Classification Of Healthcare Risk Factors Using Machine Learning

Mohammad Atiyeh[1], Mohammad Abo Rob[2], Mohammad Baker[3] and Yahya Jarrar[4]
Computer System Engineering Department, Arab American University
{m.atteyah, m.abualrob33, m.baker17, y.jarrar3}@student.aaup.edu

Abstract

Various forms of Machine Learning allow for improved analysis of data and informed decision-making across multiple areas of application. One significant area which can greatly benefit from the use of machines for decision-making is in the medical field due to numerous variables such as; complex medical datasets composed of mixed data types, numerous missing values, and the potential for noise in the dataset, therefore, allowing for limited success of traditional forms of analysis. In this paper, the authors provide a comparison of several different machine learning techniques for predicting the likelihood of heart disease in individuals based on one medical dataset. In addition to comparing the performance of machine-learning methods, the authors utilized multiple types of preprocessing (handling missing values, encoding categorical features, normalizing, determining outliers via anomaly detection, and selecting features) on the dataset before applying machine-learning methods. To better understand the scenario, clustering was performed on the data as a means of exploring and analyzing some of its internal structure.

Multiple supervised machine-learning models were used to assess and predict an individual's risk for developing heart disease including Support Vector Machine, k-Nearest Neighbours, Random Forest, Naive Bayes, AdaBoost, and Neural Networks. All of the models were assessed using the industry-recognized methodologies of accuracy, area under receiver operating curve, precision, recall, and F1-score. Overall, from the experimental results, the best two performing Machine Learning algorithms were k-Nearest Neighbours and the Neural Network model which both achieved approximately 70 percent accuracy while Random Forest and Naive Bayes produced roughly equal results at performing close to 69 percent. Both Support Vector Machines and AdaBoost produced lower performance results compared to other models on this dataset. Based on these findings it can be assumed that Machine Learning will be able to provide both reliable and realistic performance when used to analyze medical-dataset in order to assist with predicting disease risk classification.

Keywords

Machine Learning, Healthcare, Risk Classification, Support Vector Machine, KNN, Random Forest, Neural Network.

1. Introduction

Machine Learning is an exciting and fast growing area of computer science and artificial intelligence that allows us to use machines to find useful patterns in large and complex datasets. Machine Learning does not rely on creating rule based systems as in traditional programming; instead, these ML systems are able to "learn" automatically from large amounts of information and improve over time based on what you feed them. As a result, the field of Machine Learning has successfully been applied in a variety of fields, including finance, image processing, cyber security, and healthcare.

There are two main categories of Machine Learning models: supervised and unsupervised. Supervised learning techniques use the knowledge of what type of job they will perform, by using a labelled dataset, to train a model to complete its task, i.e., a classifier predicts an output (label) based on the input. An example of a classifier would be a Medical Image Diagnosis classifier, which classifies medical images (x-rays, CT scans, etc.) and a weather prediction model, which predicts a weather forecasting label based on various types of weather data sources. On the other hand, unsupervised learning seeks to find hidden structures and relationships in datasets without having any prior labels available for the dataset being analyzed in unsupervised learning algorithms, such as k-means clustering and principal component analysis.

Many common supervised and unsupervised machine learning algorithms are available and have been created over the years that can assist you in determining the most appropriate model(s) to use for your own experiments and building reliable and meaningful models.

2. RELATED WORKS

In [1], the authors employed machine learning methodologies to categorize glycaemic control risk factors in individuals with type 2 diabetes mellitus. The research utilized data from 647 patients, examining demographic, psychological, and physiological variables. We used regression tree analysis to find interactions between these risk factors. Several machine learning algorithms were tested, and the random forest classifier achieved the highest performance with 84% accuracy and 95% AUC. The results show that machine learning models can effectively classify T2DM patients when multiple risk factors are considered.

