Journal of Theoretical and Applied Information Technology
31st July 2025. Vol.103. No.14

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AN INTELLIGENT SYSTEM FOR HAJJ CROWD MANAGEMENT USING DATA MINING TECHNIQUES

BASHAIR FAHAD[1] , ISLAM R. ABDELMAKSOUD [1], HAZEM EL-BAKRY [1]

[1] Department of Information Systems, Faculty of Computers and Information, Mansoura University, Egypt

E-mail: [1] wnus9@hotmail.com

ABSTRACT

Every year, millions of Muslims come to Mecca to participate in the Hajj pilgrimage. Due to the large number of attendees, there are a variety of logistical and safety concerns that must be resolved. Many times, modern crowd management techniques can be inefficient, resulting in tragic great loss of life events such as the 2015 stampede. In this paper, we explore the use of machine learning methods to predict crowd concentration during the Hajj and Umrah pilgrimages. For this purpose, we apply the Hajj and Umrah Crowd Management dataset available on Kaggle. Our aim is to classify remote sensing crowd density features into three classes, Low, Medium, and High based on certain conditions such as time of the year, weather, and health data. The dataset requires preprocessing such as rescaling, imputation of missing values, and encoding of categorical variables. Feature selection is performed using mutual information to eliminate irrelevant factors that do not aid in predicting crowd density. Hold-out and 5-Fold Cross-Validation techniques are used to train and assess five classification models: Random Forest, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Decision Trees, and Logistic Regression. The results show that Random Forest performs better than the other models. When feature selection is used, it attains maximum accuracy and F1-scores. The outcomes show how well machine learning predicts crowd density, and Random Forest turns out to be the most dependable model for handling sizable crowds during the Hajj and Umrah.

Keywords: *Hajj Crowd Management, Feature Selection, Data Mining Techniques, Mutual Information, Machine Learning in Crowd Management.*

1. INTRODUCTION

The annual Hajj pilgrimage is a principal religious practice for Muslims and one of the five pillars of Islam. It is a deeply religious experience that brings together millions of Muslims around the globe. Nonetheless, with the massive scale of the event also comes massive challenges, especially crowd management and safety and comfort for everyone involved [1]. With more than 2 million people congregated in a relatively small, defined geographic area, the risks of overcrowding, accidents, and organizational collapse are high. Controlling mass crowds at Hajj is also uniquely difficult due to many considerations. First, pilgrim movements are normally organized, especially during central rituals such as the Tawaf (process of circumambulation of the Kaaba) and the ritual of stoning the Jamarat. These coordinated movements are very susceptible to severe congestion at certain points. Second, the duration of the event is short with millions of persons attending and departing in a very short time that it is very resource-intensive and difficult to optimize. Third, human behavior and decision-making varies between individuals, especially regarding their age, health, culture and prior Hajj experience, which adds to this ever-changing dimension that is typically very hard to model with conventional approaches [2,3].

Multiple systems have been proposed for improved crowd management during Hajj. For example, Alazbah and Zafar proposed a model to reduce congestion at the path of stoning the Al-Jamarat. The proposed model used a dataset of 550 images. These images were split into 420 images for training and 130 for testing. The dataset was used to train and evaluate a convolutional neural network (CNN) model that classified the input images into three classes: crowded, semi-crowded, and normal [4]. Similarly, Albattah et al. [5] introduced a model that classifies crowd images into five categories: heavy crowd, crowded, semi-crowded, light crowd, and normal using a CNN. Each of these categories is associated with a color as an alarm to the severity of the situation. For

