ISSN: 2227-7994

FORUM A+P

INTERDISCIPLINARY JOURNAL OF ARCHITECTURE AND BUILT ENVIRONMENT

Revolutionizing Healthcare: Disease Prediction Through Machine Learning Algorithms

ANDIA VLLAMASI
POLIS University
KLEJDA HALLAÇI
POLIS University

Abstract

A crucial field of medical research is disease prediction, which has the potential to improve early diagnosis and therapy that can have a major impact on the course of treatment. By dramatically raising the standard of patient care and the effectiveness of the healthcare system as a whole, disease prediction plays a critical role in contemporary healthcare. Early detection of illnesses or medical issues, even before symptoms appear, is a key component of this proactive approach to healthcare management. This enables prompt interventions, better treatment outcomes, and better resource allocation. In this study, we use four different machine learning techniques to predict diseases using large datasets. Our main goal is to evaluate the effectiveness of different algorithms and determine which one performs best at accurately predicting the condition. To guarantee data quality and significance, the study makes considerable use of feature selection, engineering, and data preparation. Across various illness datasets, four machine learning algorithms, K-Nearest Neighbors, XG Boost, Ada Boost with SVM and Logistic Regression, are thoroughly examined. Accuracy, precision, recall, F1-score, and receiver operating characteristic area under the curve (AUC-ROC) are just a few of the performance criteria used to rate these algorithms.

The comparative study not only identifies the algorithm with the best predicted accuracy, but it also offers insightful information about the benefits and drawbacks of each strategy.

This study has significant healthcare impacts. We provide medical professionals with an effective tool for early detection and intervention by determining the algorithm that performs best at disease prediction. Improved disease prediction accuracy can result in earlier and more efficient treatment, which may save lives and lower healthcare costs. Additionally, this research opens the door for the application of sophisticated machine learning methods to clinical practice, ushering in a new era in healthcare where data-driven predictions support clinical judgment. In conclusion, by utilizing the potential of machine learning algorithms for more precise and timely disease prediction, our research supports the continual evolution of healthcare.

Keywords

AI (artificial intelligence); machine learning; healthcare; datasets; algorithm

Introduction

Artificial Intelligence (AI) is having a profound impact on many different industries and aspects of our daily lives, including work and social interactions. Artificial intelligence (AI) and machine learning (ML) have revolutionised the healthcare industry by radically changing the ways in which diseases are identified and treated. In the medical domain, artificial intelligence mostly focuses on developing algorithms and techniques to evaluate if a system's behaviour in identifying diseases is accurate, which also affects disease diagnosis and treatment approaches. A medical diagnosis determines the disease or diseases that cause a person's symptoms and indicators. Usually, diagnostic information is obtained through a physical examination and the patient's history. Because many indications and symptoms are unclear and may only be diagnosed by trained medical specialists, this process is frequently difficult.

Given that humans are prone to making mistakes, it is not surprising that a patient would receive an incorrect diagnosis more frequently. Overdiagnosis can result in problems such as needless medical intervention, which can impact people's health. A uncommon disease's condition, which leads the illness to be incorrectly dismissed from consideration, and a lack of relevant symptoms, which are sometimes undetectable, are two reasons why a misdiagnosis may occur. A branch of artificial intelligence called machine learning (ML) employs multi-dimensional clinical data as its input resource. A multitude of patient data sources, like as genetic information, imaging scans, medical records, and even lifestyle characteristics, are utilised by machine learning algorithms to uncover subtle patterns and connections that might not be immediately obvious to human clinicians. This makes it possible to diagnose illnesses like cancer, heart problems, and neurological issues earlier, which improves patient outcomes and allows for more prompt interventions. The application of machine learning (ML) in the healthcare industry is bringing about a paradigm change by using predefined mathematical functions to provide classification or regression results that frequently outperform human capabilities. Large volumes of patient data are analyzed by these ML algorithms, which then uncover complex connections and patterns that are frequently challenging for humans to achieve. The promise of inexpensive and quick machine-learning-based disease diagnosis (MLBDD) has lead to incorporation of ML in healthcare.