In the article referenced by [2], the authors applied supervised machine learning techniques to classify obesity levels using anthropometric indices obtained from bioelectrical impedance analysis. The dataset included 5372 adults, and several body composition indicators such as BMI, fat mass index, fat free mass index, and skeletal muscle index were used as input features. Six classification models were trained and evaluated using performance metrics including accuracy, F1-score, and AUC-ROC. The random forest classifier achieved the best results with an accuracy of 84.2% and an AUC-ROC of 0.947 among all tested models, The study also applied SHAP analysis, showing that FMI and FFMI were the most important predictors. This work demonstrates that machine learning, especially tree based models, can effectively classify obesity levels with high accuracy.

In paper [3], the authors proposed a machine learning approach for the detection of musculoskeletal pain risk using demographic, physical, lifestyle, and occupational variables. They used a dataset with 350 samples that underwent preprocessing steps which included missing data imputation, class label balancing with SMOTE, and scaling. They performed the task using a feedforward network, where Particle Swarm Optimization (PSO) technique was applied for the optimization process for the network's weights and biases. It performed excellently with an accuracy that ranged from 95.8% to 100%, hence able to carry out the task effectively as a risk classifier. Although the research had its limitations, it contributed towards realizing the effectiveness of optimization techniques in machine learning for the assessment of musculoskeletal pain risk.

In [4], the authors leveraged machine learning methods to figure out and forecast risk factors that are linked with Long COVID-19. The research hinged on longitudinal data for 601 individuals who had a confirmed case of COVID-19 infection. To forecast the risk of Long COVID-19, a random forest classification model was created. The model was very sensitive (97.4%) and had moderate specificity (65.4%).

The research singled out an accumulation of risk factors through the analysis, for instance, older age, more working hours per week, infection at the early stage of the pandemic, and being financially insecure. This work is a brilliant example of how machine learning can be used to unravel complex health risk factors and also be a tool for public health decision making.

The authors of paper [5] has been used sophisticated method of machine learning, aiming to predict and detect valvular heart diseases. The model they used is a two steps system that involves feature extraction and classification techniques. The goal is to increase the accuracy of diagnosis, and therefore, several machine learning algorithms were considered, being combined into hybrid approaches. The outcome of this work showed that the combination of Support Vector Machine and Principal Component Analysis proved the most successful, reaching an overall accuracy of 96.97% with a fair balance between F-measure, ROC area, sensitivity, and specificity measures. The majority voting method (MV5) was also proved to be a close rival method, thus providing high accuracy, high sensitivity, and high specificity simultaneously, providing important evidence relating to the effectiveness of combining dimensionality reduction techniques and robust classification algorithms concerning the diagnosis of cardiovascular diseases.

In [6], the authors have applied machine learning methods to create diagnostic prognostic models using the electronic health records of COVID-19 patients. The model developed aims to predict several clinical outcomes, such as positive test result for COVID-19 infection, the requirement of mechanical ventilation, death, hospital stay, and ICU stay. The proposed model had excellent predictive power with an AUC of 91.6% for the positive test result, 99.1% for the requirement of ventilation, and 97.5% for the probability of death. In the case of hospital stay and ICU stay, the model had excellent low mean absolute errors of 0.752 and 0.257 days, respectively. The paper reveals the importance of using machine learning techniques to provide predictive tasks within the context of clinical health.

In [7], the researchers proposed machine learning models for predicting the risk of ICU readmission and future ICU admissions. They used the MIMIC-III database, which contains the data of 42,307 patients who were readmitted to the ICU for several years. Thirteen variables were used from the database, which were the risk factors, and the three supervised learning models considered were Multilayer Perceptron, Random Forest, and Support Vector Machine. Among these models, the model that performed well had an accuracy of 86.4% with an AUC of 0.642.

3. METHODOLOGY

The data set contains health records of patients with varying numerical and categorical features, the target feature is the risk column. A data preprocessing applied to handle missing values, normalize features, and prepare the data for further analysis. The goal of this dataset is to apply machine learning techniques to predict the health risk level of individuals based on their health information.

3.1. DATA UNDERSTANDING

The dataset contains 30000 rows with 18 columns, including 17 features and one target variable. The target variable has two classes 0 and 1. The features include 2 nominal categorical attributes, 2 binary attributes, and 13 continuous numerical attributes.

