IJITCE

ISSN 2347- 3657

International Journal of Information Technology & Computer Engineering

Email : ijitce.editor@gmail.com or editor@ijitce.com

International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 13, Issue 2, 2025

Patients' Health Analysis using Machine Learning

[1]Mr.Abdul Rais, [2] Abdul Rizwan Naveed, [3] Beeram Manichandu, [4] Shaik Abdullah Abdul Jabbar.
[1]Assistant Professor,Department of CSE-AIML, Lords Institute of Engineering & Technology.
[234] Student Department of CSE-AIML, Lords Institute of Engineering & Technology.

Abstract—

Examining patient health through the lens of Machine Learning (ML) was the primary objective of this research. We accomplished this by using the auto-ML-Pycaret and Extreme Gradient Boost (XGBoost) classifiers. This article details the three-step process we used to construct the XGBoost model: data analysis, feature engineering, and model development. Data science tools like Google Colab (GC) and the Jupyter notebook were used for these assignments. Our next topic is the auto-ML-Pycaret model, a powerful instrument for ML applications. At last, we compare the two models' performance according to their accuracy levels. We got 88% accuracy with the auto ML Pycaret model, whereas the initial ML model was 87% accurate. When comparing the XGBoost and auto-ML Pycaret models, we found that the latter had superior performance in terms of accuracy percentages and time factor.

Keywords— Machine Learning (ML), Pycaret, Accuracy, Health Pattern Check, Extreme Gradient Boost (XGBoost).

I. INTRODUCTION

Significant strides have been achieved in healthcare thanks to technological advancements. Knowing the patient's health state is crucial. As individuals age or get ill, many of their lives are put in jeopardy. Heart attacks, hypertension, cancer, diabetes, pneumonia, influenza, and respiratory illnesses are the leading killers [1]. The leading killer of both sexes is heart attacks [2]. There are a lot of individuals that end up in hospitals every day. To determine what caused the death, an accurate investigation is required. A lot of researchers have their own way of determining how healthy a patient is. On the other hand, this study demonstrates how to examine the patient's health by using state-of-the-art data science and machine learning techniques, and then draws conclusions from those findings. Doctors and academics may find new techniques to diagnose and forecast solutions if all the patient data were connected every day. There are multiple applications for patient data. the development of National Health Schemes (NHS) and related policies; the enhancement of patient safety; and the improvement of diagnosis. Have a better grasp of what puts people at risk for illness and mortality. Critical tasks include managing and making predictions based on massive amounts of patient data. Clinical decision making relies heavily on Artificial Intelligence (AI) and Machine Learning to accomplish this complicated job [3, 4]. Pfizer is only one of several companies that employ ML technology to find new drugs. The authors of [5] suggested using data mining methods to estimate how long cardiac patients will live. The healthcare industry makes extensive use of data mining tools. Data mining employs a variety of methods, including decision trees, ensemble models, and artificial neural networks (ANNs). Many machine learning models in healthcare try to forecast patient status based on variables like age, sex, variables such as body mass index (BMI), blood pressure (systolic and diastolic), temperature, respiratory rate, peripheral oxygen saturation (SP O2), urine output, platelets, neutrophils, basophils, lymphocytes, creatine kinase, creatinine, urea nitrogen, glucose, potassium, sodium, calcium, chloride, anion gap, magnesium ions, depression, hyperlipidemia, renal failure, atrial fibrillation, and hypertension [6]. By considering every possible attribute, one may build an ML model to predict future events. The application of ML models in these contexts is not limited to injury prediction and hospital discharge decisions [7]. Here, we developed a machine learning model to predict how sick patients would become. A person's medical history is what the model uses to derive their prediction score. The Extreme Gradient Boost (XGBoost) classifier, an ensemble method, was first used to construct the model. The model was constructed using a sequential method, as shown in Figure 1. Next, we compared the XGBoost model with an auto-ML-Pycaret model using accuracy ratings, and we choose the best one.

