Hindawi
Applied Computational Intelligence and Soft Computing
Volume 2023, Article ID 8063846, 13 pages
https://doi.org/10.1155/2023/8063846

Research Article

Machine Learning Approaches to Predict Patient's Length of Stay in Emergency Department

Mohammad A. Shbool,[1] **Omar S. Arabeyyat**,[2] **Ammar Al-Bazi**,[3] **Abeer Al-Hyari**,[4] **Arwa Salem**,[1] **Thana' Abu-Hmaid**,[1] **and Malak Ali**[1]

[1]*Industrial Engineering Department, School of Engineering, The University of Jordan, Amman 11942, Jordan*
[2]*Project Management Department, Faculty of Business, Al-Balqa Applied University, Al-Salt 19117, Jordan*
[3]*Aston Business School, Aston University, Birmingham B4 7ER, UK*
[4]*Computer Engineering Department, Faculty of Engineering, Al-Balqa Applied University, Al-Salt 19117, Jordan*

Correspondence should be addressed to Mohammad A. Shbool; m.shbool@ju.edu.jo

Received 30 March 2023; Revised 17 September 2023; Accepted 14 October 2023; Published 27 October 2023

Academic Editor: Aniello Minutolo

As the COVID-19 pandemic has afflicted the globe, health systems worldwide have also been significantly affected. This pandemic has impacted many sectors, including health in the Kingdom of Jordan. Crises that put heavy pressure on the health systems' shoulders include the emergency departments (ED), the most demanded hospital resources during normal conditions, and critical during crises. However, managing the health systems efficiently and achieving the best planning and allocation of their EDs' resources becomes crucial to improve their capabilities to accommodate the crisis's impact. Knowing critical factors affecting the patient length of stay prediction is critical to reducing the risks of prolonged waiting and clustering inside EDs. That is, by focusing on these factors and analyzing the effect of each. This research aims to determine the critical factors that predict the outcome: the length of stay, i.e., the predictor variables. Therefore, patients' length of stay in EDs across waiting time duration is categorized as (low, medium, and high) using supervised machine learning (ML) approaches. Unsupervised algorithms have been applied to classify the patient's length of stay in local EDs in the Kingdom of Jordan. The Arab Medical Centre Hospital is selected as a case study to justify the performance of the proposed ML model. Data that spans a time interval of 22 months, covering the period before and after COVID-19, is used to train the proposed feedforward network. The proposed model is compared with other ML approaches to justify its superiority. Also, comparative and correlation analyses are conducted on the considered attributes (inputs) to help classify the LOS and the patient's length of stay in the ED. The best algorithms to be used are the trees such as the decision stump, REB tree, and Random Forest and the multilayer perceptron (with batch sizes of 50 and 0.001 learning rate) for this specific problem. Results showed better performance in terms of accuracy and easiness of implementation.

1. Introduction

In healthcare systems, the emergency department (ED) plays a vital role as it provides emergency services to patients who report to this department during their stay. According to [1], length of stay (LOS) can be defined as the time interval from a patient's arrival to the ED until the patient leaves the ED (the total hospitalization time). The waiting time includes all times for triage, testing, obtaining test results, and waiting for the doctor and nursing assessment. The pandemic significantly affected the number of emergency cases for reasons like COVID-19 and other medical reasons. A mathematical model for estimating the probable outbreak size of COVID-19 clusters as a function of time was presented by [2]. This leads to exhausting hospital resources such as staff members, medical equipment, and beds [3], significantly affecting patients' waiting time to receive the required medical assistance. This will increase the risk to patients' lives due to the shortage of healthcare systems to handle the increasing number of patient cases. Risks of

infection include the lengthy waiting time and the clustering inside a closed environment, such as the ED. [4] studied the effect of nonpharmaceutical interventions and clustering on the number of infections inside the ED using agent-based simulation. Therefore, classifying and then predicting the right patient's length of stay would enable hospital officials to manage the resources of their departments more effectively.

This research aims to determine the essential factors, represented by predictor variables, influencing patients' LOS in the ED during the COVID-19 outbreak. Accurate prediction of ED LOS is crucial for several reasons. Firstly, ED LOS is a critical metric in healthcare as it directly impacts patient care and resource management. Excessive LOS can lead to delays in treatment, potentially compromising patient outcomes. It also affects the overall efficiency of the ED, as prolonged stays can lead to overcrowding and strain on resources. Secondly, the definition of excessive LOS may vary for different patients and conditions. Understanding what constitutes excessive LOS for specific cases is vital for timely and effective care.

Furthermore, prolonged LOS can have significant implications for patients and the department. For patients, it may result in increased discomfort, stress, and dissatisfaction with their healthcare experience. For the ED department, it can lead to decreased throughput, increased operational costs, and challenges in managing patient flow.

Currently, ED LOS is used as a critical performance indicator for ED management and resource allocation. Hospitals rely on this metric to assess their ability to meet patient demand and make informed decisions about staffing, bed availability, and resource distribution.

Therefore, this research aims to determine the critical factors that predict the outcome: the length of stay, i.e., the predictor variables. Therefore, patients' length of stay in EDs across waiting time durations will be categorized as (low, medium, and high) using supervised machine learning (ML) approaches. The purpose is to determine significant factors in predicting ED LOS accurately, enabling healthcare systems to address crucial factors contributing to prolonged LOS proactively and thus design interventions that could reduce LOS, enhance the overall patient experience, and optimize resource allocation based on those significant factors. By doing so, we contribute to more effective healthcare service delivery, particularly during pandemics such as the COVID-19 outbreak, when the demands on the ED are especially pronounced.

The rest of this paper is organized as follows: the first section will present related studies in which the knowledge gap covered in this work will be discussed. The primary AI framework model with details about the ML model architecture is described in the methodology section. Then, a case study based in Jordan is presented. After that, the results and discussion section includes a discussion of the study results and comparisons. Finally, conclusions and future work will be given.

2. Literature Review

This section reviews the applications of machine learning approaches in predicting patients' LOS in hospitals, especially in EDs, before and after the COVID-19 outbreak.