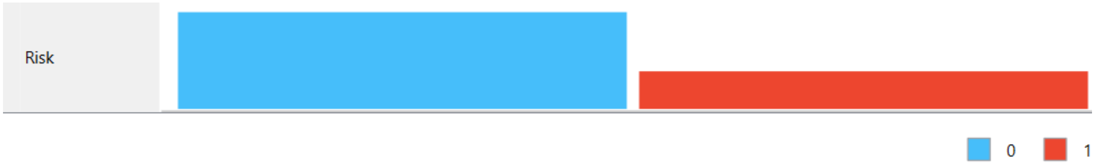


Figure 1. Distribution of risk column

The figure illustrates the distribution of the Risk variable, showing two classes 0 and 1 with class 0 being more frequent.

3.2. HANDLING MISSING VALUES

In the dataset, 16645 rows of the 30000 rows are contain missing values in one or more attributes. Since missing data can negatively affect data quality, appropriate preprocessing techniques were applied to handle this issue.

First, all rows containing missing values were removed in order to obtain a clean dataset suitable for further analysis, except a subset of 100 rows that contained missing values in the Age attribute was intentionally retained to evaluate an alternative missing value handling approach. For this subset, missing values were replaced using the average (mean) value. In the end of this process, we got a dataset contains 13455 rows without any missing data.

3.3. OUTLIERS DETECTION

Before moving to the Transformation stage, we identify the abnormal values that negatively influence the learning process of machine learning models by performing outlier detection using scatter plot to visualize and inspect the relationship between Physical Activity and Oxygen Saturation.

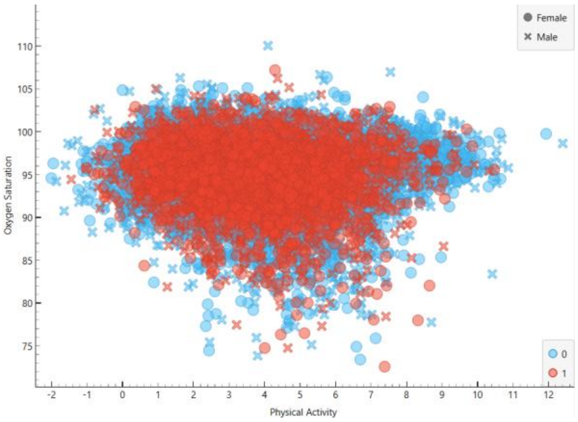


Figure 2. Outliers Detection

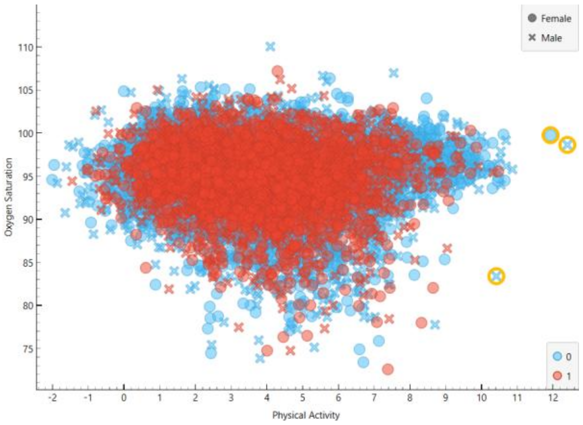


Figure 3. Outliers selected

As shown in Figure 2, most of the data are concentrated within a dense central region which indicating normal behaviour. However, you can simply see a small number of data appear isolated from the main distribution and deviate significantly from the overall data pattern.

3.4. Data Transformation

3.4.1. Encoding

Categorical features values must be converted into numerical values to optimize the machine learning algorithms work on the dataset. Nominal categorical attributes such as Gender and Medical Condition were encoded.

Table 1. Gender feature encoded values

Gender Category	Encoded Value
Female	0
Male	1

Table 2. Medical Condition feature encoded values

Medical Condition Category	Encoded Value
Arthritis	0
Asthma	1
Cancer	2
Diabetes	3
Healthy	4
Hypertension	5
Obesity	6

3.4.2. Normalizing

Most numerical features their values were scaled into a common range, specifically [0,1].

