A Performance Analysis on Patient Severity score based on Supervised and Unsupervised Learning

Asir Abrar[1], A. N. M. Sajedul Alam[1], Rimi Reza[1], Salsabil Ahmed[1], Tanvir Ahmed[1], Shihab Sharar[1], Annajiat Alim Rasel[1]

[1]*Computer Science and Engineering, BRAC University, Dhaka, Bangladesh*

{ *asir.abrar, a.n.m.sajedul.alam, rimi.reza, salsabil.ahmed, tanvir.ahmed8, shihab.sharar*}*@g.bracu.ac.bd*

, annajiat@bracu.ac.bd

Keywords: Supervised Learning, Unsupervised Learning, Clustering, Dimension Reduction, Patient Severity, Clustering, Gaussian Mixture Model, Supervised learning, Unsupervised learning, Machine Learning, Severity Score, Decision tree, Random forest

Abstract: This paper presents an analysis that is aimed at predicting the severity of patients according to their electronic health records in a specific time span by implementing supervised and unsupervised machine learning techniques. The proposed approach analyses the severity score prediction using an EHR dataset obtained from an open-source platform. The data pre-processing was done using tools such as openRefine. To implement the approach, different supervised learning algorithms such as Random Forest, XGBoost, LightGBM, KNN, CatBoost, and unsupervised learning algorithms such as K-means clustering, Spectral Clustering, Gaussian Mixture Model, Hierarchical Clustering and Principal Component Analysis are being used. The experimental results demonstrate that supervised learning outperforms unsupervised learning on the dataset for the expected outcomes.

1 INTRODUCTION

As of 2021, World Health Organization (WHO) had received reports of 228,206,384 confirmed cases of COVID-19 worldwide, with 4,687,066 deaths (who, 2021). In many cases, hospitals have been overwhelmed by the outpouring number of COVID-19 patients (ReliefWeb, 2021), (Hu, 2021). As a result, hospital authorities and doctors had to decide in many cases which patient should be given priority for ICU, CCU or which patient should be treated at home or at the hospital (Pandey, 2021). Given the gravity of the pandemic, Electronic Health Record (EHR) can be an excellent way to predict the health score of a patient. EHR is a digitalized or electronic rendition of a traditional paper chart of a patient's history (CMS, 2012). An EHR system not only contains a patient's medical and treatment history, it is also designed to go beyond traditional clinical data collected in a provider's office and can encompass a broader view of a patient's care (hea, 2019). Previously, EHR has been used for predicting several kinds of diseases and also the severity of the disease (Shickel et al., 2018). Some notable works include predicting which patients are the most vulnerable to venothromboembolism [VTE] after a hospital stay (Kawaler E, 2012), prediction of pancreatic cancer (Zhao and Weng, 2011), risk of heart failure (Taslimitehrani et al., 2016), in-hospital mortality in Sepsis patients in the emergency department (Taylor et al., 2016), diagnosis of stroke (Teoh, 2018) and estimation of delirium risk (Wong et al., 2018) and so on. During the COVID-19 pandemic, a great deal of research was done, and is still being done, to analyze, detect, forecast and diagnose COVID-19-related difficulties (Heldt et al., 2020), (Lee et al., 2021), (Razavian et al., 2020), (Vaid et al., 2020), (Sottile et al., 2021), (Paiva Proença Lobo Lopes et al., 2021).

2 PROPOSED MODEL

The goal of this research paper is to compare the performance of several machine learning models on a specific EHR dataset. The objective is to determine the influence of supervised and unsupervised learning on patient severity scores while identifying them. For supervised learning, we used Random Forest, XGBoost, LightGBM, KNN, CatBoost, and K-means clustering, Spectral Clustering, Gaussian Mixture Model, Hierarchical Clustering, and Princi-

pal Component Analysis. For unsupervised learning, we used K-means clustering, Spectral Clustering, Gaussian Mixture Model, Hierarchical Clustering, and Principal Component Analysis.

2.1 Dataset

The dataset was taken from 'Mendeley Data' (Sadikin, 2020), which is an open-source platform for datasets. The dataset was Electronic Health Record Predicting obtained from a private hospital in Indonesia, according to the article. Contains the patient's laboratory test results, which are used to determine the future course of treatment, whether the patient is in or out of care. The task included in the dataset is classification prediction.