example, dark red represents heavy crowd while green represents normal. The basic idea of their model is to activate an alarm when the crowd exceeds a certain level to reduce the potential damage of reaching dangerous crowd levels. Felemban et al. [6] reviewed multiple technologies that can be utilized to improve the Hajj crowd management. These technologies included immersive technologies, wireless, data analytics, crowd modeling and simulation, computer vision, and mobile applications. The authors provided a comparison of these technologies to better understand their impacts in managing massive crowds. The use of flying adhoc network (FAN) that utilizes flying drones to provide multimedia data to the control centers responsible for hajj crowd management was explored in [7]. The proposed model employed priority-based routing framework for FAN to expedite image transmission from flying drones to base stations. The effectiveness of the proposed model over other conventional frameworks was evaluated using the Cooja simulation environment. Al-Shaery et al. [8] generated a multimodal dataset to improve Hajj crowd management. A mobile application was used to collect the data from 64 participants that performed different Hajj activities. Different models were proposed to identify the type of Hajj activity, emotional states, and level of fatigue. The best model resulted in accuracies of 41.71% for the type of Hajj activity, 82.47% for emotional states, and 85.27% for level of fatigue. The growing complexity of crowd movements demands more advanced techniques than common human surveillance and manual crowd management methods. By combining data mining and AI for Hajj management, not only is the overall experience of the pilgrims enriched, but the risk of accidents is also reduced and the effectiveness of crowd flow is increased. The ultimate objective is to ensure that the religious value of the event is not compromised by logistics or safety problems. In this study, several machine learning techniques are applied to improve crowd management during the Hajj and Umrah pilgrimages through crowd density level forecasting. A large dataset is utilized, and careful preprocessing, feature engineering, and model validation are conducted to provide valuable knowledge of the effectiveness of different machine learning models. In the course of the study, Random Forest has been identified as the most effective model in the prediction of crowd concentration. The main contributions of the proposed model are:

- Applying machine learning techniques to predict crowd intensity for Hajj and Umrah pilgrims from a dataset with 10,000 simulated entries across various variables like crowd intensity, weather conditions, health status, and pilgrim activities.
- Applying required data preprocessing such as handling missing values, converting categorical features, and scaling numerical features to have the dataset in a shape that was ready for analysis using efficient machine learning.
- Applying mutual information (MI)-based feature selection to find the key features to predict crowd density in order to improve model performance by removing the unnecessary features.
- Reporting that Random Forest had higher performance metrics than other evaluated models, and thus it is recommended to be used to predict pilgrim crowd density.
- Mitigating any model performance evaluation bias by evaluating the proposed model using 5-fold cross validation that produces more stable results compared to the hold-out technique
- Illustrating the capacity of machine learning in improving crowd safety and crowd management in major events such as Hajj.

2. METHODOLOGY

The proposed classification algorithm uses features of the Crowd Management Hajj and Umrah Dataset to predict levels of crowd density into three classes: Low, Medium, and High. The proposed method consists of the following three main steps: Data preprocessing, feature selection and classification. Data preprocessing involves handling missing data, encoding categorical attributes, and scaling numerical attributes. An MI-based feature selection technique is then applied to identify the most significant features. Finally, five different classifiers: Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Decision Trees, and Random Forest are trained and evaluated using both hold-out and 5-fold cross validation to identify the best model. Figure 1 illustrates these basic steps.

2.1 Data Collection

Kaggle offers an openly available dataset comprising 10,000 instances of artificial data across

various categories including technology interactions, health indicators, environment variables, pilgrim activities and crowd. With both continuous and categorical variables, the dataset is a valuable resource to learn about crowd behavior and develop more effective management methods [9].

2.2 Data Preprocessing

Data preparation is crucial for transforming unstructured data into a format that classification algorithms can use effectively. The essential steps generally include:

Handling missing data

One could also decide to discard any rows or columns that have an appreciable amount of missing values, but only if losing them won't invalidate the analysis entirely [10]. If temperature values have missing entries, then an imputation process may consist in computing and substituting the gaps with the mean temperature computed using the available data above.

Encoding categorical attributes

Categorical attributes, such as Crowd_Density, Activity_Type, Weather_Conditions, and Health_Condition, contain non-numeric data that must be encoded into numerical values before they can be used by machine learning algorithms [11]. Label Encoding: For binary categorical features (e.g., Emergency_Event, AR_System_Interaction), label encoding will be used to convert the categories into 0 and 1. Example: For Crowd_Density: 'Low' → 0, 'Medium' → 1, 'High' → 2.