3.5. Removing Outliers

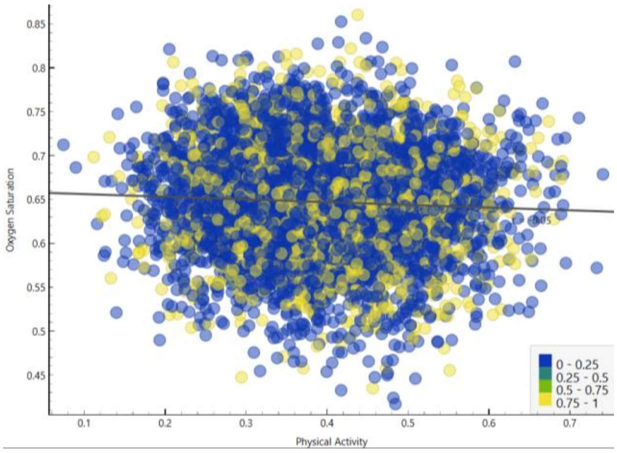


Figure 3. Removed Outliers

In this step, the best result was achieved using One-Class SVM, used as an unsupervised outlier detection method. The model learns the boundary of normal data values and identifies the abnormal data. Removing these outliers helped improve data quality and model robustness.

3.6. Clustering

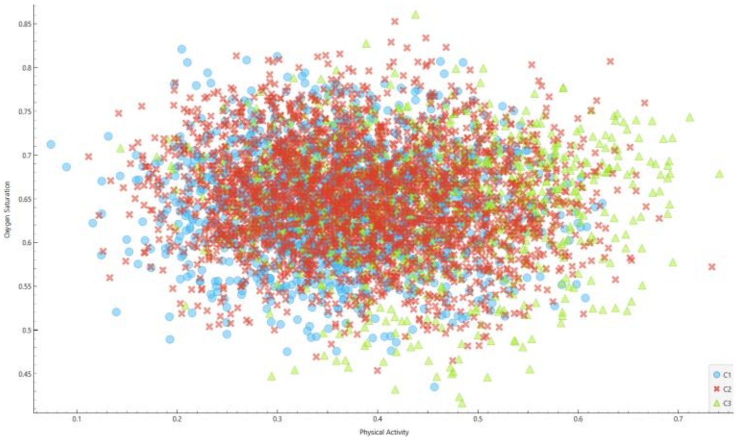


Figure 4. Clustering visualization

We applied clustering using k-means, one of the most popular unsupervised learning algorithms, dividing the data into three clusters (c1-c3) as shown in the Figure 4. Then, we used a scatter plot to display the distribution of this data by placing Physical Activity on the X-axis and Oxygen Saturation on the Y-axis.

3.7. Feature Selection

According to the Orange Tool ranking for the features, and searching through the web about what are the most important features that we must keep in our dataset, the set of high importance features contains 13 features, 4 features were removed.

Table 3. Removed Features

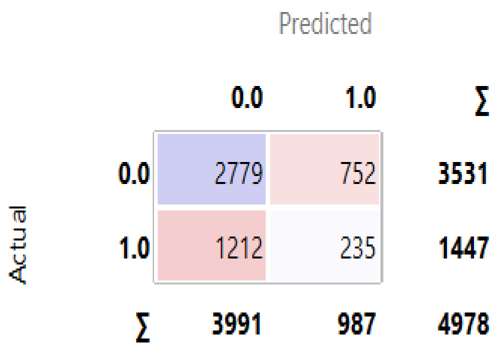
Feature Name	Decision
Gender	Removed
Medical Condition	Removed
LengthOfStay	Removed
Alcohol	Removed

4. RESULTS AND DISCUSSION

Table 4. Models Results

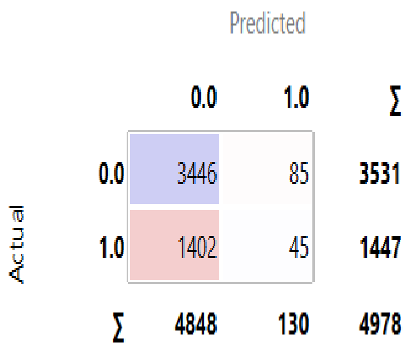
Model	AUC	Accuracy	F1	Precision	Recall
SVM	53%	60%	58%	56%	0.605
KNN	53%	70%	60%	60%	0.701
Random Forest	59%	69%	63%	63.5%	0.696
Naive Bayes	60%	69%	63%	63.7%	0.697
AdaBoost	52%	60%	60%	61%	0.605
Neural Network	60%	70%	58%	50%	0.709