International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 13, Issue 2, 2025

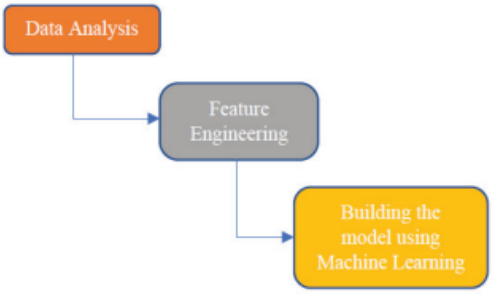


Fig. 1. Procedure for building the model using Machine Learning

Because of its ease of use and minimal coding requirements, the Pycaret ML library was chosen. Data scientists may greatly benefit from Pycaret's ability to automate the ML process. Anomaly detection, clustering, regression, and classification are just some of the many uses for the Pycaret package.

II. DATASET

We used data from https://healthdata.gov/, a resource with several features, to help with our forecast. The number of patients is 1177 and the key attributes are: id, outcome, age, gender, BMI, hypertensive, atrial fibrillation, CHD with no MI, diabetes, deficiency anemias, depression, hyperlipemia, renal failure, chronic obstructive pulmonary disease (COPD), heart rate, systolic blood pressure, diastolic blood pressure, respiratory rate, temperature, SP O2, urine output, hematocrit, red blood cells (RBC), mean corpuscular hemoglobin (MCH), mean cell hemoglobin concentration (MCHC), mean corpuscular volume (MCV), red cell distribution width (RDW), leucocyte, platelets, neutrophils, basophils, lymphocyte, prothrombin time (PT), n-terminal (NT)-proBNP, creatine kinase, creatinine, urea nitrogen, glucose, blood potassium, blood sodium, blood calcium, chloride, anion gap, magnesium ion, PH, bica The dataset with a few features is shown in Table I.

III. DATA ANALYSIS AND PRE-PROCESSING

Jupyter notebook and Google Colab (GC) were used for all data pre-processing and analysis. An ML model describing a patient's health state is constructed in this part using the XGBoost classifier model. Zero indicates that the individual is alive according to the model, while one indicates that they have passed away. Figure 2 shows the progression of the model. All necessary library resources: Data processing and computation were carried out using Pandas and NumPy. Seaborn, matplotlib, pyplot, plotly, and express are tools for visualizing statistical data. The dataset's missing data was handled by importing SimpleImputer. In order to prevent overfitting and achieve regularization, the XGBoost classifier was used. Using the StandardScaler scikit-learn, the data was standardized.

TABLE I. DATASET WITH FEW ATTRIBUTES

International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 13, Issue 2, 2025

group	ID	outcome	age	gender	BMI	hypertensive	atrial fibrillation	CHD with no MI	diabetes	deficiency anemias	depression	hyperlipemia
1	125047	0	72	1	37.58818	0	0	0	1	1	0	1
1	139812	0	75	2	NA	0	0	0	0	1	0	0
1	109787	0	83	2	26.57263	0	0	0	0	1	0	0
1	130587	0	43	2	83.26463	0	0	0	0	0	0	0
1	138290	0	75	2	31.82484	1	0	0	0	1	0	0
.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.
1	153461	0	83	2	22.31111	1	1	0	1	1	0	0

heart rate	systolic blood pressure	diastolic blood pressure	respiratory rate	temperature	SP O2	urine output	hematocrit	RBC	MCH	MCHC
68.83783784	155.8666667	68.33333333	16.62162162	36.7142857	98.394737	2155	26.2727273	2.96	28.25	31.52
101.3703704	140	65	20.85185185	36.6825397	96.923077	1425	30.78	3.138	31.06	31.66
72.31818182	135.3333333	61.375	23.64	36.4337037	95.291667	2425	27.7	2.62	34.32	31.3
94.5	126.4	73.2	21.85714286	36.287037	93.846154	8760	36.6375	4.2775	26.0625	30.4125
67.92	156.56	58.12	21.36	36.7619048	99.28	4455	29.9333333	3.286667	30.66667	33.66667
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
84.66666667	141.1304348	46.91304348	18.4	36.6736111	97.875	3039	28.8	2.867143	33.21429	33.74286