2.2 Data Preprocessing

For unsupervised learning approach the data has been scaled and normalized for avoiding biased results while clustering.

2.3 Unsupervised Learning Approach

Unsupervised learning, also known as unsupervised machine learning, analyzes and clusters unlabeled datasets using machine learning techniques. Without the need for human intervention, these algorithms uncover hidden patterns or data groupings. It is the best solution for exploratory data analysis, cross-selling techniques, consumer segmentation, and image identification because of its capacity to detect similarities and differences in information (Education, 2020).

Here clustering models were implemented to group data in such a way that the severity of the patients can be predicted by individual health records. For each clustering model the processes will be described below:

2.3.1 Algorithms

- K-Means Clustering: When you have unlabeled data, K-means clustering (Scikit, 2021) is a sort of unsupervised learning (i.e., data without defined categories or groups). The purpose of this technique is to locate groups in the data, with K representing the number of groups. Based on the attributes provided, the algorithm assigns each data point to one of K groups iteratively. Data points are grouped together based on how comparable their features are.[5]

  Elbow diagram: In K-means clustering, the elbow method was implemented on the dataset to justify how many clusters should be taken in the count to proceed for getting highly efficient clustering results(GeeksforGeeks, 2021), (Trevino, 2016).

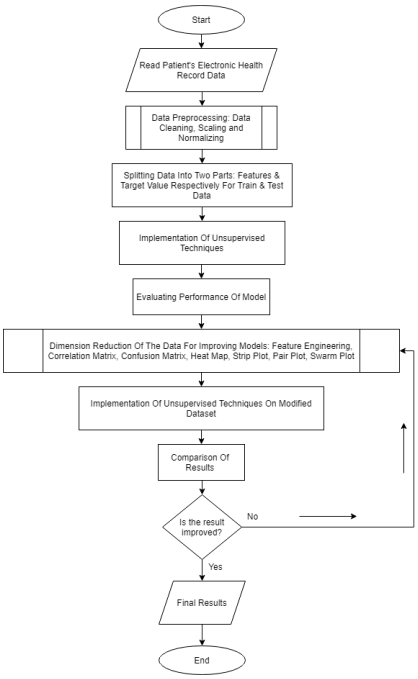

Figure 1: Unsupervised Machine Learning Workflow.

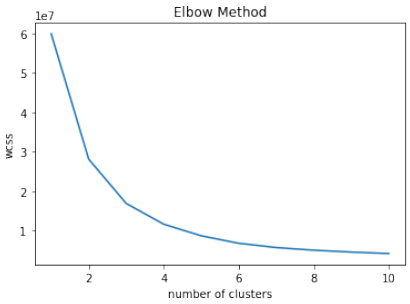

Figure 2: Elbow Diagram before Dimension Reduction.

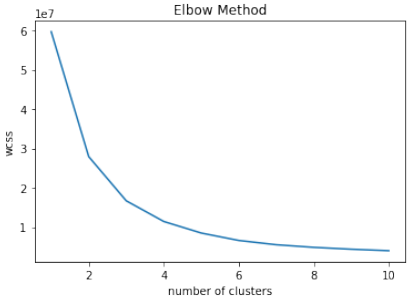

Figure 3: Elbow Diagram after Dimension Reduction.

- Hierarchical Clustering: Prior to using Hierarchical Clustering (Sharma, 2019), the data should be

Table 1: Dataset Description

Name	Data Type	Value Sample	Description
HAEMATOCRIT	Continuous	35.1	Patient laboratory test result of haematocrit
HAEMOGLOBINS	Continuous	11.8	Patient laboratory test result of haemoglobins
ERYTHROCYTE	Continuous	4.65	Patient laboratory test result of erythrocyte
LEUCOCYTE	Continuous	6.3	Patient laboratory test result of leucocyte
THROMBOCYTE	Continuous	310	Patient laboratory test result of thrombocyte
MCH	Continuous	25.4	Patient laboratory test result of MCH
MCHC	Continuous	33.6	Patient laboratory test result of MCHC
MCV	Continuous	75.5	Patient laboratory test result of MCV
AGE	Continuous	12	Patient age
SEX	Nominal	Binary	Binary/F/ Patient gender
SOURCE	Nominal	in,out	The class target in.= in care patient, out = out care patient

normalized such that each variable has the same scale. The significance of this is that the scales of the variables aren't the same, the model may be skewed toward variables of greater size.