The models did not give the same accuracy. KNN and Neural Network reached 70%. Random Forest and Naive Bayes were almost the same with about 69%. SVM and AdaBoost were lower and around 60% on this data.



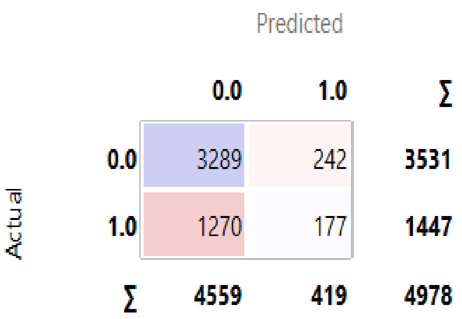
		Predicted		
		0.0	1.0	∑
Actual	0.0	2779	752	3531
	1.0	1212	235	1447
	∑	3991	987	4978

Figure 4. SVM Confusion Matrix



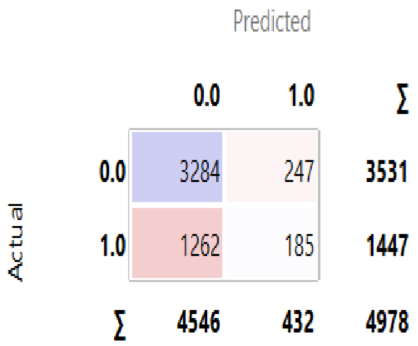
		Predicted		
		0.0	1.0	∑
Actual	0.0	3446	85	3531
	1.0	1402	45	1447
	∑	4848	130	4978

Figure 5. KNN Confusion Matrix



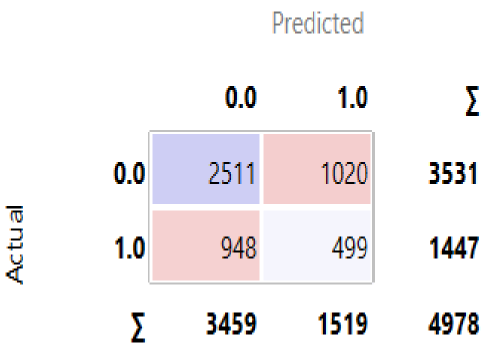
		Predicted		
		0.0	1.0	∑
Actual	0.0	3289	242	3531
	1.0	1270	177	1447
	∑	4559	419	4978

Figure 6. Random Forest Confusion Matrix



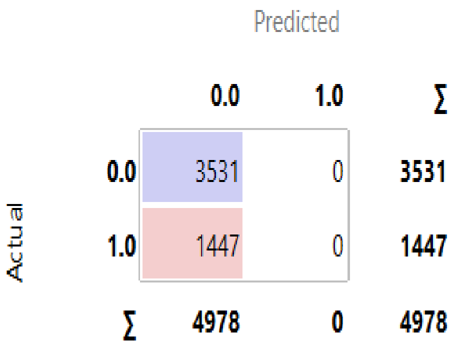
		Predicted		
		0.0	1.0	∑
Actual	0.0	3284	247	3531
	1.0	1262	185	1447
	∑	4546	432	4978

Figure 7. Naive Bayes Confusion Matrix



		Predicted		
		0.0	1.0	∑
Actual	0.0	2511	1020	3531
	1.0	948	499	1447
	∑	3459	1519	4978

Figure 8. AdaBoost Confusion Matrix



		Predicted		
		0.0	1.0	∑
Actual	0.0	3531	0	3531
	1.0	1447	0	1447
	∑	4978	0	4978

Figure 9. Neural Network Confusion Matrix

5. Conclusions

In general, KNN and Neural Network gave the best results for predicting healthcare risk, they perform 70% accuracy.