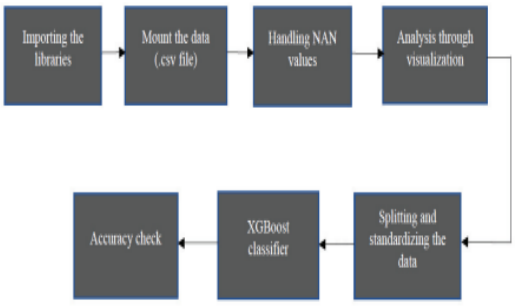
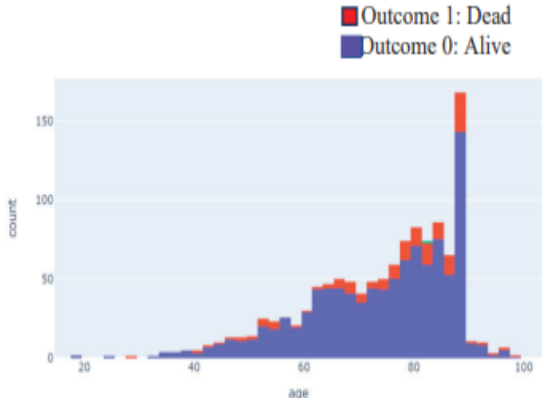


Figure 2 shows the evolution of the XGBoost model. As seen in Fig. 3, most patients died within 80-89 age group. There is a well-documented high mortality rate among the elderly.

International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 13, Issue 2, 2025



Data on ages (Fig. 3). Most patients with a body mass index (BMI) of 30–31 have passed away, as seen in Figure 4. Obesity was the cause of death for a few individuals whose body mass indexes were between 24 and 29.9. The body mass index (BMI) values are presented in Table II

Underweight	< 18.5
Healthy	18.5-24.9
Overweight	25-29.9
Obesity	≥ 30

.

TABLE II. RANGE BODY MASS INDEX (BMI) [8]

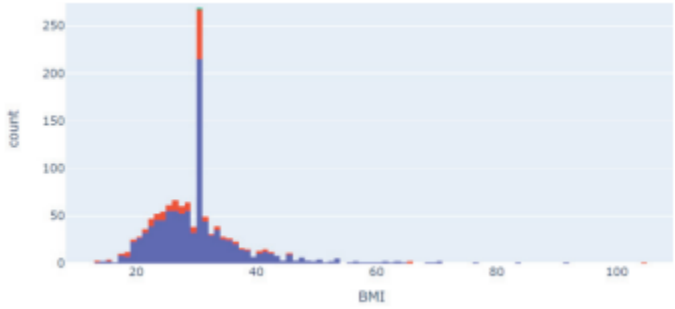


Fig. 4. BMI data.

Figures 5 and 6 provide graphical representations of the patient data, including SpO2 and heart rate readings. Percentages ranging from 95 to 100 are considered normal for oxygen levels. In most cases, a significant risk factor is indicated by a SpO2 value below 95%. The saturation of oxygen, or SpO2, is another name for it. The majority of patients were alive, with a SpO2 reading in the 95–100 range. Other causes of death occurred in people with normal SpO2 values. In a healthy adult, one might expect to hear 60–100 beats per minute. We also included more individuals whose heart rates were within the normal range in our esearch.

International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 13, Issue 2, 2025

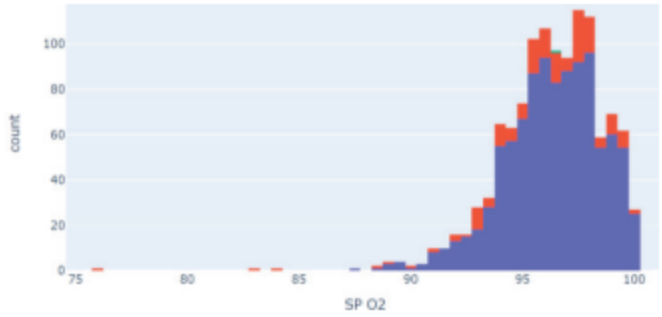


Fig. 5. SP O2 data.

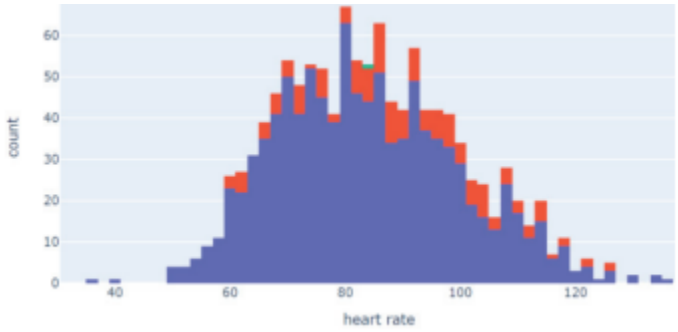


Fig. 6. Patients heart rate data.