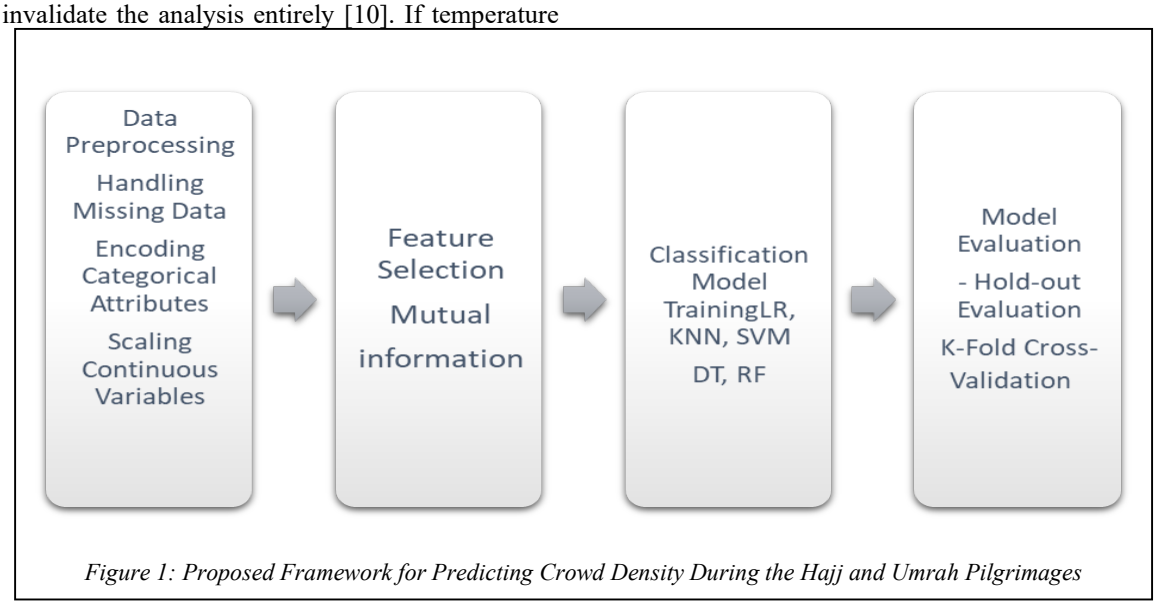

Figure 1: Proposed Framework for Predicting Crowd Density During the Hajj and Umrah Pilgrimages

Scaling continuous variables

Some continuous variables (such as Temperature, Movement_Speed, Queue_Time_minutes, etc.) may need to be scaled to ensure that the range of values does not impact the performance of algorithms sensitive to magnitude, such as KNN or Logistic Regression. Standardization (Z-score scaling): This will be used to scale continuous variables to have a mean of 0 and a standard deviation of 1. Example: Temperature: Before scaling, it may range from 30°C to 45°C. After scaling, it will be normalized to a mean of 0 and standard deviation of 1.

2.3 Feature Selection

Feature selection is critical to improve model performance by selecting only the most relevant features. In this case, we aim to select features that are most strongly correlated with the output label (Crowd_Density). In this work, a filter-based feature selection method, namely MI is used. Filter-based methods evaluate the importance of features based on their statistical relationship with the target variable. They are independent of any machine learning model. These methods are fast, computationally efficient, and interpretable. They work well with high-dimensional data and are independent of machine learning algorithms. They also reduce overfitting by eliminating irrelevant features [12, 13].

- **MI**

MI is a measure of dependence between two variables. It quantifies how much knowing one

variable reduces uncertainty about another. In the context of feature selection, MI helps determine how much information a feature provides about the target variable [14, 15]. It is defined as:

$$I(X,Y) = \sum_{x,y} p(x,y) log(\frac{p(x,y)}{p(x)p(y)})$$

where p(x,y) is the joint probability of feature X and target Y, p(x) is the probability of feature X, and p(y) is the probability of target Y.

If I(X,Y)=0, it means X and Y are independent, so the feature X does not help in predicting Y. Higher MI values indicate stronger relationships between the feature and the target.

MI has the following advantages:

- It works for both categorical and numerical features, making it highly versatile.
- It captures non-linear relationships.
- It has no assumptions about data distribution.

2.4 Classification Models Training

Once preprocessing and feature selection are complete, we proceed to train various classical machine learning classification algorithms. The algorithms will be evaluated using two techniques:

1- Hold-out evaluation (80-20 split): The dataset is split into a training set (80%) and a test set (20%). The model will be trained on the training set, and the performance will be evaluated on the test set.