Traditional diagnosis procedures take a long time, are expensive, and frequently involve human intervention. Traditional diagnosis methods are constrained by the patient's capacity, but ML-based systems are unrestricted and do not experience human weakness. Also, there is often an absence of qualified healthcare workers in many low-resource environments, especially specialists with training in particular illness diagnosis. Algorithms for machine learning can function as virtual assistants, offering diagnostic assistance even in places where access to qualified medical personnel is restricted. As a result, a technique for diagnosing disease with unexpectedly high patient numbers in the healthcare setting might be created. This is more benefitial for developing nations like Albania that do not have enough medical professionals that are specialized for specific kind of diseases for their populations and struggle to provide appropriate diagnostic procedures for their maximum patient populations. Additionally, diagnostic procedures frequently call for medical tests, which low-income individuals frequently find to be pricey and difficult to afford. All things considered, machine learning-based illness diagnoses hold the potential to completely transform the way that healthcare is provided in developing nations by increasing access to knowledge, cutting expenses, improving patient outcomes, and enabling remote patient care. These technologies have the potential to alleviate healthcare inequities that impoverished groups experience globally as they develop and become more widely available.

Data, Algorithms, and Methods

In the current world, data is becoming an increasingly important resource. Data is turning into an essential part of decision-making in business, government, research, and other areas. Data is becoming a ubiquitous and extremely valuable resource that permeates almost every aspect of our daily lives in the age of technology. The work of American mathematician Claude Shannon, who is regarded as the pioneer of information theory, is where the idea of data originated in the context of computers. We have selected two different datasets for this research, with one containing information connected to diabetes and the other centered on heart-related data. Our analysis is based on these databases, which enable us to investigate trends, patterns, and connections related to these different health issues. We have carefully applied and assessed the results of four different machine learning algorithms—K-Nearest Neighbors, XG Boost, Ada Boost with SVM, and Logistic Regression—in each of these datasets. With this all-encompassing method, we hope to determine the effectiveness and appropriateness of every algorithm in correctly forecasting results and trends in the context of diabetes and heart-related data.. Our goal is to derive significant conclusions about these algorithms' performance and efficacy in tackling particular healthcare issues by analyzing their efficiency in light of multiple characteristics.

Data

The National Institute of Diabetes and Digestive and Kidney Diseases provided the diagnostic measurements for the first dataset that is being examined. This dataset is essential for estimating the probability that a patient would develop diabetes. This dataset is unique in that it focuses on a particular demographic: female patients who identify as Pima Indian and who are at least 21 years old. Multiple important health markers are carefully tracked in this extensive dataset. These contain the number of births in the country, serum insulin levels, diastolic blood pressure, plasma glucose concentration, body mass index, triceps skin fold thickness, diabetes pedigree function, age, and the final result expressed as a binary class variable. Specifically, this binary variable assumes a value of 0 to denote the absence of diabetes and 1 to signify the presence of diabetes.

This dataset is significant since it focuses on a specific population subset and has a wide range of health characteristics. Focusing on women of Pima Indian descent who are 21 years of age or older, the dataset allows for a more in-depth investigation of diabetes risk variables in this particular population. Consequently, the dataset offers significant insights that support a more focused and knowledgeable approach to comprehending and managing diabetes in this specific demographic. Adding a variety of health measures to the dataset enhances its quality and provides a comprehensive view of the various factors that influence the prognosis of diabetes in these people.

The second dataset is a combination of four different databases from Long Beach V, Cleveland, Hungary, and Switzerland. This dataset is unique in that it includes 76 variables in total, including the predictive feature, in contrast to numerous research that usually concentrate on a small group of 14 distinct traits. This dataset's main goal is to determine whether patients have heart disease, which is the main target variable. Heart disease is represented binaryly, with 0 signifying no disease and 1 signifying the presence of the condition. Several significant characteristics have been carefully documented in this large dataset to provide a thorough comprehension of the cardiovascular system. Age, sex, kind of chest pain, resting blood pressure, serum cholesterol levels, fasting blood sugar, ECG readings, maximal heart rate achieved, exercise-induced angina, exercise-induced ST depression, slope of the peak exercise ST segment, number of major vessels coloured by fluoroscopy, and a variable labelled "thal," which suggests the presence of heart defects, are some of these. The careful anonymization procedure used in this dataset to protect patient privacy and data security is a crucial component. Sensitive personal information is kept private by replacing patient names and social security numbers with dummy values. In addition to upholding moral principles, this calculated anonymization promotes the safe and ethical use of data for analysis and study. This dataset becomes a useful tool for researching and comprehending the subtleties of heart disease across a range of populations while upholding the highest standards of data protection and confidentiality by combining a wide range of variables and giving privacy measures priority