Data pre-processing includes tasks such as data standardization, scaling, and splitting. At first, we just utilized the columns that were not labeled as "Id" for making predictions.

IV. RESULTS

We used the XGBoost classifier model, which is a classification model, to determine whether the patient was alive or dead using the outcomes 0 and 1, respectively. The patient is either dead or in grave danger if the result is 1, and they are alive or safe if the result is 0. While training the model, we used a random state value of 42 and imported the XGBoost classifier. The XGBoost model took the default settings into account. The following is an array displaying the prediction results: alive is denoted by zero and danger by one.

International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 13, Issue 2, 2025

```
array ([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,
        0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0,
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0,
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
        0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0,
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0,
        0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0,
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0,
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1,
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,
        0, 0])
```

As can be seen in Table III, the model's performance was checked and validated using a classification report. The precision score for 0 was 88%, while for 1, it was 68%. There was a 98% recall for 0 and a 25% recall for 1. And for a value of 1 the f1-score percentage is 36%, whereas for a value of 0 it is 93%. And 87% is the total success rate.

TABLE III. CLASSIFICATION REPORT.

	precision	recall	f1-score
0	0.88	0.98	0.93
1	0.68	0.25	0.36
Accuracy	0.87		

We ran a test on y_test and y_pred to see how the model predicts; the results are in Table IV. The model got the sequence number 353 wrong, but else it was spot on.

TABLE IV. MODEL PREDICTION ON TEST DATA

sq.no	y_test	y_pred
0	0	0
1	0	0
2	0	0
3	0	0

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 13, Issue 2, 2025

4	0	0
.	.	.
.	.	.
349	1	1
350	0	0
351	0	0
352	0	0
353	1	0

To evaluate the classification model's efficacy, one may make use of the Area Under the Curve (AUC) and the receiver operating characteristic (ROC) curve. Auxiliary unit of receiver operating characteristic curve, or AUROC for short. The accuracy of the prediction model improves as the AUC rises. As shown in Figure 7, our study's model achieved an AUC score of 84%. This proves the model was accurate.

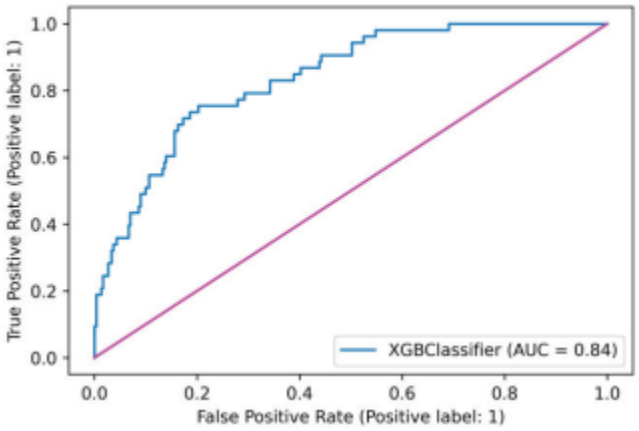


Fig. 7. AUROC curve.

V. AUTO ML-PYCARET

To train machine learning algorithms, it is a free and open-source Python library. A lengthy sequence of code is not necessary. Data is all that's needed for this. There was more than enough data to run the model. Training and predicting the results required less time. Table V provides the comparison of all models. An impressive 88% accuracy and 96% precision were achieved using the auto-ML-Pycaret model. In addition, the autoML model outperformed the XGBoost model in terms of training and prediction time.

Noteworthy points:

1. Before running the first model, we remove the 'Id' column. Nevertheless, in order to execute the Pycaret model, all columns were necessary in the auto-ML model.

2. Got errors because we did not restart the runtime. After restarting the runtime, the system worked perfectly

. **VI. DISCUSSION**

The auto-ML model outperformed the XGBoost model in terms of accuracy (88% vs. 87%), making it the superior model for outcome prediction. Not only that, auto-ML-Pycaret required less time to construct than the XGBoost model.

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 13, Issue 2, 2025

TABLE V. COMPARISON WITH OTHER CLASSIFIER MODELS.