Dendrogram: Dendrogram is a type of tree that represents hierarchical grouping relationships between similar data sets. They are commonly used in biology to show gene pools or patterns, but they can represent any type of clustered data (Glen, 2021). The dendrogram helps to decide the number of clusters for solving this particular problem.

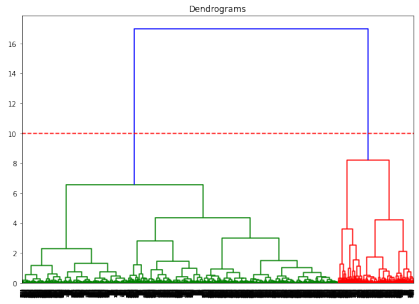

Figure 4: Dendrogram for Hierarchial Clustering.

- Spectral Clustering: Spectral clustering (Gupta, 2019) is an EDA approach for breaking down large multidimensional datasets into smaller groups of comparable data in rarer dimensions. The main goal is to categorize all unstructured data points into numerous categories based on their similarity (Chatterjee, 2020).The dataset was scaled into specific numerical values rather than scattered mixture values of different ranges. Furthermore, the data was normalized. After that, it was converted into a NumPy array into a pandas data frame. In the end, the dimensions of the data were reduced by PCA analysis[13].

  Principal component analysis (PCA): An unsupervised learning technique used in machine learning to reduce dimensionality. This research includes Spectral Clustering based on the PCA technique. After PCA analysis (Jaadi, 2021), the spectral clustering model is trained on that dataset.

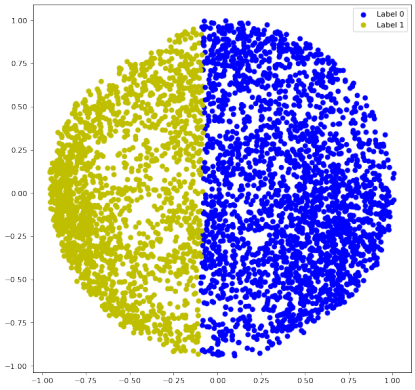

Figure 5: Spectral diagram before dimension reduction while constructing affinity matrix using radial basis function kernel.

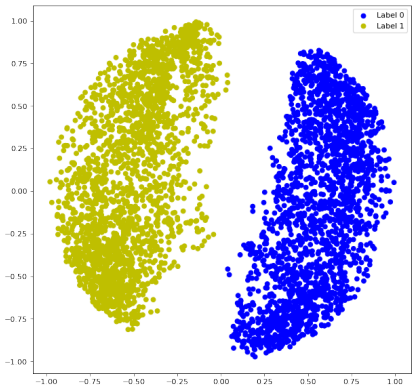

Figure 6: Spectral diagram after dimension reduction while constructing affinity matrix using radial basis function kernel.

- Gaussian Mixture Model: It presumes that a fixed

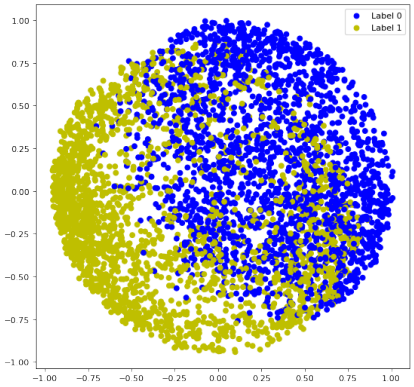

Figure 7: Spectral diagram before dimension reduction while constructing affinity matrix by computing a graph of nearest neighbors.

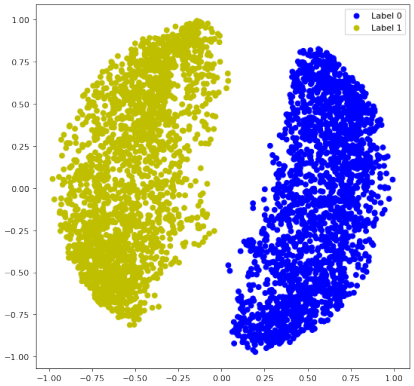

Figure 8: Spectral diagram after dimension reduction while constructing affinity matrix by computing a graph of nearest neighbors.

number of Gaussian distributions exist, each of which represents a cluster. As a result, the data points belonging to a single distribution tend to be grouped together in a Gaussian Mixture Model (Singh, 2020).