2- 5-fold cross validation: Used to ensure robust model evaluation by splitting the dataset into five equal parts, where the model is trained on four folds and tested on the remaining one, repeating the process five times.

The following classification algorithms are evaluated for crowd density prediction:

1- Logistic regression – This is a linear model used for classification tasks, primarily for binary and multi-class problems. It estimates probabilities using the logistic function and applies a decision threshold [16, 17]. Despite being simple, it performs well for linearly separable data. Regularization techniques like L1 (Lasso) and L2 (Ridge) help prevent overfitting.

2- KNN – This is a non-parametric, instance-based algorithm that classifies a data point by considering the majority class of its nearest neighbors. The number of neighbors (k) is a crucial parameter affecting accuracy. It relies on distance metrics like Euclidean, Manhattan, or Minkowski. KNN is simple but computationally expensive for large datasets [18, 19].

3- SVM – This algorithm finds the optimal hyperplane that best separates data points into different classes. It maximizes the margin between classes and uses kernel functions (linear, polynomial, RBF) to handle non-linearly separable data. SVM is highly effective for high-dimensional spaces. However, it can be slow for large datasets [20, 21].

4- Decision Trees – A tree-based model that predicts outcomes by recursively splitting data based on feature values. 1. The model is very interpretable because each node represents a decision rule. Although it can be prone to overfitting, it handles both regression and classification applications. To increase performance and generalization, pruning techniques are applied [22, 23].

5- Random Forest – This ensemble learning technique combines several decision trees to improve resilience and accuracy. It prevents overfitting by taking averages of the decisions of many trees learned from random subsets of data [24, 25]. Random Forest is perhaps less understandable than one decision tree, but it works very well on missing values and high-dimensional data. Given the nature of the dataset, Tree-Based Methods, such as Random Forest, are particularly suitable to this classification method for a variety of reasons:

- **Dealing with mixed data types:** Both continuous (e.g., Movement_Speed, Temperature) and categorical variables (e.g., Activity_Type, Weather_Conditions) exist in the dataset. Tree-based models can deal with mixed data types without extra preprocessing.
- **Feature importance:** Tree-based models, particularly Random Forest, provide feature importance scores, which are useful to identify which features are most important in predicting Crowd_Density. This can directly be used for feature selection.
- **Robustness to overfitting:** Random Forest is less prone to overfitting than some other methods, especially in cases where there are many features in the data. Pruning techniques are employed to improve performance and extend the range of generalization.

3. EXPERIMENTAL RESULTS

3.1 Dataset Description

This dataset supports AI and AR-based research for managing Hajj and Umrah crowds in Mecca by analyzing pilgrim activity, environmental conditions, and security incidents. It includes 10,000 simulated records to aid in crowd prediction, safety management, and AR-based guidance.

3.2 Evaluation Metrics

Accuracy: The proportion of correct predictions out of the total predictions made.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Precision: The proportion of true positive predictions out of all positive predictions.

$$Precision = \frac{TP}{TP + FP}$$

Recall, or sensitivity: denotes the proportion of actual positive instances that the model correctly identifies. It is calculated using the formula:

$$Recall = \frac{TP}{TP + FN}$$

F1-Score: provides a balanced measure between precision and recall by taking their harmonic mean. This metric is particularly useful when there is an uneven distribution between classes or when both precision and recall hold equal importance. The formula for calculating the F1-Score is:

$$F1 - Score = \frac{2 \times (Precision \times Recall)}{Precision + Recall}$$

3.3 Results

1. Hold-out evaluation results:

For each model, we evaluate its performance on 20% of the dataset after training it on the remaining 80%. Below is a summary of each model's effectiveness, detailing its accuracy and classification report, both without (Table 1) and with the use of feature selection (Table 2). A visual representations of the results of these tables are shown in Figure 2 and Figure 3, respectively.

2. 5-fold cross-validation results:

Table 3 presents the precision, recall, F1-score, and accuracy metrics for each model, categorized by Low, Medium, and High crowd density classes. These findings offer a thorough assessment of the classifiers' performance across the various segments of the dataset. Figure 4 provides an illustration of the results in Table 3.