A work done by [5] focused on the effect of prolonged LOS in hospitals on poor functional outcomes and hospital-acquired infections. Thus, it is critical to focus on predicting and reducing LOS in hospitals, specifically in the ED. For example, [6] studied the impact of delirium on patients' LOS in the ICU and hospital. A prediction model based on a light gradient boosting machine for indoor patients was developed by [7]. A work by [8] addressed the idea that healthcare services might benefit from new technologies like artificial intelligence (AI), big data and machine learning, and the Internet of Things (IoT) to fight COVID-19 (coronavirus) and other pandemics. The authors in [9] highlighted how AI and other factors can be incorporated into a model to predict patients' length of stay. These improved information systems will facilitate hospital EDs' services and reduce the overcrowding of patients in these departments.

The authors in [10] applied AI algorithms and data mining tools, including logistic regression (LR), decision trees (DT), and gradient boosted machines (GBM), to predict hospital admissions with patient data collected from the ED. In order to reduce the hospital LOS, automated patient discharge predictions were presented and incorporated by [11], yielding over 12 hours reduction in the LOS of some units of the hospital. The authors in [12] built an artificial neural network to predict the length of stay and need for postacute care for coronary syndrome patients. The proposed ANN consists of four layers: an input layer, two hidden layers, and one output layer. Due to the effect of LOS on hospital resources and staffing, accurately predicting the LOS is an essential step for healthcare givers, insurance companies, and medical teams. The authors in [13] used general admission features to predict LOS accurately. Several ML models were used, which are neural networks (NN), classification trees (CT), tree bagger (TB), Random Forest (RF), fuzzy logic (FL), support vector machine (SVM), K-nearest neighbor (KNN), regression tree (RT), and Naive Bayes (NB). The model was able to obtain 90.04% accuracy using the CT model.

The authors in [14] investigated the feasibility of using artificial neural network ensembles to predict ED disposition for infants and toddlers with bronchiolitis and their length of stay. The authors in [15] adopted artificial neural networks and genetic algorithms to predict renal colic in EDs. Machine learning classification techniques were of high interest to researchers during the COVID-19 pandemic; for example, [16] implemented machine learning classifiers to classify the mortality of people with underlying health conditions. The authors in [17] aimed at forecasting patients' length of stay using artificial neural network (ANN) within the predictive input factors such as patient age, gender, mode of arrival, treatment unit, medical tests, and the needed inspection in the ED. This method can also provide insights to ED medical staff to decide the patient's length of stay. The authors in [18] applied an established Random Forest (RF) algorithm to rank variables according to the power of AI and machine learning over clinical scores in predicting inpatient mortality for ED sepsis patients. The authors in [19] examined the factors that might influence the ED and length of stay for old patients. Factors that affect LOS in the ICU were investigated by [20].

A study by [21] in a diverse urban hospital found that a machine learning model, gradient boosting, accurately predicted the length of stay in the ED for COVID-19 patients based on clinical factors, aiding resource planning and informing patients about expected waiting times. Another work performed by [22] analyses electronic health records (EHR) of COVID-19 patients to predict infection severity based on the length of stay, utilizing oversampled data and an artificial neural network (ANN) with optimized hyperparameters, ultimately selecting the model with the highest $F1$ score for evaluation and discussion. The authors in [23] developed and validated a prediction model using a decision tree algorithm to accurately predict patients with an ED LOS of more than 4 hours, identifying key risk factors such as waiting for specific consultations, providing valuable insights for health managers to implement targeted interventions, and suggesting the potential utility of real-time risk display at the point-of-care.

Although the above studies investigated how to estimate patients' length of stay in the EDs, the impact of the COVID-19 outbreak and other patients on the LOS of patients in the EDs has not been investigated yet. In addition, this work focuses on the critical factors affecting the LOS. Machine learning algorithms were used to address predictor variables crucial in determining and classifying the LOS of patients in the ED. The reason behind selecting such algorithms is attributed to the nature of different input variables (gender, insurance, triage level, etc.) and the unawareness of the type of relationships between these variables and the LOS. As the above literature presented, machine learning has proved to be efficient in solving such complexity inherited in this kind of problem.

3. Methodology

3.1. The Proposed Prediction-Classification Framework. This section presents the prediction model development framework. A conceptual overview is given in Figure 1. The first step is to determine the input attributes and collect-related data. Details of the input attributes, definitions, and types of each attribute are summarized in Table 1.

Figure 1 shows that unsupervised, followed by supervised algorithms, were applied. The unsupervised algorithm's purpose was to cluster LOS times into range categories, followed by implementing the supervised algorithm after the categories had been generated to predict the correct range category. The supervised part of the data (input and output) was used to learn the pattern and classify the LOS.

3.2. Data Acquisition and Analysis. The LOS of patients at the ED represents the total time a patient spends in the ED before leaving home or being admitted to further healthcare services inside other hospital departments. The ED process starts with patients' arrival and ends with their departure. The patient might need to go through several activities, each consuming a specific amount of time reflecting their entire LOS at the ED. The LOS can be schematically depicted, as shown in Figure 2.

Figure 2 shows that the time spent in the ED starts with the patient's arrival, either by ambulance or as an ambulatory case. Then, the patient must be checked in at the reception by providing information, including the mode of arrival, date, day, gender, insurance, and age. After check-in, medical care starts with immediate treatments for urgent cases. Depending on case urgency, the triage level is determined to assess the next level of needed care. All required tests and imaging are then decided by the medical staff members, which go in parallel with the medication. The final step is the consultation before leaving the ED. The workload (staff) is assumed to be constant.

3.2.1. Data Collection. The dataset used in this study was collected from hospital's records. The data covers two years, from 2019 to 2020. A sample of data for the busiest days during the month was collected. These days are 1, 2, 9, 12, 13, 15, 18, 22, 23, and 28 of each month from January 2019 until October 2020. The final dataset contains a total number of 400 randomly selected patients' records. Patient privacy is critical, so we consented to collect raw data without patient identification information. Patients were not interviewed or asked about this data; the research team reviewed historical records from the hospital database under the supervision of the records responsible, with patient identification information masked. The hospital management granted the research team access to the data with consent to use the anonymous records solely for research purposes. Data were collected from emergency forms with categorical and numerical types for input in the ML model. Forty-two attribute data points were collected for the randomly selected 400 patients. Table 1 shows dataset definitions and details. The last attribute (LOS) is the response/output we want to estimate the LOS.

The ED process starts with patients' arrival and ends with their departure. After the patient has arrived, the check-in data needs to be undertaken. In this process, the receptionist will give the patient an ID number and record the date, day, arrival time, gender, insurance information, and patient age. Immediately after check-in, treatment will occur, starting with a nurse assessing the patient's case urgency level to put him in the right triage level. Then, the patient will be cared for by a physician to start the medication process, be prescribed all required tests to be correctly diagnosed, and be given the proper medication and consultation. When the medication process ends, the patient leaves the ED or is admitted to the hospital; thus, the LOS is calculated at this point. Data and inputs handled in this research are shown in detail with definitions in Table 1.