Model	Accuracy	AUC	Recall	Precision	F1-score	Time(sec)
Auto-ML-Pycaret	0.8807	0.8559	0.8906	0.9661	0.9268	0.027
Extreme Gradient Boost	0.8703	0.8400	0.8769	0.9356	0.9194	0.319
Random Forest Classifier	0.8747	0.8192	0.1136	0.7000	0.1927	0.615
Ridge Classifier	0.8746	0.0000	0.2083	0.5917	0.3057	0.017
Extra Trees Classifier	0.8650	0.8054	0.0523	0.5000	0.0936	0.507
Logistic Regression	0.8649	0.7456	0.1462	0.4667	0.2198	0.725
Gradient Boosting Classifier	0.8649	0.8030	0.2614	0.5239	0.3435	0.572
Dummy Classifier	0.8613	0.5000	0.0000	0.0000	0.0000	0.014
Ada Boost Classifier	0.8552	0.7335	0.2886	0.4534	0.3482	0.207
K Neighbors Classifier	0.8504	0.5642	0.0167	0.2000	0.0308	0.125
Naive Bayes	0.8333	0.7569	0.3826	0.4253	0.3938	0.021
Decision Tree Classifier	0.8005	0.6038	0.3318	0.3010	0.3133	0.043
SVM - Linear Kernel	0.7930	0.0000	0.1402	0.1428	0.0768	0.026
Quadratic Discriminant Analysis	0.2384	0.4739	0.8152	0.1331	0.2285	0.024

VII. CONCLUSION

This article will be useful for analyzing the hospital's mortality rate, both historically and in light of recent events, such as the COVID-19 pandemic. Additionally, this article will provide some helpful pointers on how to tackle ML classification issues. The statistical visual depiction of the data offers information on medical terms. We achieved impressive levels of accuracy when experimenting with the XGBoost and Auto ML-Pycaret models. Less time was needed for training and result prediction when constructing the Pycaret model. Quickly and easily set up and train your data without the need for lengthy code. In light of these findings, we determined that the Pycaret model provided the most accurate results prediction.

REFERENCES

[1] A. Baldominos, A. Puello, H. Oğul, T. Aşuroğlu and R. ColomoPalacios, “Predicting infections using computational intelligence – A systematic review,” IEEE Access, vol. 8, pp. 31083-31102, Feb. 2020.

[2] E. Longato, G. P. Fadini, G. Sparacino, A. Avogaro, L. Tramontan and B. D. Camillo, “A deep learning approach to predict diabetes cardiovascular complications from administrative claims,” IEEE Journal of Biomedical and Health Informatics, vol. 25, pp. 3608- 3617, Sept. 2021.

[3] F. Ahamed and F. Farid, “Applying Internet of things and machinelearning for personalized healthcare: issues and challenges,” 2018 International Conference on Machine Learning and Data Engineering (ICMLDE), pp. 19-21, Jan. 2019.

[4] D. V. K, T. K. Ramesh and S. A, “A machine learning based ensemble approach for predictive analysis of healthcare data,” 2020 2nd PhD Colloquium on Ethically Driven Innovation and Technology for Society (PhD EDITS), pp. 1-2, Jan. 2021.

[5] P. R. Hachesu, M. Ahmadi, S. Alizadeh, and F. Sadoughi, “Use of data mining techniques to determine and predict length of stay of cardiac patients,” Healthcare informatics research, vol. 19, pp. 121– 129, Jun. 2013.

[6] E. L Thacker, B. McKnight, M. M. Psaty, W. T. Longstreth Jr, S. Dublin, P. N. Jensen, K. M. Newton, N. L. Smith, D. S. Siscovick, and S. R. Heckbert, A. “Association of body mass index, diabetes, hypertension, and blood

pressure levels with risk of permanent atrial fibrillation," Journal of General Internal Medicine, vol. 25, pp. 247- 53. Sep. 2012.

[7] Z. S. H. Abad, D. M. Maslove and J. Lee, "Predicting discharge destination of critically ill patients using machine learning," IEEE Journal of Biomedical and Health Informatics, vol. 25, pp. 827-837, Mar. 2021.

[8] R. Kumar, P. K. Dubey, A. Zafer, A. Kumar, and S. Yadav, "Design and Development of a Temperature-Compensated Body Mass Index Measuring System," Journal of Metrology Society of India (MAPAN), vol. 36, pp. 287–294, Apr. 2021.