2.3.2 Dimension Reduction

With a high number of dimensions in the feature space, the volume of that space can be rather enormous, and the points (rows of data) that we have in that space frequently reflect a tiny and non-representative sample. Techniques for decreasing the number of input variables in training data are referred to as dimensionality reduction. One of the techniques of dimension reduction can be 'Feature Selection Method' (Brownlee, 2020).

Correlation Matrix: A strong correlation between dependent and independent variables is desirable, whereas a strong correlation between two independent variables is undesirable (Acharya, 2020). This notion is used for feature extraction to apply dimension reduction.

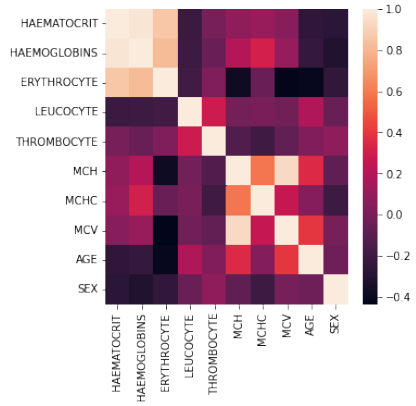

Figure 9: Correlation Matrix of Independent Variables.

2.4 Supervised Learning Approach

The machine learning job of learning a function that translates input to an output based on example input-output pairs is known as supervised learning. It derives a function using labeled training data, which consists of a collection of training instances.

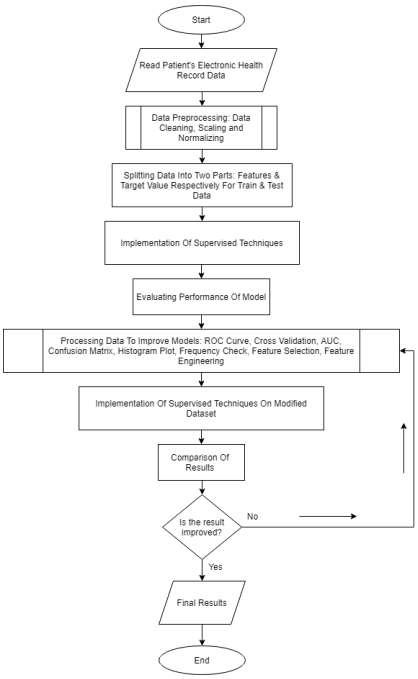

Figure 10: Supervised Machine Learning Workflow.

2.4.1 Algorithms

- Support Vector Machine (SVM) classifier: Both linear and non-linear data may be classified us-

ing this technique. It begins by mapping data into an n-dimensional feature space, with n being the number of features(Joachims, 1998). The hyperplane that divides the data items into two classes is then identified (Monien and Decker, 2005), (Liu et al., 2017).

- Decision Tree (DT): This is a non-parametric supervised learning approach that may be applied to both classification and regression tasks (Quinian, 1993), (Pedregosa et al., 2012).Internal nodes contain dataset characteristics, branches represent decision rules, and each leaf node provides the conclusion in this tree-structured classifier (Shah et al., 2020).
- XGBoost: This is a machine learning approach that can handle missing information without the need for imputation preprocessing and is extensively used for classification issues (Rusdah and Murfi, 2020).
- AdaBoost: By adaptively modifying the weak learning cycle, this technique allows weak classifiers to increase their performance (Li et al., 2005). It works well on well-balanced datasets but underperforms when noise is present (Misra and Li, 2020).
- LightGBM: Another gradient boosting approach that employs tree-based learning algorithms is this one. In terms of computing speed and memory usage, it outperforms XGBoost (Ke et al., 2017).
- KNN: KNN is known as a lazy learning algorithm because instead of learning from the training set directly, it stores the dataset and executes an action on it when it comes time to classify it (Kataria and Singh, 2013), (Wettschereck et al., 1997), (Abu Alfeilat et al., 2019).
- Random forests: Random forests are a set of tree predictors in which the values of a random vector selected separately and with the same distribution for all trees in the forest are used to forecast the behavior of each tree (Breiman, 2001), (Caruana and Niculescu-Mizil, 2006), (Menze et al., 2009).
- CatBoost: CatBoost is an open-source implementation of the Gradient Boosted Decision Tree (GBDT) for Supervised Machine Learning applications. CatBoost works effectively on machine learning challenges requiring category, heterogeneous data as a Decision Tree-based method (Hancock and Khoshgoftaar, 2020).
- Naive Bayes Classification: Given the category variable, the Naive Bayes (NB) classifier is a probabilistic classifier based on the premise that all characteristics are independent of one another (Xu, 2016).