An analysis of how the classification models perform are detailed so that readers can determine what kind of classifiers are most appropriate for predicting crowd densities. This is of especial importance for crowd management at events as big as Hajj and Umrah. Random Forest does an astonishing job compared to other classification models even in situation without using feature selecting the. It was the best of all by far for both F1-scores, accuracy (ceiling at 73%), precision and recall. This demonstrates the astonishingly good feature-selection in that model that it is able to capture some complex patterns from the outliers where number of features revolves around a high number. F1-score was up in the second metric, greatly showcasing that our model's performance has increased in a lot predicting low crowd density. Results of the Hold-Out method and the 5-Fold Cross-Validation approach were very similar. However, 5-fold cross validation has an advantage of providing more stable results. This is because cross-validation helps mitigate the potential bias of a single hold-out test set. Random Forest followed by Logistic Regression and SVM resulted in the highest performance.

This study on Hajj crowd management has some limitations. The classification of crowd density into only three levels (Low, Medium, High) oversimplifies the problem and may lead to ignoring potential dangerous situations. Moreover, the feature selection algorithm based on MI may not capture non-linear dependencies and discard important features. The proposed models have to be combined with other technologies such as sensor and drones to be implemented in real-time. Additionally, some human behavioral factors such as emotional state are not taken into account in this research, which play a central role in effective crowd management.

4. CONCLUSION

This study examines the effectiveness of various machine learning classification algorithms in predicting crowd density during the Hajj and Umrah pilgrimages—events that draw millions of participants annually and demand precise crowd management to ensure safety and operational efficiency. The research follows a structured methodology that includes data preprocessing, feature selection using mutual information techniques, and rigorous training and evaluation of multiple machine learning models. Results indicate that the Random Forest algorithm performs better than other classifiers with consistency in major performance metrics. One of the major findings is

the degree to which feature selection methods impact model accuracy and reliability. This impact is especially significant in the Random Forest and Logistic Regression models, emphasizing the importance of feature engineering in machine learning workflows. By both 5-Fold Cross-Validation and Hold-out validation tests, Random Forest performed better in accuracy, precision, recall, and F1-scores, validating its power in predictive models. Future work might focus on exploring methods such as deep learning and transformer-based frameworks or incorporating additional real-time variables like environmental and behavioral variables to further enhance predictive capability

REFERENCES:

[1] Alasmari AM, Farooqi NS, Alotaibi YA. Recent trends in crowd management using deep learning techniques: a systematic literature review. J Umm Al-Qura Univ Eng Archit. 2024;1-29.

[2] Aldahawi HA. Big Data Analytics Strategy Framework: A Case of Crowd Management During the Hajj Pilgrimage Mecca Saudi Arabia. Biosci Biotechnol Res Commun. 2021;14(4):1975-84.

[3] Shah AA. Enhancing Hajj and Umrah Rituals and Crowd Management through AI Technologies: A Comprehensive Survey of Applications and Future Directions. IEEE Access. 2024;12:161820-41.

[4] Alazbah A, Zafar B. Pilgrimage (hajj) crowd management using agent-based method. Int J Found Comput Sci Technol. 2019;9(1):1-17.

[5] Albattah W, Khel MHK, Habib S, Islam M, Khan S, Abdul Kadir K. Hajj crowd management using CNN-based approach. Comput Mater Contin. 2020;66(2):2183-97.

[6] Felemban EA, Rehman FU, Biabani SAA, Ahmad A, Naseer A, Majid ARM, et al. Digital revolution for Hajj crowd management: A technology survey. IEEE Access. 2020;8:208583-609.

[7] Felemban E, Sheikh AA, Naseer A. Improving response time for crowd management in Hajj. Comput. 2021;10(4):46.

[8] Al-Shaery AM, Ahmed SG, Aljassmi H, Al-Hawsawi AN, Maksoud N, Tridane A, et al. Open dataset for predicting pilgrim activities for crowd management during Hajj using wearable sensors. IEEE Access. 2024;12:72828-46.

[9] Hajj and Umrah crowd management dataset. Kaggle. Available from: https://www.kaggle.com/datasets/ziya07/hajj-and-umrah-crowd-management-dataset/data

[10] Zhang Y, Thorburn PJ. Handling missing data in near real-time environmental monitoring: A system and a review of selected methods. Future Gener Comput Syst. 2022;128:63-72.