3.3. Data Preprocessing and Transformation. Real-world data is often incomplete, inconsistent, or lacking in specific ways and is likely to contain many errors, and here comes the researcher's role in resolving these issues. Data preprocessing is an essential step in machine learning. This process ensures that the data will be in a format the model understands to obtain the output. Data preprocessing is a data mining technique that involves cleaning and

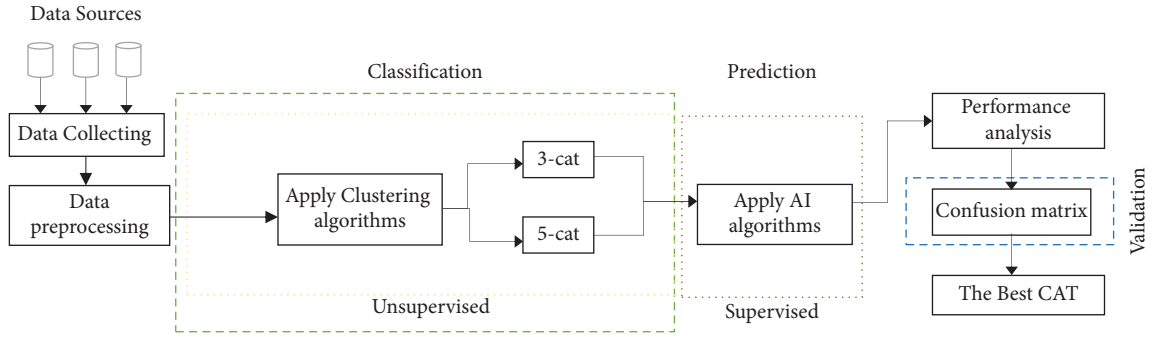

Figure 1: Prediction-classification framework.

transforming raw data into an acceptable form. Data preprocessing includes cleaning, instance selection, normalization, transformation, feature extraction, categorical values, sampling, etc. [24]. Cleaned data were divided into training and testing sets.

Data preprocessing is an essential step in machine learning. The phrase "garbage in, garbage out" is particularly suitable for data mining and machine learning projects; it emphasizes data preprocessing. Real-world data is often incomplete, inconsistent, or lacking in specific ways and is likely to contain many errors, and here comes the researcher's role in resolving these issues.

Data samples from the raw data considered outliers were removed, including those who died in the ED, less than 1-year-old infants (because they have different procedures), inpatients who left without being seen, and incomplete records. Also, qualitative attributes were labeled into quantitative data. Tables 2 and 3 in the following show some descriptive measures of the numerical and categorical variables, respectively.

The LOS is, on average, 68.1 minutes with a standard deviation of 49.6 minutes (see Table 2); this shows a significant number of cases that take more than 100 minutes. The time is considered high for two main reasons. First, patients visiting the ED are, in most cases, in need of immediate service, even if the case is not life-threatening. Second, in pandemics like COVID-19, high waiting time means a large queue, and as it is already well established, crowding is the primary factor for virus transmission and, thus, infection [4].

From a simple management perspective, data in Table 3 can be divided into two main categories: controlled and uncontrolled. The controlled variables are those we can decide in advance, while the others are those collected and found based on the decision of the controlled. We tried to distribute the controlled data uniformly. The output is given based on the two numbers of categorization tested; this will be discussed later.

In this research, a LOS prediction model was developed to determine the appropriate LOS time range using unsupervised machine learning techniques. Specifically, the data was clustered into five categories using the EM (Expectation-Maximization) algorithm implemented in Weka. The EM algorithm applied unsupervised clustering to group the data based on similarities or patterns. The resulting five categories were defined as follows: Category 1 represented LOS times ranging from 0 to 60 minutes, Category 2 encompassed LOS times from 61 to 120 minutes, Category 3 covered LOS times from 121 to 180 minutes, Category 4 included LOS times from 181 to 240 minutes, and Category 5 spanned LOS times from 241 to 300 minutes.

By leveraging the power of unsupervised machine learning, this LOS prediction model enabled the accurate classification of data points into the appropriate time ranges. Such an approach provides valuable insights into LOS patterns and facilitates decision-making in various domains, allowing for more effective resource allocation and patient management.

3.4. Attribute Correlation Analysis. Features selection, also identified as variable selection, attribute selection, or variable subset selection, is the process of choosing a subset of relevant features (variables and predictors) for use in model building. We used the Correlation Attribute Evaluation to assess the worth of an attribute by measuring the correlation (Pearson's) between it and the class. Nominal attributes are considered on a value-by-value basis by treating each value as an indicator.

3.5. Classification. Artificial intelligence (AI) is the intelligence demonstrated by machines. We used machine learning (ML), a branch of artificial intelligence that enables a model to learn from past data or experiences without being explicitly programmed. Machine learning uses a massive amount of structured and semistructured data, so a machine learning model can generate accurate results or give predictions based on that data. It can be divided into three types: supervised learning, reinforcement learning, and unsupervised learning. We use supervised learning, the machine learning task of learning a function that maps an input to an output based on previous cases (input-output pairs).

Unsupervised learning is a type of machine learning that looks for previously undetected patterns in a dataset with no preexisting labels and with a minimum of human supervision. Two main methods used in unsupervised learning are principal component and cluster analysis. Cluster analysis is used in unsupervised learning to group or segment datasets with shared attributes to extrapolate algorithmic relationships. Cluster analysis is a branch of machine learning that groups the data that has not been labeled, classified, or categorized [25]. In our project, we use clustering analysis.

Table 1: Input data set attributes, types, and definitions.

