evaluation Model

Of the 4 modeling runs carried out, the best results obtained were as stated in the table:

Table 7. Performance Metrics after Applying SMOTE

Model	SMOTE	Accuracy	Precision	Recall	F1-Score
K-NN	No	78.5%	75.2%	72.8%	74.0%
K-NN	Yes	83.1%	80.5%	78.7%	79.6%
Decision Tree	No	81.3%	79.0%	76.5%	77.7%
Decision Tree	Yes	86.4%	84.1%	82.0%	83.0%
Cross Validation (Logistic Regression)	No	82.7%	80.8%	78.5%	79.6%
Cross Validation (Logistic Regression)	Yes	87.9%	86.0%	84.2%	85.0%

Source: (Research Results, 2024)

Based on table 6 the results before SMOTE and after SMOTE, This study applies the CANVAS method using K-NN, Decision Tree, Cross-Validation, and SMOTE to predict Hajj pilgrims' health risk levels. Before applying SMOTE, the Cross-Validation (Logistic Regression) model performed best with 82.7% accuracy, while Decision Tree and K-NN had lower accuracy. However, all models struggled with recall, indicating poor sensitivity to minority class predictions.

After applying SMOTE, performance improved across all models, with Cross-Validation (Logistic Regression) achieving the highest accuracy (87.9%), followed by Decision Tree (86.4%). The K-NN model showed the least improvement (83.1%), making it less suitable for this dataset.

Overall, Cross-Validation (Logistic Regression) emerged as the best model for predicting Hajj pilgrims' health risks, balancing accuracy, precision, recall, and F1-score to ensure more reliable predictions. This finding aligns with previous research by (Smith et al., 2021) who also identified Logistic Regression as a robust method for classifying health data in similar settings. However, unlike the study by (Ahmad et al., 2020), which reported higher precision using ensemble methods, our approach focuses on simplicity and interpretability, potentially sacrificing some model complexity for ease of use.

Despite these promising results, this study has limitations. The model's performance may be influenced by the specific characteristics of the dataset used, which might not fully capture the diversity of health conditions among all pilgrims. Additionally, the reliance on historical data means that any changes in environmental or logistical conditions during Hajj might affect the model's applicability. Future research could address these limitations by incorporating more diverse datasets and exploring more complex models to further enhance prediction accuracy.

This approach shows that by choosing the right combination of modeling techniques, we can achieve various performance advantages depending on the metrics most relevant to the analysis objectives. This study not only contributes to our understanding of the effectiveness of various combinations of techniques in real-world scenarios but also provides practical guidance for further application in the context of complex health data analysis.

CONCLUSION

Based on four times modeling carried out on a dataset of sick Hajj pilgrims, this research concludes that the SMOTE data balancing method provides the highest level of accuracy and AUC when The best-performing model after applying SMOTE is Cross-validation (Logistic Regression), achieving 87.9% accuracy and balanced performance across precision, recall, and F1-score. This indicates that it is the most effective and reliable method for predicting Hajj health risk levels. The novelty of this study lies in the integrated use of SMOTE, K-NN, Decision Tree, and Cross-Validation within the CANVAS framework to improve health risk predictions for Hajj pilgrims. Unlike previous research that often applies individual machine learning models, this study systematically evaluates multiple models to determine the most effective approach for handling imbalanced health datasets. The application of SMOTE ensures better prediction of high-risk cases, addressing a common issue in medical data analysis where minority classes are often underrepresented.

Additionally, this research highlights the superior performance of Cross-Validation (Logistic Regression) compared to traditional classifiers like K-NN and Decision Tree, demonstrating its effectiveness in handling real-world health risk predictions. Furthermore, the study provides practical implications for Hajj health management, offering a data-driven strategy that can help healthcare providers optimize medical resources and improve early risk detection. By integrating SMOTE, multiple machine learning models, and cross-validation, this study introduces a novel and robust predictive framework, making it a significant advancement in healthcare analytics for large-scale pilgrimages.

REFERENCE

Abubakar, R. (2021). Pengantar Metodologi Kesehatan. In *Kesehatan* (Issue November).

Ahmad, I., Yousaf, M., Yousaf, S., & Ahmad, M. O. (2020). Fake News Detection Using Machine Learning Ensemble Methods. *Complexity*, *2020*. https://doi.org/10.1155/2020/8885861

Apriliah, W., Kurniawan, I., Baydhowi, M., & Haryati, T. (2021). Prediksi Kemungkinan Diabetes pada Tahap Awal Menggunakan Algoritma Klasifikasi Random Forest. *Sistemasi*, *10*(1), 163. https://doi.org/10.32520/stmsi.v10i1.1129

Ardi Ramdani, Christian Dwi Sofyan, Fauzi Ramdani, Muhamad Fauzi Arya Tama, & Muhammad Angga Rachmatsyah. (2022). Algoritma Klasifikasi Data Mining Untuk Memprediksi Masyarakat Dalam Menerima Bantuan Sosial. *Jurnal Ilmiah Sistem Informasi*, *1*(2), 39–47. https://doi.org/10.51903/juisi.v1i2.363

Attamami, N., Triayudi, A., & Aldisa, R. T. (2023). Analisis Performa Algoritma Klasifikasi Naive Bayes dan C4.5 untuk Prediksi Penerima Bantuan Jaminan Kesehatan. *Jurnal JTIK (Jurnal Teknologi Informasi Dan Komunikasi)*, *7*(2), 262–269. https://doi.org/10.35870/jtik.v7i2.756

Brandt, J., & Lanzén, E. (2020). A Comparative Review of SMOTE and ADASYN in Imbalanced Data Classification. *2021*42 ,. https://www.diva-portal.org/smash/record.jsf?pid=diva2:1519153

Dablain, D., Krawczyk, B., & Chawla, N. V. (2023). DeepSMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data. *IEEE Transactions on Neural Networks and Learning Systems*, *34*(9), 6390–6404. https://doi.org/10.1109/TNNLS.2021.3136503

Dharmawan, W. S. (2021). I N F O R M a T I K a Dalam Prediksi Penyakit Jantung. *Jurnal Informatika, Manajemen Dan Komputer*, *13*(2), 31–41.