[11] Breskuvienė D, Dzemyda G. Categorical feature encoding techniques for improved classifier performance when dealing with imbalanced data of fraudulent transactions. Int J Comput Commun Control. 2023;18(3).

[12] Rakesh DK, Jana PK. A general framework for class label specific mutual information feature selection method. IEEE Trans Inf Theory. 2022;68(12):7996-8014.

[13] Zhou G, Li R, Shang Z, Li X, Jia L. Multi-label feature selection based on minimizing feature redundancy of mutual information. Neurocomputing. 2024;607:128392.

[14] Gong H, Li Y, Zhang J, Zhang B, Wang X. A new filter feature selection algorithm for classification task by ensembling Pearson correlation coefficient and mutual information. Eng Appl Artif Intell. 2024;131:107865.

[15] He J, Qu L, Wang P, Li Z. An oscillatory particle swarm optimization feature selection algorithm for hybrid data based on mutual information entropy. Appl Soft Comput. 2024;152:111261.

[16] DeMaris A, Selman SH. Logistic regression. In: Converting data into evidence: A statistics primer for the medical practitioner. 1st ed. 2013. p. 115-36.

[17] Charizanos G, Demirhan H, İçen D. A Monte Carlo fuzzy logistic regression framework against imbalance and separation. Inf Sci. 2024;655:119893.

[18] Abu Alfeilat HA, Hassanat AB, Lasassmeh O, Tarawneh AS, Alhasanat MB, Eyal Salman HS, Prasath VS. Effects of distance measure choice on k-nearest neighbor classifier performance: a review. Big Data. 2019;7(4):221-48.

[19] Xie J, Xiang X, Xia S, Jiang L, Wang G, Gao X. MGNR: A multi-granularity neighbor relationship and its application in KNN classification and clustering methods. IEEE Trans Pattern Anal Mach Intell. 2024.

[20] Sevakula RK, Verma NK. Support vector machine for large databases as classifier. In: Proceedings of the Third International Conference on Swarm, Evolutionary, and

Memetic Computing (SEMCCO 2012), Bhubaneswar, India, December 20-22, 2012. Springer; 2012. p. 303-13.
[21] Lai Z, Liang G, Zhou J, Kong H, Lu Y. A joint learning framework for optimal feature extraction and multi-class SVM. Inf Sci. 2024;671:120656.
[22] Priyanka, Kumar D. Decision tree classifier: a detailed survey. Int J Inf Decis Sci. 2020;12(3):246-69.
[23] Sun Z, Wang G, Li P, Wang H, Zhang M, Liang X. An improved random forest based on the classification accuracy and correlation measurement of decision trees. Expert Syst Appl. 2024;237:121549.
[24] Halabaku E, Bytyçi E. Overfitting in Machine Learning: A Comparative Analysis of Decision Trees and Random Forests. Intell Autom Soft Comput. 2024;39(6).
[25] Sun Z, Wang G, Li P, Wang H, Zhang M, Liang X. An improved random forest based on the classification accuracy and correlation measurement of decision trees. Expert Syst Appl. 2024;237:121549.