Still, the results depend on the data used and the type of problem.

Acknowledgements

We would like to thank Dr. Mohammad Awad for his help and guidance during this work. We also thank our team members for their effort and cooperation in completing this project.

REFERENCES

[1] Y.-L. Cheng, Y.-R. Wu, K.-D. Lin, C.-H. R. Lin, and I.-M. Lin, “Using machine learning for the risk factors classification of glycemic control in type 2 diabetes mellitus,” Healthcare, vol. 11, no. 8, p. 1141, 2023, doi: 10.3390/healthcare11081141.

[2] R. Yáñez-Sepúlveda, A. Vásquez-Bonilla, R. Olivares, et al., “Supervised machine learning algorithms for the classification of obesity levels using anthropometric indices derived from bioelectrical impedance analysis,” Scientific Reports, vol. 15, p. 30681, 2025, doi: 10.1038/s41598-025-15264-6.

[3] D. M. Fouad, M. M. Mahfouz, M. M. Mohamed, et al., “Classification of musculoskeletal pain using machine learning,” Scientific Reports, vol. 15, p. 27158, 2025, doi: 10.1038/s41598-025-12049-9.

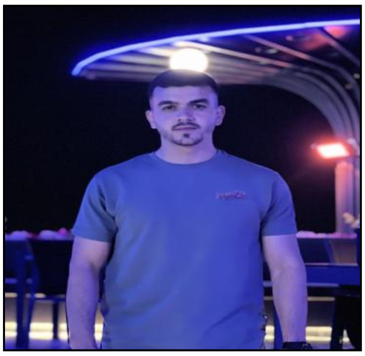
[4] R. Machado, R. Soorinarain Dodhy, A. Sehgal, K. Rattigan, A. Lalwani, and D. Waynforth, “A machine learning approach to identifying risk factors for long COVID-19,” Algorithms, vol. 17, no. 11, p. 485, 2024, doi: 10.3390/a17110485.

[5] M. U. Aslam, S. Xu, S. Hussain, et al., “Machine learning-based classification of valvular heart disease using cardiovascular risk factors,” Scientific Reports, vol. 14, p. 24396, 2024, doi: 10.1038/s41598-024-67973-z.

[6] M. Nasir, N. S. Summerfield, S. Carreiro, et al., “A machine learning approach for diagnostic and prognostic predictions, key risk factors and interactions,” Health Services and Outcomes Research Methodology, vol. 25, pp. 1–28, 2025, doi: 10.1007/s10742-024-00324-7.

[7] A. R. B. Junqueira, F. Mirza, and M. M. Baig, “A machine learning model for predicting ICU readmissions and key risk factors: analysis from longitudinal health records,” Health and Technology, vol. 9, pp. 297–309, 2019, doi: 10.1007/s12553-019-00329-0.

AUTHORS' BIOS

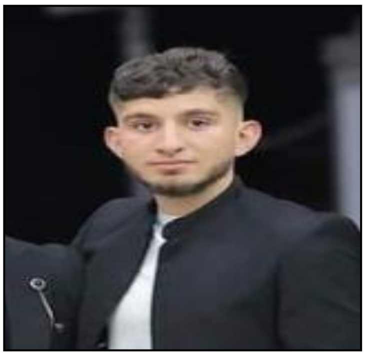


Mohammad Iyad Atiyeh is an undergraduate student in Computer Systems Engineering.His interests include machine learning, data preprocessing. He contributed to data preprocessing, feature selection, outlier detection, and classification model development.

Mohammad Awni Baker is an undergraduate student in Computer Systems Engineering.

His interests include machine learning, data preprocessing. He contributed to data preprocessing, feature selection, outlier detection, and classification model development.

Yahya Mohammad Jarrar is an undergraduate student in Computer Systems Engineering.His interests include machine learning, data preprocessing. He contributed to data preprocessing, feature selection, outlier detection, and classification model development.



Mohammad Riyad Abu Alrob is an undergraduate student in Computer Systems Engineering.His interests include machine learning, data preprocessing. He contributed to data preprocessing, feature selection, outlier detection, and classification model development.