The performance of machine learning algorithms is heavily influenced by the number and diversity of datasets they are trained on. The importance of huge datasets cannot be overstated when employing semi-supervised, supervised, or unsupervised learning techniques. When using supervised learning, where models learn from labelled examples to make predictions, large datasets provide an abundant supply of varied examples. (Brownlee, J. (2016) This improves the accuracy with which algorithms are able to recognise nuances and connections between input characteristics and target labels. Likewise, larger datasets offer a greater variety of data points, facilitating the discovery of more significant and representative patterns in the context of unsupervised learning, when computers identify latent structures or groupings in unlabeled data. Large datasets also aid in semi-supervised learning, which uses both labelled and unlabeled data, by providing an abundance of instances for model enhancement and generalisation. Now that deep learning has become popular, large datasets are also necessary since deep neural networks require a lot of data to build hierarchical representations of complex data domains. Large datasets are fundamentally what allow machine learning algorithms—regardless of paradigm—to achieve higher levels of accuracy, robustness, and generalisation, hence increasing the potential and applications of artificial intelligence in a variety of sectors.

Algorithms

In the field of data analysis and decision-making, the proper algorithms chosen with care are crucial since they affect the quality of decisions made and the ability to extract valuable insights. Through an examination of four different algorithms and their applications to two different datasets, this research seeks to delve into the complexities of algorithm selection. By means of this investigation, we hope to demonstrate the flexibility and efficiency of these algorithms in tackling healthcare issues, highlighting their usefulness in yielding outcomes that are in line with the distinct features and requirements of every dataset. Our research will clarify the complex ways in which these algorithms can be used to address issues in healthcare. We aim to demonstrate the algorithmic prowess in deriving solutions that are accurate and customised to the nuances of the healthcare domain by taking into account the unique qualities and subtleties embedded in each dataset. In addition, the analysis will methodically reveal the benefits and limitations related to every algorithm, providing a thorough comprehension of how well they function in practical scenarios. We hope to give readers insightful knowledge about the complex field of algorithm selection in data analysis by means of a comparison analysis.

In the end, this investigation into algorithm selection makes a significant addition to the larger conversation about data-driven healthcare decision-making. The purpose of this work is to equip researchers, analysts, and decision-makers with the information necessary to successfully negotiate the complex terrain of data analysis by shedding light on the factors and consequences that go into selecting algorithms. By means of a detailed investigation of algorithmic applications in the healthcare domain, we want to augment the collective cognizance of optimal practices, so enabling well-informed decisions that propel progress in the domain of data-driven healthcare solutions.

The algorithms that we used are: XGBoost, AdaBoost, K-nearest neighbor (KNN) and Logistic regression (LR).

XGBoost

One popular and effective open-source implementation of the gradient boosted trees algorithm is XGBoost. This approach, which is classified as supervised learning, combines the estimations from several less powerful, simpler models to produce accurate predictions of a target variable. These weaker learners appear as regression trees in the context of regression with gradient boosting, where each tree translates an input data point to one of its leaves that contains a continuous score. It uses a regularised objective function to improve prediction accuracy. This function includes a penalty term that accounts for model complexity, specifically the regression tree functions, and a convex loss function that measures the difference between the target and projected outputs. Iteratively, the training process introduces new trees that forecast the residuals or errors of earlier trees. The final forecast is then created by combining these new predictions with the results of earlier trees. The algorithm's use of a gradient descent strategy to minimise loss while integrating new models into the ensemble is reflected in the name "gradient boosting" (xgboost, u.d.).

XGBoost's success can be attributed mostly to its scalability across all contexts. In distributed or memory-constrained environments, the system expands to billions of samples and operates more than ten times quicker than popular solutions now available on a single machine. XGBoost is scalable because of a number of significant systems and algorithmic improvements. Among these developments are: a theoretically justified weighted quantile sketch process that allows handling instance weights in approximate tree learning; a unique tree learning algorithm for handling sparse data. Rapid learning with parallel and distributed computing facilitates faster model exploration. The XGBoost algorithm's primary goal is to minimise the objective function that is comprised of the regularisation terms and loss function:

$$L^{t} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t\left(x_i\right) \right) + \Omega\left(f_t\right). \quad (1)$$

These days, XGBoost is the most popular approach for creating predictive models because of its exceptional precision, effectiveness, and flexibility. (T. Chen, 2016)

AdaBoost

AdaBoost, is an ensemble machine learning algorithm that can be used in a wide variety of classification and regression tasks. It is a supervised learning algorithm that is used to classify data by combining multiple weak or base learners (e.g., decision trees) into a strong learner. (Schapire in 2013).