3 RESULT

This research study employs a variety of evaluation measures. Precision, recall, accuracy, and F1 score have been seen in the supervised learning strategy, whereas additional clustering assessment approaches such as silhouette score, BIC score, and log-likelihood score have been observed in the unsupervised learning approach. These metrics vary depending on the number of dimensions reduced and other changes that occurred when the model hyper parameters were changed.

- Precision: The number of True Positives divided by the total number of True Positives and False Positives equals precision.
- Recall: The number of True Positives divided by the number of True Positives and False Negatives is the recall.
- Accuracy: This function computes subset accuracy in multilabel classification: the set of labels predicted for a sample must perfectly match the corresponding real label value (learn, 2021).
- F1 score: The F1 score represents the balance between accuracy and recall(Brownlee, 2019), (Joshi, 2016).
- Silhouette score: Silhouette refers to a way of interpreting and validating consistency inside data clusters. The method generates a simple graphical depiction of how effectively each object was categorized. When compared to other clusters, the silhouette value is a measure of how similar an object is to its own cluster(Wikipedia, 2021), (Kumar, 2021).
- BIC score: The Bayesian Information Criterion (BIC) is a scoring and selection tool for models. It takes its name from the field of research from which it arose: Probability and inference in Bayesian theory. As the model's complexity rises, the BIC value rises, and as the likelihood rises, the BIC value falls. As a result, the lower the number, the better.
- Log-likelihood score: The log likelihood value is a measure of the goodness of fit of each model. The higher the value, the better the model. We must remember that the logarithmic probability can be between Inf and + Inf. Therefore, the absolute consideration of the value cannot make a

statement. We can only compare log probability values between various models (Datalab, 2019).

Dimension reduction(DR) was done for unsupervised learning techniques by looking at the heatmap (Contributor, 2011) that showed the correlation matrix (Institute, 2019) between distinct features of the dataset. When clustering, a high correlation among independent features might lead to biases in the findings. As a result, strongly linked features were eliminated from the study. The correlation chart shows that 'HAEMATOCRIT', 'HAEMOGLOBINS', 'ERYTHROCYTE', 'MCH', and 'MCHC' are strongly linked characteristics, therefore they were excluded. In addition, the accuracy of this condensed dataset was compared to prior dataset results.

Table 2: K-Means Clustering Evaluation Metrics

DR	Precision	Recall	Accuracy	F1
Before	0.3036	0.3498	0.3251	0.4127
After	0.4922	0.6502	0.5602	0.5873

Table 3: K-Means Clustering Silhouette Score Analysis

DR	Silhouette Score
Before	0.3074
After	0.4672

Table 2, 3 shows that after dimension reduction all scores - Precision, Recall, Accuracy, F1 and Silhouette scores improved significantly.

From Table 4 and 5, it can be easily observed that, without dimension reduction, the linkage average hyper parameter is performing best and the second best is the linkage complete hyper parameter according to the Recall score. On the other hand, after dimension reduction, for both of the linkage hyper parameter the Recall score was improved. Additionally, according to Silhouette, F1, Accuracy and Precision scores, the scores for these two linkage respectively complete and average were improved for all scores.

Table 6 and 7 describes the evaluation metrics for Gaussian Mixture models. Prior to dimension reduction, the implementation of hyper parameters improved just the recall score. Furthermore, adding hyper parameters has no effect on any scores after dimension reduction.