Category	Attribute	Definition
Check-in data	Date	Day, month, and year of arrival
	Day	The name of the day (Sunday, Monday … etc.)
	ID	Identity document of the patient in the hospital
	Gender	Male\female
	Insurance	Insurance info
	Mode of arrival	Patient's arrival mode
	Age	Age of the patient
Medical procedure	Immediate treatment	Immediate treatment requirements
	Triage level	Urgency case level (1–5)
	Medication	Medication needed (yes, no)
	Consultation	Consultation needed (yes, no)
Time	T arrive	The arrival time
	T triage assessment	Triage assessment time
	T NURS assessment	Nurse assessment time
	T doctor assessment	Doctor assessment time
	T departure	Patient's departure time
Medical tests	Twenty-three tests, including urine analysis, CBC, cardiac enzymes, stool analysis, X-ray, ultrasound, CT scan, and MRI	Tests
Others	Number of nurses	Available number of nurses
	Crowding	Number of patients in the ED
	Lockdown	Lockdown status
	LOS	(T departure-T arrival)

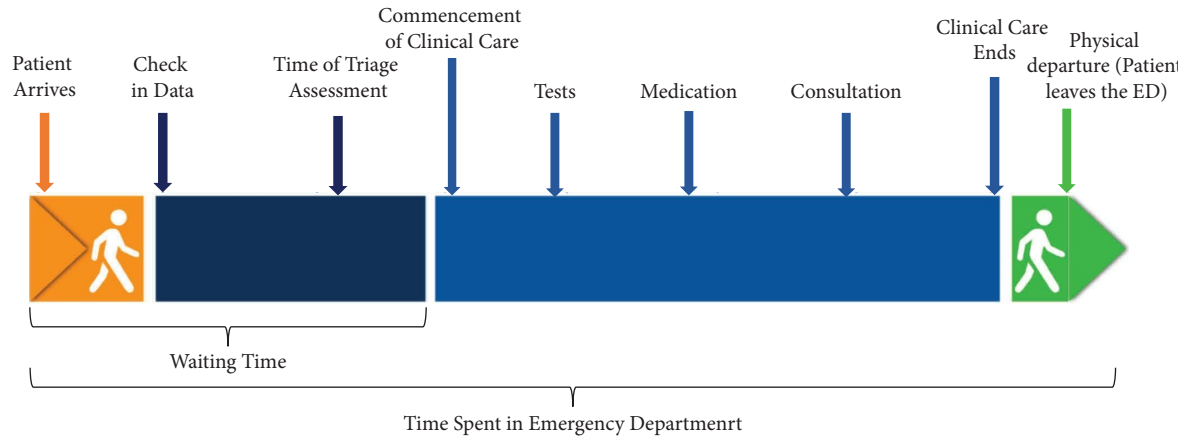

Figure 2: Length of stay breakdown in the emergency department.

Table 2: Descriptive statistical results of numeric variables.

Attribute	Type	Details	Min	Max	Mean	SD
Age	Input	The age of the patient	1	93	32.9	19.2
Nurses	Input	The number of nurses on duty upon patients' arrival	4	10	7.3	1.5
Crowding	Input	Number of patients in the ER at the same hour	1	33	10.1	5.9
LOS	Output	(T arrive-T departure) in minutes	8	294	68.1	49.6

Classification is done in this research using an unsupervised procedure (clustering analysis). This involves grouping data into categories based on inherent similarity or a distance measure. Unsupervised learning allows the system maximum flexibility in creating its own classification rules and hopefully finding hidden patterns unknown to humans (Ethem Alpaydin, 2014). Our work mainly uses clustering analysis to determine the number of categories. Implementation of Expectation Maximization Clustering EM assigns a probability distribution to each instance, indicating the probability of it belonging to each cluster. EM can decide how many clusters to create by cross-validation, or we may specify how many clusters to generate (Frank et al., 2017).

The next step is to implement one of the supervised learning algorithms to predict the right LOS category in terms of the attributes mentioned earlier. Finally, the performance evaluation and validation of the model are illustrated. The details of this step will be given in the results and discussion section. But first, let us provide some explanation of the main algorithms used as follows:

(i) Logistic Regression (logistic function):

It is a classification algorithm, used when the target variable's value is categorical. Logistic regression is a supervised classification algorithm. In a classification problem, the target variable (or output), y, can take only discrete values for a given set of features (or inputs), X, using the sigmoid function:

$$g(z) = \frac{1}{1 + e^{-z}}. \quad (1)$$

(ii) Naïve Bayes:

It is a classification algorithm for binary (two-class) and multiclass classification problems. The technique is easiest to understand when described using binary or categorical input values. In machine learning, we are often interested in selecting the best hypothesis (h) given data (d). In a classification problem, our hypothesis (h) may be the class to assign for a new data instance (d).

(iii) Random Forest:

It is a machine-learning classifier based on choosing random subsets of variables for each tree and using the most frequent tree output as the overall classification. It consists of many individual decision trees that operate as an ensemble. As we mentioned, each tree in the random forest spits out a class prediction, and the class with the most votes becomes our model's prediction.

(iv) Decision Stump:

A decision stump is a machine-learning model consisting of a one-level decision tree. It is a decision tree with one internal node (the root) immediately connected to the terminal nodes (its leaves). A decision stump makes a prediction based on the value of just a single input feature. Sometimes, they are also called 1-rules. Decision stumps are often used as components (called "weak learners" or "base learners") in machine learning ensemble techniques such as bagging and boosting.

Table 3: Descriptive statistical results of categorical variables.

Categorical attribute	Attribute type	Details	Occurrence (%)	Number of records
Day of the month	Input	1, 2, 9, 12, 13, 15, 18, 22, 23, 28	10.0 each	40 each
Month	Input	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	10.0 each	40 each
Day	Input	1: Friday	13.0	52
		2: Saturday	15.0	60
		3: Sunday	13.5	54
		4: Monday	13.5	54
		5: Tuesday	15.0	60
		6: Wednesday	15.5	62
		7: Thursday	14.5	58
Year	Input	2019	50.0	200
		2020	50.0	200
Gender	Input	1: female	46.5	186
		2: male	53.5	214
Insurance	Input	1: insured	74.0	296
		0: cash	26.0	104
Mode of arrival	Input	1: ambulatory	53.5	214
		2: ambulance	46.5	186
Immediate treatment	Input	0: not taken	99.5	398
		1: taken	0.5	2
Triage level	Input	1: level 1	0.5	2
		2: level 2	2.0	8
		3: level 3	20.5	82
		4: level 4	57.3	229
		5: level 5	19.7	79
Medication	Input	1: taken	54.0	216
		0: not taken	46.0	184
Consultation	Input	1: taken	75.0	300
		0: not taken	25.0	100
CBC	Input	1: taken	25.7	103
		0: not taken	74.3	297
KFT	Input	1: taken	17.5	70
		0: not taken	82.5	330
LFT	Input	1: taken	4.7	19
		0: not taken	95.3	381
Cardiac enzymes	Input	1: taken	6.8	27
		0: not taken	93.2	373
RBS	Input	1: taken	3.0	12
		0: not taken	97.0	388
CRB	Input	1: taken	4.5	18
		0: not taken	95.5	382
Amylase	Input	1: taken	1.0	4
		0: not taken	99.0	396
Lipase	Input	1: taken	2.3	9
		0: not taken	97.7	391
Urine analysis	Input	1: taken	6.0	24
		0: not taken	94.0	376
Stool analysis	Input	1: taken	1.3	5
		0: not taken	98.7	395
ABGS	Input	1: taken	0.5	2
		0: not taken	99.5	398
PT, INR	Input	1: taken	1.8	7
		0: not taken	98.2	393
PTT	Input	1: taken	0.3	1
		0: not taken	99.7	399

Table 3: Continued.