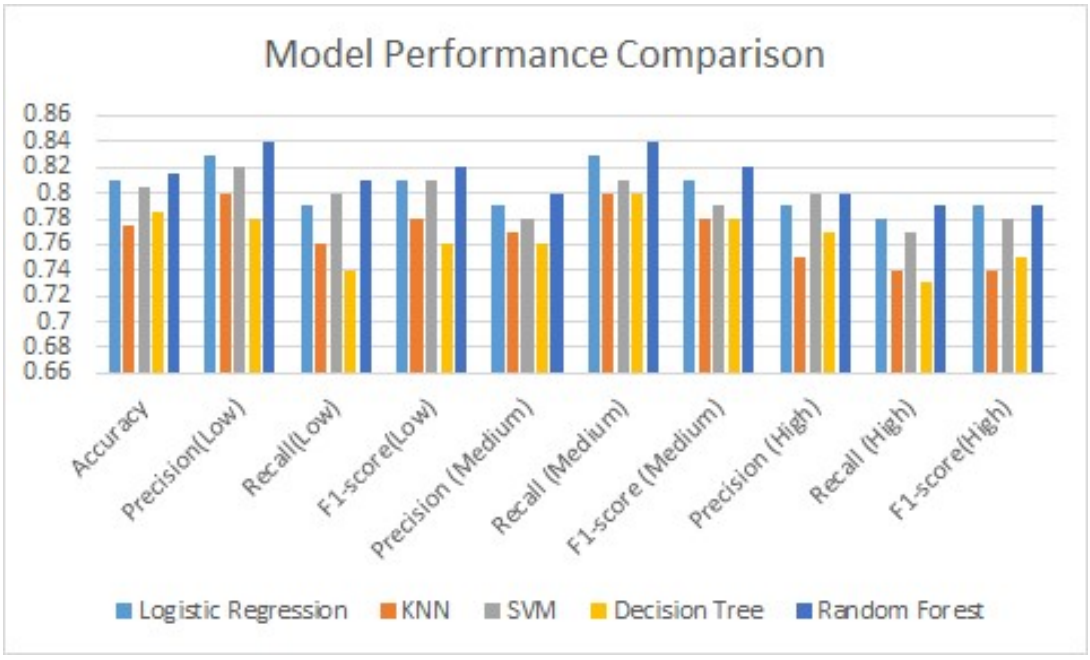

Figure 2: Hold-Out Evaluation Results for Five Different Models Without Feature Selection

Table 1: Hold-Out Results (Without Feature Selection).

Model	Accuracy	Precision(Low)	Recall(Low)	F1-score(Low)	Precision (Medium)	Recall (Medium)	F1-score (Medium)	Precision (High)	Recall (High)	F1-score (High)
Logistic Regression	0.81	0.83	0.79	0.81	0.79	0.83	0.81	0.79	0.78	0.79
KNN	0.78	0.80	0.76	0.78	0.77	0.80	0.78	0.75	0.74	0.74
SVM	0.81	0.82	0.80	0.81	0.78	0.81	0.79	0.80	0.77	0.78
Decision Trees	0.79	0.78	0.74	0.76	0.76	0.80	0.78	0.77	0.73	0.75
Random Forest	0.82	0.84	0.81	0.82	0.80	0.84	0.82	0.80	0.79	0.79

Table 2: Hold-Out Results (With Feature Selection).

Model	Accuracy	Precision(Low)	Recall(Low)	F1-score(Low)	Precision (Medium)	Recall (Medium)	F1-score (Medium)	Precision (High)	Recall (High)	F1-score (High)
Logistic Regression	0.83	0.86	0.82	0.84	0.81	0.85	0.83	0.80	0.82	0.81
KNN	0.80	0.82	0.79	0.80	0.79	0.82	0.80	0.77	0.75	0.76
SVM	0.82	0.85	0.83	0.84	0.80	0.84	0.82	0.81	0.79	0.80
Decision Trees	0.80	0.80	0.77	0.78	0.78	0.81	0.79	0.76	0.74	0.75
Random Forest	0.85	0.88	0.85	0.86	0.83	0.87	0.85	0.83	0.81	0.82