Table 8 refers to the spectral clustering analysis based on BIC score and log-likelihood score. There are no changes in any scores for introducing hyper parameters without dimension reduction. Although, without applying hyper parameters, the Log-likelihood score deteriorated after dimension reduction, and with hyper parameters modified after dimension reduction, the BIC score dropped.

Figure 11: Supervised Learning Analysis Chart.

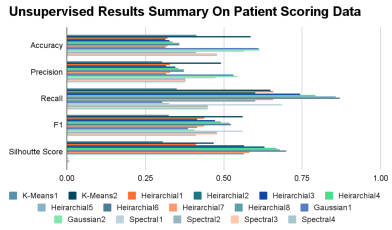

Figure 12: Unsupervised Learning Analysis Chart.

From the Table 9, it is observed that the Random Forest outperforms all other algorithms in case of accuracy which is 76%. On the other hand, the K Nearest Neighbor algorithm achieves the highest precision score 76%, where the precision score for Random Forest is 75%. In the case of Recall, we get the highest score from CatBoost, AdaBoost, and LightGBM which is 61%. The lowest recall score we get is for KNN which is 46%. Also, the CatBoost and LightGBM confer a good F1 score which is 67%. We see from the table that Naive Bayes does not make any sound score in any parameter for this dataset.

4 FUTURE WORK

As this work shows performance analysis based on unsupervised and supervised classification techniques, in the future, we will apply regression techniques on these datasets to see how it performs. Also, we will try to improve the classification techniques by doing feature selection and feature engineering. Through our work, we have come across a point where the Gaussian Mixture model works more limited after dimension reduction, which we want to work on later. This future task will include the improvement of the BIC score and the log-likelihood score of the Gaussian Mixture Model.

For dimension reduction, we have relied on feature extraction approaches. Other dimension reduction techniques, such as Matrix Factorization, Man-

Table 4: Hierarchical clustering evaluation before Dimension Reduction

Hyper Parameter: linkage	Precision	Recall	Accuracy	F1 score	Silhouette score
ward	0.3296	0.65863	0.32026	0.4393	0.6296
complete	0.3569	0.7926	0.3386	0.4922	0.62967
average	0.3726	0.8587	0.3583	0.5197	0.6781

Table 5: Hierarchical clustering evaluation after Dimension Reduction

Hyper Parameter: linkage	Precision	Recall	Accuracy	F1 score	Silhouette score
ward	0.3168	0.6003	0.3150	0.4147	0.5660
complete	0.3569	0.7926	0.3386	0.4922	0.6677
average	0.3739	0.8699	0.3585	0.5231	0.7004

Table 6: Gaussian Mixture Model before dimension reduction

HP	Precision	Recall	Accuracy	F1
Default	0.6638	0.4361	0.6826	0.5263
Changed	0.3104	0.5639	0.3173	0.4004

Table 7: Gaussian Mixture Model after dimension reduction

HP	Precision	Recall	Accuracy	F1
Default	0.3107	0.5779	0.3109	0.4041
Changed	0.3107	0.5779	0.3109	0.4041

ifold Learning, and Autoencoder Methods, can be used for this task.

Furthermore, similar studies may be performed on various EHR datasets or any COVID19 pandemic-related EHR datasets to see how these algorithms perform on various datasets.

5 CONCLUSION

In this paper, a performance analysis of predicting the patient severity scores has been proposed and successfully implemented by applying supervised and unsupervised machine learning techniques upon an EHR dataset. This analysis is planned and executed with consideration of blood test report data such as the patient's amount of Hematocrit, Hemoglobin, Erythrocyte, Thrombocytes and all the other measurements of different cells in blood. Our proposed approach can assess and forecast the severity of patients in cases when there is a shortage of medical beds in pandemics and we can't provide all of the patients beds in the hospital, so we have to make quick decisions about who needs a medical bed and who doesn't. In this instance, instead of human brain decisions, this strategy can make things faster as an automated procedure. Patients with less severe conditions can simply be alerted about their health status and taken home; on the other hand, when patients are at high risk, their families can take required steps to improve the patient's position. As a result of our research, we discovered that supervised classification algorithms outperformed unsupervised algorithms in predicting patient severity score.