Categorical attribute	Attribute type	Details	Occurrence (%)	Number of records
Urine culture	Input	1: taken	0.3	1
		0: not taken	99.7	399
Urea	Input	1: taken	0.5	2
		0: not taken	99.5	398
Creatinine	Input	1: taken	1.0	4
		0: not taken	99.0	396
Troponin	Input	1: taken	1.0	4
		0: not taken	99.0	396
Xray	Input	1: taken	15.8	63
		0: not taken	84.2	337
Ultrasound	Input	1: taken	1.8	7
		0: not taken	98.2	393
CT scan	Input	1: taken	2.8	11
		0: not taken	97.2	389
MRI	Input	1: taken	0.8	3
		0: not taken	99.2	397
Others	Input	1: taken	23.3	93
		0: not taken	76.7	307
BMP	Input	1: taken	2.8	11
		0: not taken	97.2	389
Lockdown	Input	0: no lockdown	63.0	252
		1: lockdown	6.0	24
		2: partial ban from 6 pm to 10 am, walking on foot	1.5	6
		3: part-time and work permits granted	4.0	16
		4: odd, even cars from 8 am to 7 pm	3.5	14
		5: partial ban 11 pm–6 am	0.5	2
		6: partial ban 1 pm–6 am	2.5	10
		7: partial ban 12 pm–6 am	19.0	76
CAT	Output	*3 categories:		
		1: 0–100	83.0	332
		2: 101–200	14.3	57
		3: 201–300	2.7	11
		*5 categories:		
		1: 0–60	53.0	212
		2: 61–120	34.5	138
		3: 121–180	8.0	32
		4: 181–240	2.8	11
		5: 241–300	1.7	7

3.6. Model Evaluation. Evaluation of classification models can be performed using multiple metrics [26], including (not ordered in terms of importance): (1) confusion matrix (2) Accuracy, Recall, and Precision (3) *F*1 Score, and (4) log loss.

The confusion matrix (primary evaluation method used in this research), also known as an error matrix, is a specific table layout that allows visualization of the performance of an algorithm, typically for supervised learning. Each row of the matrix represents the instances in a predicted class, while each column represents the instances in an actual class. It uses two-word terminology. The first word indicates the correctness of the decision, while the second indicates the prediction result. Regarding the model validation, the general layout for a three-category model can be summarized in Table 4. This means that the first diagonal elements are desired. For example, True A means that the model correctly classified A as A, while False A means that the model classified B or C as A, which is incorrect output.

3.6.1. Five Categories vs. Three Categories. The 5-CAT results were unsatisfactory; the best accuracy is 65.75% for the Naïve Bayes algorithm. Thus, a 3-CAT classification was suggested by the researchers after trying a few other category scenarios. The time intervals are divided into three categories labeled by the numbers 1, 2, and 3, in which each number represents a category as follows (1: 0–100, 2: 101–200, 3: 201–300 minutes). It represents the general human-used classification: low, medium, and high.

3.7. Validation. The most critical indicator to consider in machine learning is cross-validation, a resampling procedure used to evaluate machine learning models on limited data samples. The procedure has a single parameter called *k*, which refers to the number of groups a given data sample is split into. In a prediction problem, a model is usually given a dataset of known data on which training is run (the

Table 4: Confusion matrix layout.

		Actual		
		A	B	C
Predicted	Predicted A	True A	False A	False A
	Predicted B	False B	True B	False B
	Predicted C	False C	False C	True C

training dataset) and a dataset of unknown data (or first-seen data) against which the model is tested (called the validation dataset or testing set). Cross-validation aims to test the model's ability to predict new data not used in the model development. In this model, 90% of the data were used in training the model, which comes from the pre-COVID dataset, while the rest (10%) of the data were used for testing the model, which includes COVID-19 data in addition to part of the pre-COVID dataset. This is because the main aim is to investigate the critical factors in predicting LOS, and COVID-19 is a temporary issue that is not considered a fundamental element in the model.

In k-fold cross-validation, the original sample is randomly partitioned into k equal-sized subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining "k−1" subsamples are used as training data. The cross-validation process is then repeated k times, with each k subsample used exactly once as the validation data. The k results can then be averaged to produce a single estimation. The advantage of this method over repeated random subsampling is that all observations are used for training and validation, and each observation is used for validation exactly once. 10-fold cross-validation is commonly used, and we use it in our models, but in general, k remains an unfixed parameter [27]. Figure 3 shows an explanation of the concept using k of 5.

4. Case Study

In order to justify the developed ML-based classification model established to predict LOS classes before and after the COVID-19 pandemic, a case study was selected and run. This case study was carried out in one of the local hospitals in Jordan. Established in 1994, it positioned itself among the top medical destinations and leading referral hospitals for local, regional, and international patients. The hospital strives to provide high-quality healthcare to all of its patients.

This hospital includes 14 specialized medical units consisting of 145 inpatient beds and 89 covering emergency, resuscitation, operations, newborns, and other outpatient services (with a total capacity of 234 beds). It offers a full range of medical and surgical services, covering all specialties. Their ED contains rooms dedicated to pediatric cases and for those with infectious diseases. The department treats an average of 200 patients per day. There are doctors on call, covering all subspecialties 24/7.

This section includes data collection, preparation, and descriptive statistical analysis and presents datasets and definitions.