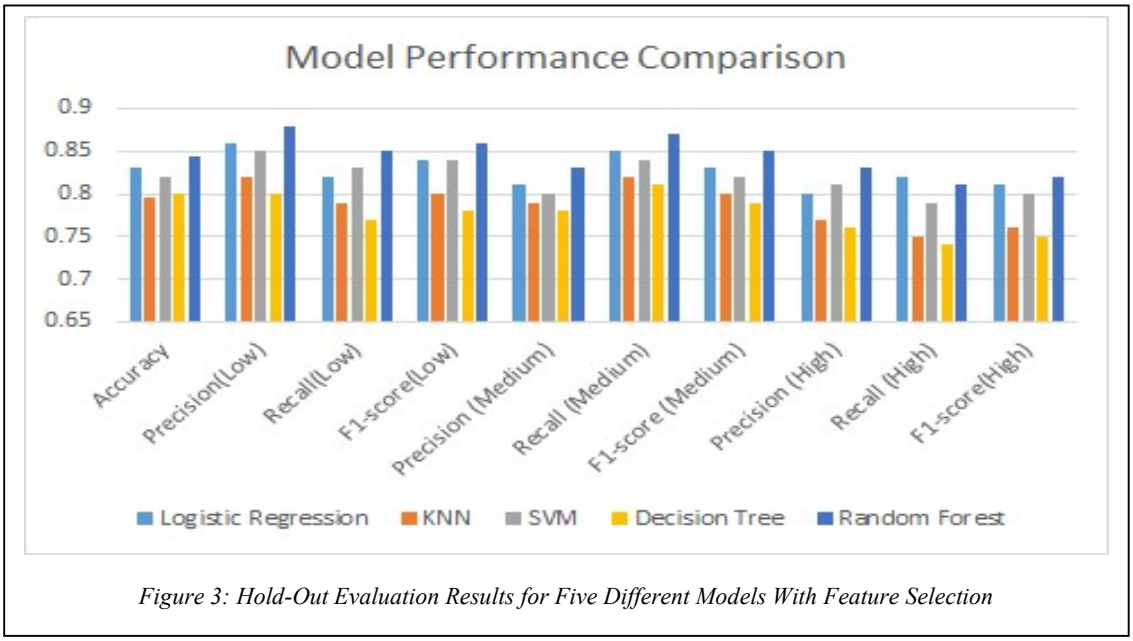

Figure 3: Hold-Out Evaluation Results for Five Different Models With Feature Selection

Table 3: 5-Fold Cross-Validation Results (Without Feature Selection).

Model	Accuracy	Precision (Low)	Recall(Low)	F1-score(Low)	Precision (Medium)	Recall (Medium)	F1-score (Medium)	Precision (High)	Recall (High)	F1-score(High)
Logistic Regression	0.81	0.84	0.79	0.80	0.79	0.80	0.80	0.81	0.79	0.80
KNN	0.79	0.79	0.75	0.77	0.75	0.76	0.74	0.77	0.75	0.75
SVM	0.81	0.82	0.80	0.79	0.78	0.80	0.78	0.80	0.79	0.79
Decision Trees	0.77	0.72	0.68	0.75	0.68	0.72	0.71	0.70	0.70	0.72
Random Forest	0.81	0.85	0.82	0.83	0.81	0.82	0.79	0.83	0.80	0.81

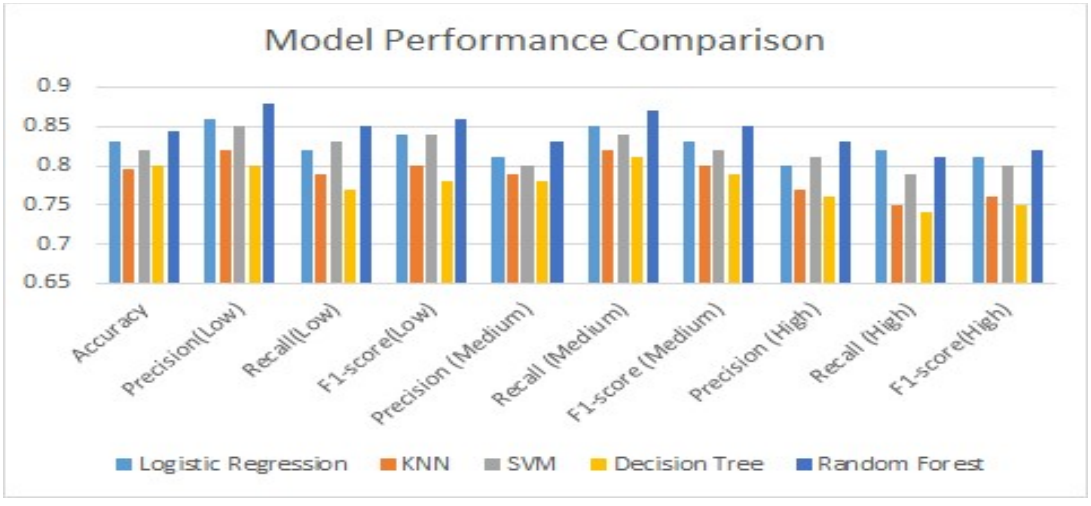

Figure 4: 5-Fold Cross Validation Evaluation Results for Five Different Models Without Feature Selection