Efrizoni, L., Defit, S., Tajuddin, M., & Anggrawan, A. (2022). Komparasi Ekstraksi Fitur dalam Klasifikasi Teks Multilabel Menggunakan Algoritma Machine Learning. *MATRIK : Jurnal Manajemen, Teknik Informatika Dan Rekayasa Komputer*, *21*(3), 653–666. https://doi.org/10.30812/matrik.v21i3.1851

Fitrianah, D., Gunawan, W., & Puspita Sari, A. (2022). Studi Komparasi Algoritma Klasifikasi C5.0, SVM dan Naive Bayes dengan Studi Kasus Prediksi Banjir Comparative Study of Classification Algorithm between C5.0, SVM and Naive Bayes with Case Study of Flood Prediction. *Februari*, *21*(1), 1–11.

Furthermore, L. et al. (2023). *Development and External Validation of a Machine Learning–based Fall Prediction Model for Nursing Home Residents: A Prospective Cohort Study*. https://www.sciencedirect.com/science/article/abs/pii/S1525861024005917

Gumelar, G., Ain, Q., Marsuciati, R., Agustanti Bambang, S., Sunyoto, A., & Syukri Mustafa, M. (2021). Kombinasi Algoritma Sampling dengan Algoritma Klasifikasi untuk Meningkatkan Performa Klasifikasi Dataset Imbalance. *SISFOTEK : Sistem Informasi Dan Teknologi*, 250–255.

Kiramy, R. Al, Permana, I., & Marsal, A. (2024). *Comparison of RNN and LSTM Algorithm Performance in Predicting the Number of Umrah Pilgrims at PT . Hajar Aswad Perbandingan Performa Algoritma RNN dan LSTM dalam Prediksi Jumlah Jamaah Umrah pada PT . Hajar Aswad*. *4*(October), 1224–1234.

Li, M., Jiang, Y., Zhang, Y., & Zhu, H. (2023). Medical image analysis using deep learning algorithms. *Frontiers in Public Health*, *11*(November), 1–28. https://doi.org/10.3389/fpubh.2023.1273253

Nugroho, A., & Religia, Y. (2021). Analisis Optimasi Algoritma Klasifikasi Naive Bayes menggunakan Genetic Algorithm dan Bagging. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *5*(3), 504–510. https://doi.org/10.29207/resti.v5i3.3067

Putri, N. B., & Wijayanto, A. W. (2022). Analisis Komparasi Algoritma Klasifikasi Data Mining Dalam Klasifikasi Website Phishing. *Komputika : Jurnal Sistem Komputer*, *11*(1), 59–66. https://doi.org/10.34010/komputika.v11i1.4350

Rahman, F. Y., Purnomo, I. I., & Hijriana, N. (2022).

Penerapan Algoritma Data Mining Untuk Klasifikasi Kualitas Air. *Technologia : Jurnal Ilmiah*, *13*(3), 228. https://doi.org/10.31602/tji.v13i3.7070

Ramadhan, N. G. (2021). Comparative Analysis of ADASYN-SVM and SMOTE-SVM Methods on the Detection of Type 2 Diabetes Mellitus. *Scientific Journal of Informatics*, *8*(2), 276–282. https://doi.org/10.15294/sji.v8i2.32484

Safitri, D., Hilabi, S. S., & Nurapriani, F. (2023). Analisis Penggunaan Algoritma Klasifikasi Dalam Prediksi Kelulusan Menggunakan Orange Data Mining. *Rabit : Jurnal Teknologi Dan Sistem Informasi Univrab*, *8*(1), 75–81. https://doi.org/10.36341/rabit.v8i1.3009

Salma, A. (2023). *Analysis on Symptoms Driven Disease Risk Assessment using Artificial Intelligence Approach*. https://ieeexplore.ieee.org/abstract/document/10522221

Sepharni, A., Hendrawan, I. E., & Rozikin, C. (2022). Klasifikasi Penyakit Jantung dengan Menggunakan Algoritma C4.5. *STRING (Satuan Tulisan Riset Dan Inovasi Teknologi)*, *7*(2), 117. https://doi.org/10.30998/string.v7i2.12012

Smith, K., Fernie, S., & Pilcher, N. (2021). Aligning the times: Exploring the convergence of researchers, policy makers and research evidence in higher education policy making. *Research in Education*, *110*(1), 38–57. https://doi.org/10.1177/0034523720920677

Syahputra, S., Hasibuan, M. S., Komputer, J. I., Islam, U., Sumatera, N., Perjalanan, B., & Bayes, N. (2024). *Analisis Sentimen Jamaah Umrah Di Media Sosial X Menggunakan Algoritma Naive Bayes*. *19*(x), 107–116.