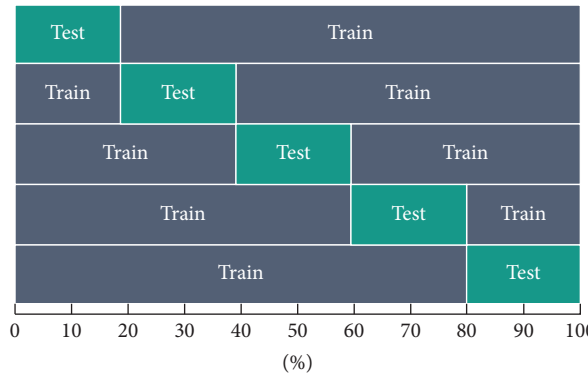

Figure 3: Five-fold (5-fold) cross-validation [27].

5. Results and Discussion

The LOS prediction model was built in this research to predict which of the LOS time ranges unsupervised algorithms have classified correctly. Time using the unsupervised clustering algorithm in Weka (EM algorithm), which clustered the data into five categories (1 ⟶ 0–60, 2 ⟶ 61–120, 3 ⟶ 121–180, 4 ⟶ 181–240, and 5 ⟶ 241–300 minutes).

We used the top-ranked classification algorithms mentioned in the method to differentiate between the three-categories model and the five-categories model using 400 patient records.

Regarding the attribute correlation analysis, attributes with high correlation significantly affect the LOS of the patients. Those attributes with a high correlation will be given more weight in the model. The ED management should focus more on these variables when reducing the LOS. It becomes more important when fighting against spread of viruses, such as the recent COVID-19 pandemic. Table 5 summarizes the correlation analysis.

One of the main factors in reducing LOS is increasing the staff and equipment assigned to the most requested tests by the doctors. As a result, each patient's time spent in the ED is decreased.

Regarding the model evaluation, Table 6 shows the output confusion matrix for the 5-CAT classification with results for all algorithms implemented in this work for comparison purposes. For example, the logistic algorithm classified 3 data points (LOS of 3 patients) as "a," and they are actually in category "a." On the other hand, it classified 15 as "a," and they are actually "b." The actual number of data points in class "a" = 3 + 16 + 7 + 2 + 0 = 28, correctly predicted 3. In other words, the diagonal of the confusion matrix represents the correctly predicted classes. For the logistic algorithm, 3 + 56 + 168 + 4 + 3 = 234 out of the 400 records were correctly predicted, resulting in a percentage of 234/400 = 58.5%. More discussion will be given later, when the 3-CAT is introduced.

The table in the following (Table 7) compares these categories for all algorithms for the five-categories vs. three-categories analysis.

Table 5: Correlation coefficients of the attributes.

Attribute	Correlation coefficient	Correlation
Day	0.03902	Low
Month	0.04942	Low
Year	0.01453	Low
Day	0.06091	Low
Gender	0.02381	Low
Insurance	0.03442	Low
Mode of arrival	0.01844	Low
Immediate treatment	0.06235	Low
Triage level	0.06123	Low
Medication	0.17187	High
Consultation	0.08824	Low
CBC	0.37436	High
KFT	0.30211	High
LFT	0.29638	High
Cardiac enzymes	0.1144	Low
RBS	0.07796	Low
CRB	0.24626	High
Amylase	0.15416	High
Lipase	0.19014	High
Urine analysis	0.1898	High
Stool analysis	0.01062	Low
ABGS	0.06144	Low
PT, INR	0.18187	High
PTT	0.10291	Low
Urine culture	0.10954	Low
Urea	0.15511	High
Creatinine	0.15416	High
Troponin 1	0.04406	Low
Xray	0.13779	High
Ultrasound	0.13197	High
CT scan	0.04266	Low
MRI	0.11438	Low
Other tests	0.18535	High
BMP	0.34567	High
Age	0.08461	Low
Number of nurses	0.00505	Low
Crowding	0.0713	Low
Lockdown	0.0645	Low

Table 7 shows correctly classified measures for all algorithms in both classification schemes. In addition, the REP tree algorithm resulted in the best performance, with an accuracy of 86.3%, followed by the decision stump with 85.8% accuracy. The main reason the 3-CAT is better than the 5-CAT is the effect of widening the scoring scale in decision problems in general. It becomes more difficult to distinguish between categories when their number increases. Thus, accuracy will increase for fewer categories, especially with a small sample size (400 is considered small in these models). The tradeoff between accuracy and informative classification is the primary criterion for selecting three categories and not two.

5.1. Before and after COVID-19. In this section, we compare the model's performance before and after COVID-19 are necessary. The before and after results of the spread of COVID-19 in Jordan are summarized in Figure 4. The pandemic reduced the model quality. This is mainly due to unusual situations that cause interruptions in healthcare services.

Table 6: Confusion matrix for 5-CAT for each algorithm.

Algorithm	Confusion matrix				
Logistic	a	b	c	d	e
	3	15	11	3	0
	16	56	60	6	0
	7	35	168	2	0
	2	2	2	4	1
	0	0	1	3	3
Naive Bayes	a	b	c	d	e
	2	16	11	2	1
	7	68	63	0	0
	1	24	187	0	0
	3	4	0	3	1
	0	1	2	3	1
SMO	a	b	c	d	e
	4	15	12	1	0
	10	62	65	1	0
	1	24	187	0	0
	3	4	0	1	3
	0	2	1	1	3
lazy.IBk	a	b	c	d	e
	0	17	15	0	0
	0	59	79	0	0
	0	13	199	0	0
	0	10	1	0	0
	0	4	3	0	0
Decision stump	a	b	c	d	e
	0	17	15	0	0
	0	59	79	0	0
	0	13	199	0	0
	0	10	1	0	0
	0	4	3	0	0
REP tree	a	b	c	d	e
	0	17	15	0	0
	0	61	77	0	0
	0	21	191	0	0
	0	10	1	0	0
	0	4	3	0	0
Random Forest	a	b	c	d	e
	0	19	13	0	0
	3	57	78	0	0
	0	43	169	0	0
	0	9	2	0	0
	0	4	3	0	0
MLP	a	b	c	d	e
	0	0	32	0	0
	0	0	138	0	0
	0	0	212	0	0
	0	0	11	0	0
	0	0	7	0	0

Figure 4 shows that all algorithms performed better for the data available before the COVID-19 pandemic. The reduced accuracy of the LOS prediction model after the COVID-19 spread can be attributed to several factors. One significant factor is the limited data availability during the pandemic compared to the prepandemic period. Most of the data used for training and testing the model was collected before the outbreak. Consequently, the model's performance may have been adversely affected as it was not explicitly trained on postpandemic patterns.