Table 8: Spectral clustering Analysis

HP	DR	BIC Score	Log-likelihood Score
Default	before	-299093.3968	34.0200
Changed	before	-299093.4942	34.0200
Default	after	-299093.3968	12.4517
Changed	after	-109529.8661	12.4517

Table 9: Result Analysis Of Supervised Classification Algorithms

Algorithms	Accuracy	Precision	Recall	F1
SVC	0.72	0.72	0.52	0.60
XGBoost	0.74	0.72	0.59	0.65
RandomForest	0.76	0.75	0.59	0.66
NaiveBayes	0.69	0.66	0.51	0.57
LightGBM	0.75	0.74	0.61	0.67
KNN	0.73	0.76	0.46	0.58
DecisionTree	0.66	0.58	0.60	0.59
CatBoost	0.75	0.74	0.61	0.67
AdaBoost	0.66	0.58	0.61	0.59

REFERENCES

(2019). What is an electronic health record (ehr)?

(2021). Who coronavirus (covid-19) dashboard.

Abu Alfeilat, H. A., Hassanat, A. B., Lasassmeh, O., Tarawneh, A. S., Alhasanat, M. B., Eyal Salman, H. S., and Prasath, V. S. (2019). Effects of distance measure choice on k-nearest neighbor classifier performance: A review. *Big Data*, 7(4):221–248.

Acharya, T. (2020). Understanding feature extraction using correlation matrix and scatter plots.

Breiman, L. (2001). *Machine Learning*, 45(1):5–32.

Brownlee, J. (2019). Classification accuracy is not enough: More performance measures you can use.

Brownlee, J. (2020). Introduction to dimensionality reduction for machine learning.

Caruana, R. and Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. *Proceedings of the 23rd international conference on Machine learning - ICML '06*.

Chatterjee, M. (2020). What is spectral clustering and how its work?

CMS (2012). Electronic health records.

Contributor, T. (2011). What is heat map (heatmap)? - definition from whatis.com.

Datalab, A. (2019). Log-likelihood- analyttica function series.

Education, B. I. C. (2020). What is unsupervised learning?

GeeksforGeeks (2021). Elbow method for optimal value of k in kmeans.

Glen, S. (2021). Hierarchical clustering / dendrogram: Simple definition, examples.

Gupta, A. (2019). Ml: Spectral clustering.

Hancock, J. T. and Khoshgoftaar, T. M. (2020). Catboost for big data: An interdisciplinary review. *Journal of Big Data*, 7(1).

Heldt, F. S., Vizcaychipi, M. P., Peacock, S., Cinelli, M., McLachlan, L., Andreotti, F., Jovanovic, S., Durichen, R., Lipunova, N., Fletcher, R. A., and et al. (2020). Early risk assessment for covid-19 patients from emergency department data using machine learning.

Hu, C. (2021). Brazil's hospitals reach breaking point as health minister blames new coronavirus variants.

Institute, C. F. (2019). Correlation matrix.

Jaadi, Z. (2021). A step-by-step explanation of principal component analysis (pca).

Joachims, T. (1998). *Making large-scale SVM learning practical*. Univ., SFB 475.

Joshi, R. (2016). Accuracy, precision, recall f1 score: Interpretation of performance measures.

Kataria, A. and Singh, M. (2013). A review of data classification using k-nearest neighbour algorithm. *International Journal of Emerging Technology and Advanced Engineering*, 3(6):354–360.

Kawaler E, Cobian A, P. P. C. D. Y. S. C. M. (2012). Learning to predict post-hospitalization vte risk from ehr data.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30:3146–3154.

Kumar, A. (2021). Kmeans silhouette score with python examples - dzone ai.

learn, S. (2021). Sklearn.metrics.accuracy$_s$*core*.

Lee, J., Ta, C., Kim, J. H., Liu, C., and Weng, C. (2021). Severity prediction for covid-19 patients via recurrent neural networks.

Li, X., Wang, L., and Sung, E. (2005). A study of adaboost with svm based weak learners. *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*

Liu, C., Wang, W., Wang, M., Lv, F., and Konan, M. (2017). An efficient instance selection algorithm to reconstruct training set for support vector machine. *Knowledge-Based Systems*, 116:58–73.

Menze, B. H., Kelm, B. M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., and Hamprecht, F. A. (2009). A comparison of random forest and its gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics*, 10(1).