Table 7: Comparison between the 3-CAT and 5-CAT for all algorithms.

Classifier	3-CAT				5-CAT			
	Correctly classified		Incorrectly classified		Correctly classified		Incorrectly classified	
Logistic regression	322	80.5%	78	19.5%	242	60.5%	158	39.5%
Naive Bayes	327	81.8%	73	18.2%	263	65.8%	137	34.2%
REP tree	345	86.3%	55	13.7%	252	63.0%	148	37.0%
SMO	333	83.3%	67	16.7%	254	63.5%	146	36.5%
lazy.IBk	310	77.5%	90	22.5%	211	52.8%	189	47.2%
Decision stump	343	85.8%	57	14.2%	258	64.5%	142	35.5%
Random Forest	330	82.5%	70	17.5%	235	58.8%	165	41.3%
MLP (0.0001)	332	83.0%	68	17.0%	212	53.0%	188	47.0%
MLP (0.001)	332	83.0%	68	17.0%	254	63.5%	146	36.5%
MLP (0.01)	317	79.3%	83	20.7%	240	60.0%	160	40.0%

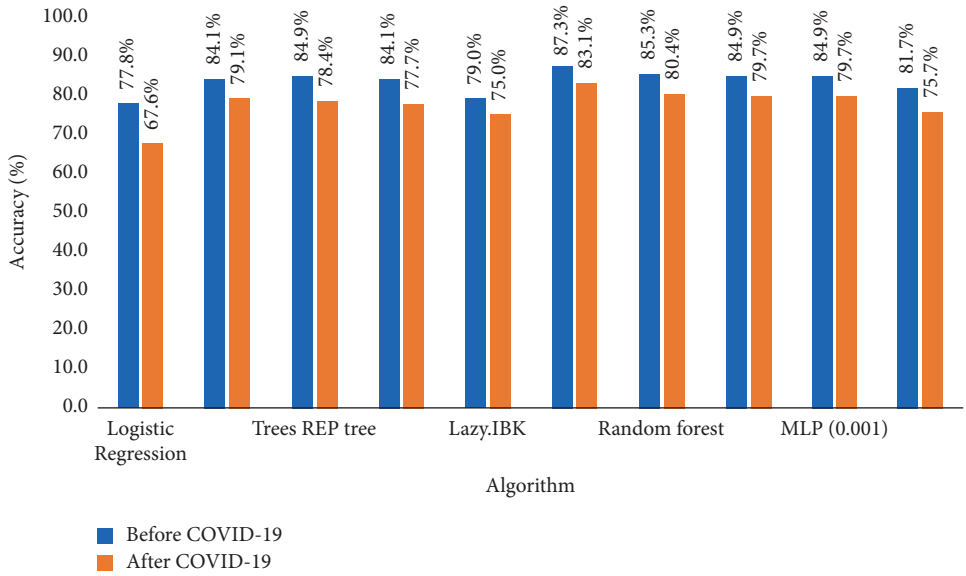

Figure 4: Correctly classified % of LOS for each algorithm before and after COVID-19.

Moreover, there was a notable decline in the number of visits to the Emergency Department (ED) following the implementation of lockdown measures and mobility restrictions during the pandemic. This decrease in patient volume resulted in a shift in the types and distribution of medical cases encountered in the ED. The reduced likelihood of accidents and infections due to restricted movements further influenced the accuracy of the model's predictions.

This work has some limitations, which might hinder the accuracy of the results if they have not been tackled in the future. Among these limitations is the limited data availability during the pandemic compared to the prepandemic period. This might denigrate the model's performance if it is not well trained on postpandemic patterns. It is worth mentioning that this study is a single-site study, which reflects the results of a special case for a single-site study rather than a general study with more than one hospital. Also, a small sample size was another limitation that impacted the validation and accuracy of the study results.

6. Conclusion and Future Work

This study aimed to determine the critical factors that predict the length of stay, i.e., the predictor variables in the ED across three predetermined time range categories (low, medium, and high), utilizing ML algorithms. These categories were determined using unsupervised algorithms and took into account the impact of COVID-19 and various factors associated with the ED process. A case study was conducted in a local healthcare facility. Regression predictive modeling was initially utilized; however, it failed in our case due to the small size of the available data. Thus, classification algorithms were used, which showed high performance in predicting the best LOS category at ED. The best performance was achieved using Trees algorithms (decision stump, REB tree, and Random Forest) and the multilayer perceptron (with batch size 50 and 0.001 learning rate). Two scenarios were tested: the five categories and the three categories. The main reason the 3-CAT is better than the 5-CAT is the effect of widening the scoring scale in decision

problems in general. It becomes more difficult to distinguish between categories when their number increases. Thus, accuracy will increase for fewer categories, especially with a small sample size (400 is considered small in these models). The tradeoff between accuracy and informative classification is the main criterion for selecting three categories and not two.

As future work, the model can be expanded to include more than one facility and a larger dataset. Other factors might be considered to capture unusual situations and crises like pandemics for more accurate prediction. It is recommended to incorporate pandemic-related factors, such as mobility measures and healthcare service interruptions, into the training and evaluation processes to improve the model's accuracy in the postpandemic period. In addition, considering staff capacity, particularly the impact on nurses, during the initial stages of the pandemic can help mitigate the effects of clustering and improve the model's performance. By accounting for these pandemic-specific factors, the LOS prediction model can better adapt to the changing healthcare landscape and provide more reliable and accurate predictions.

Data Availability

The primary data used to support the findings of this study are included within the article. Additional data used to support the results of this study are available upon request from the corresponding author.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The researchers would like to acknowledge The Arab Medical Centre's support in this research for providing sample data.

References

[1] J. L. Wiler, S. Welch, J. Pines, J. Schuur, N. Jouriles, and S. Stone-Griffith, "Emergency department performance measures updates: proceedings of the 2014 emergency department benchmarking alliance consensus summit," *Academic Emergency Medicine*, vol. 22, no. 5, pp. 542–553, 2015.

[2] M. N. Saidan, M. A. Shbool, O. S. Arabeyyat et al., "Estimation of the probable outbreak size of novel coronavirus (COVID-19) in social gathering events and industrial activities," *International Journal of Infectious Diseases*, vol. 98, pp. 321–327, 2020.

[3] C.-H. Chaou, H.-H. Chen, S.-H. Chang et al., "Predicting length of stay among patients discharged from the emergency department—using an accelerated failure time model," *PLoS One*, vol. 12, no. 1, Article ID e0165756, 2017.

[4] M. Shbool, A. Al-Bazi, L. Zureigat, and A. Mahafzah, "Developing modern agent technologies in combating covid-19 exposure: an application in a healthcare facility," in *Proceedings of the 2022 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)*, pp. 700–706, Sakheer, Bahrain, November 2022.

[5] A. J. George, A. K. Boehme, J. E. Siegler et al., "Hospital-acquired infection underlies poor functional outcome in patients with prolonged length of stay," *ISRN Stroke*, vol. 2013, Article ID 312348, 5 pages, 2013.

[6] C. Dziegielewski, C. Skead, T. Canturk et al., "Delirium and associated length of stay and costs in critically ill patients," *Critical Care Research and Practice 2021*, vol. 2021, Article ID 6612187, 8 pages, 2021.

[7] X. Zeng, "Length of stay prediction model of indoor patients based on Light gradient boosting machine," *Computational Intelligence and Neuroscience*, vol. 2022, Article ID 9517029, 14 pages, 2022.

[8] R. Vaishya, M. Javaid, I. H. Khan, and A. Haleem, "Artificial intelligence (AI) applications for COVID-19 pandemic," *Diabetes & Metabolic Syndrome: Clinical Research Reviews*, vol. 14, no. 4, pp. 337–339, 2020.

[9] P. Yoon, I. Steiner, and G. Reinhardt, "Analysis of factors influencing length of stay in the emergency department," *Canadian Journal of Emergency Medicine*, vol. 5, no. 3, pp. 155–161, 2003.

[10] Z. Hu, B. Jin, A. Y. Shin et al., "Real-time web-based assessment of total population risk of future emergency department utilization: statewide prospective active case finding study," *Interactive Journal of Medical Research*, vol. 4, no. 1, p. e2, 2015.

[11] S. Levin, S. Barnes, M. Toerper et al., "Machine-learning-based hospital discharge predictions can support multidisciplinary rounds and decrease hospital length-of-stay," *BMJ Innovations*, vol. 7, no. 2, pp. 414–421, 2021.

[12] H. Kulkarni, M. Thangam, and A. P. Amin, "Artificial neural network-based prediction of prolonged length of stay and need for post-acute care in acute coronary syndrome patients undergoing percutaneous coronary intervention," *European Journal of Clinical Investigation*, vol. 51, no. 3, Article ID e13406, 2021.

[13] M. A. Abd-Elrazek, A. A. Eltahawi, M. H. Abd Elaziz, M. N. Abd-Elwhab, E. Abd, and M. N. Abd-Elwhab, "Predicting length of stay in hospitals intensive care unit using general admission features," *Ain Shams Engineering Journal*, vol. 12, no. 4, pp. 3691–3702, 2021.

[14] P. Walsh, P. Cunningham, S. J. Rothenberg, S. O'Doherty, H. Hoey, and R. Healy, "An artificial neural network ensemble to predict disposition and length of stay in children presenting with bronchiolitis," *European Journal of Emergency Medicine*, vol. 11, no. 5, pp. 259–264, 2004.

[15] M. A. Jenny, R. Hertwig, S. Ackermann et al., "Are mortality and acute morbidity in patients presenting with nonspecific complaints predictable using routine variables?" *Academic Emergency Medicine*, vol. 22, no. 10, pp. 1155–1163, 2015.

[16] A. Mustafa, R. Mohammad, M. Aljabri, M. Aboulnour, S. Mirza, and A. Ahmad, "Classifying the mortality of people with underlying health conditions affected by COVID-19 using machine learning techniques," *Applied Computational Intelligence and Soft Computing*, vol. 2022, Article ID 3783058, 12 pages, 2022.

[17] M. Gul and A. F. Guneri, "Yapay sinir ağlari kullanilarak acil servis hasta kaliş süresinin tahmini," *Journal of Aeronautics and Space Technologies (Havacilik ve Uzay Teknolojileri Dergisi)*, vol. 8, no. 2, p. 15, 2015.

[18] R. Taylor, J. R. Pare, A. K. Venkatesh et al., "Prediction of in-hospital mortality in emergency department patients with sepsis: a local Big data-driven, machine learning approach,"

Academic Emergency Medicine, vol. 23, no. 3, pp. 269–278, 2016.

[19] M. Street, M. Mohebbi, D. Berry, A. Cross, and J. Considine, "Influences on emergency department length of stay for older people," *European Journal of Emergency Medicine*, vol. 25, no. 4, pp. 242–249, 2018.

[20] E. Ağaçayak, R. Bugday, N. Peker et al., "Factors affecting ICU stay and length of stay in the ICU in patients with HELLP syndrome in a tertiary referral hospital," *International Journal of Hypertension*, vol. 2022, Article ID 3366879, 9 pages, 2022.

[21] E. E. Etu, L. Monplaisir, S. Arslanturk et al., "Prediction of length of stay in the emergency department for COVID-19 patients: a machine learning approach," *IEEE Access*, vol. 10, Article ID 42243, 51 pages, 2022.

[22] Z. Farahany, J. Wu, K. M. S. Islam, and P. Madiraju, "Oversampling techniques for predicting COVID-19 patient length of stay," in *Proceedings of the 2022 IEEE International Conference on Big Data (Big Data)*, Osaka, Japan, December 2022.

[23] M. A. Rahman, B. Honan, T. Glanville, P. Hough, and K. Walker, "Using data mining to predict emergency department length of stay greater than 4 hours: derivation and single-site validation of a decision tree algorithm," *Emergency Medicine Australasia*, vol. 32, no. 3, pp. 416–421, 2020.

[24] S. García, J. Luengo, and F. Herrera, "Tutorial on practical tips of the most influential data preprocessing algorithms in data mining," *Knowledge-Based Systems*, vol. 98, pp. 1–29, 2016.

[25] V. Roman, "Unsupervised machine learning: clustering analysis," 2019, https://towardsdatascience.com/unsupervised-machine-learning-clustering-analysis-d40f2b34ae7e.

[26] S. Wu, "How to evaluate my classification model results," 2021, https://towardsdatascience.com/top-5-metrics-for-evaluating-classification-model-83ede24c7584.

[27] M. Grootendorst, "Validating your machine learning model," 2020, https://towardsdatascience.com/validating-your-machine-learning-model-25b4c8643fb7.