Misra, S. and Li, H. (2020). Chapter 9 - noninvasive fracture characterization based on the classification of sonic wave travel times.

Monien, K. and Decker, R. (2005). Strengths and weaknesses of support vector machines within marketing data analysis. *Studies in Classification, Data Analysis, and Knowledge Organization*, page 355–362.

Paiva Proença Lobo Lopes, F., Kitamura, F. C., Prado, G. F., Kuriki, P. E., and Garcia, M. R. (2021). Machine learning model for predicting severity prognosis in patients infected with covid-19: Study protocol from covid-ai brasil. *PLOS ONE*, 16(2).

Pandey, V. (2021). India covid: Delhi hospitals plead for oxygen as more patients die.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., and Louppe, G. (2012). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12.

Quinian, R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann.

Razavian, N., Major, V. J., Sudarshan, M., Burk-Rafel, J., Stella, P., Randhawa, H., Bilaloglu, S., Chen, J., Nguy, V., Wang, W., and et al. (2020). A validated, real-time prediction model for favorable outcomes in hospitalized covid-19 patients. *npj Digital Medicine*, 3(1).

ReliefWeb (2021). International medical corps india covid-19 situation report 1, may 5, 2021 - india.

Rusdah, D. A. and Murfi, H. (2020). Xgboost in handling missing values for life insurance risk prediction. *SN Applied Sciences*, 2(8).

Sadikin, M. (2020). Ehr dataset for patient treatment classification.

Scikit (2021). 2.3. clustering.

Shah, D., Patel, S., and Bharti, S. K. (2020). Heart disease prediction using machine learning techniques. *SN Computer Science*, 1(6).

Sharma, P. (2019). A beginner's guide to hierarchical clustering and how to perform it in python.

Shickel, B., Tighe, P. J., Bihorac, A., and Rashidi, P. (2018). Deep ehr: A survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE Journal of Biomedical and Health Informatics*, 22(5):1589–1604.

Singh, A. (2020). Gaussian mixture models: Clustering algorithm python.

Sottile, P. D., Albers, D., DeWitt, P. E., Russell, S., Stroh, J., Kao, D. P., Adrian, B., Levine, M. E., Mooney, R., Larchick, L., and et al. (2021). Real-time electronic health record mortality prediction during the covid-19 pandemic: A prospective cohort study.

Taslimitehrani, V., Dong, G., Pereira, N. L., Panahiazar, M., and Pathak, J. (2016). Developing ehr-driven heart failure risk prediction models using cpxr(log) with the probabilistic loss function. *Journal of Biomedical Informatics*, 60:260–269.

Taylor, R. A., Pare, J. R., Venkatesh, A. K., Mowafi, H., Melnick, E. R., Fleischman, W., and Hall, M. K. (2016). Prediction of in-hospital mortality in emergency department patients with sepsis: A local big data-driven, machine learning approach.

Teoh, D. (2018). Towards stroke prediction using electronic health records. *BMC Medical Informatics and Decision Making*, 18(1).

Trevino, A. (2016). Introduction to k-means clustering.

Vaid, A., Jaladanki, S. K., Xu, J., Teng, S., Kumar, A., Lee, S., Somani, S., Paranjpe, I., De Freitas, J. K., Wanyan, T., and et al. (2020). Federated learning of electronic health records improves mortality prediction in patients hospitalized with covid-19.

Wettschereck, D., Aha, D. W., and Mohri, T. (1997). A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms. *Artificial Intelligence Review*, 11(1):273–314.

Wikipedia (2021). Silhouette (clustering).

Wong, A., Young, A. T., Liang, A. S., Gonzales, R., Douglas, V. C., and Hadley, D. (2018). Development and validation of an electronic health record–based machine learning model to estimate delirium risk in newly hospitalized patients without known cognitive impairment. *JAMA Network Open*, 1(4).

Xu, S. (2016). Bayesian naïve bayes classifiers to text classification. *Journal of Information Science*, 44(1):48–59.

Zhao, D. and Weng, C. (2011). Combining pubmed knowledge and ehr data to develop a weighted bayesian network for pancreatic cancer prediction. *Journal of Biomedical Informatics*, 44(5):859–868.