It was first presented by (Freund & Schapire in 1997). The method weights every example equally and begins with a weak learner. Boosting iterations is the process of applying weights w_1,w_2,...w_n to every training sample in order to achieve this. All of the weights are initially adjusted evenly to w_i=1/n. The training dataset is then used to train a weak learner on the original data. In the training phase, the new distribution weight is lowered when the error rate falls and vice versa, with an increase in the sample's distribution weight occurring when it does. After that, samples are continuously trained using weights from an unknown distribution. By lowering the error of the subsequent machine and ultimately achieving higher accuracy rates, the goal is to receive positive feedback. It is easy to locate the AdaBoost algorithm's process, and one such study is in (Lu, Hu, & Bai, 2015). (Schapire, 2003) explains a basic AdaBoost algorithm; relevant Python packages are readily available in (Api reference).

The AdaBoost algorithm of this study works through the following steps:

1. Initially, training and test subsets are randomly created and assigned using Adaboost.
2. By choosing the training set, iteratively trains the model.
3. In order to provide incorrectly categorised observations a higher probability of categorization in the following iteration, it gives them a larger weight.
4. The algorithm uses the customised classifier to assign the weight to the classifier at the end of each iteration.
5. The process keeps going until all of the training data fits perfectly or there are no more estimators left.
6. To classify, conduct a "vote" among classifiers and select the result based on the constructed model.

K- nearest neighbor (KNN)

K- nearest neighbor (KNN) is created by Evelyn Fix and Joseph Hodges in 1951 and is a nonparametric classification method which is simple but effective in many cases according to the seminal work by Hand, Mannila, and Smyth (2001), titled Principles of Data Mining. Both classification and regression analysis can be done with KNN. Class membership is the conclusion of KNN classification. A voting system is utilized to categorize the thing. The distance between two data samples is calculated using Euclidean distance methods, which may be computed using the following formula :

$$D(p,q) = \sqrt{(p_1 - q_1)^2 + \ldots + (p_n - q_n)^2} \quad (2)$$

where p and q are the ones with n attributes that need to be compared. This is what the knn() function uses by default (Weinberger, 2009). There are alternative ways to compute distance as well, like the Manhattan distance (Breiman 2001) .

The category or class of any particular dataset can be easily solved with the help of the K-NN algorithm. The following is an understanding of how the K-NN algorithm operates:

Step 1: Choose the neighbour with the K-number first.

Step 2: For every K neighbours, the Euclidean distance must be determined.

Step 3: Using the computed Euclidean distance, select the K closest neighbours.

Step 4: Determine how many data points there are in each category among these K neighbours.

Step 5: The category with the greatest number of neighbours will receive new data points.

Step 6: The K-NN classification model will then be prepared in this manner.

Logistic regression (LR) is a ML technique that is employed to address classification problems. With projected values ranging from 0 to 1, the LR model has a probabilistic framework. Malignant tumor detection, the identification of spam emails, and the detection of online fraud transactions are a few examples of LR-based ML. LR employs the cost function, also referred to as the sigmoid function. Every real number between 0 and 1 is transformed by the sigmoid function. Logistic Regression can be expressed as,

$$log \ (\frac{p(x)}{1-p(x)}) = \beta_0 + \beta_1 X \quad (3)$$

where the left-hand side is referred to as the logit or log-odds function, and the expression p(x)/(1-p(x)) is called odds. The ratio of success to failure chances is known as the odds. Consequently, a linear combination of inputs is converted to log(odds) in logistic regression, yielding an output of 1. A straight line that depicts the association between the variables and clean data with no significant surprises between data points are prerequisites for using logistic regression. This model is applicable to any datasets; however, some assumptions must be taken into account in order to get optimal performance. In binary logistic regression, the dependent variable needs to be binary. Include only the variables that are meaningful. There must be no correlation between the independent variables. In other words, the model should have very little or no multicollinearity. The independent variables have an equal relationship with the log odds. Logistic regression requires large sample sizes. (Fix, E., & Hodges, J.L. 1989)

Methods

A preliminary data analysis revealed that the dataset did not contain missing values. Additionally, data anomaly identification, an analysis for the examination of the data components and descriptive statistics of the data have all been provided. We then split the data in two sets, the first 80% is used for model training, and the 20% is used for model prediction. The decision to treat the modelling as 80-20 was taken after the discussion on the lack of data available. Another possible approach would be using rolling forecasts, which fix the horizon for the test set prediction and use an expanding set of training data as more data become available. We decided to use the first option since the algorithms that we use incorporates the concept of rolling forecasts. In fact, the algorithm implements a method known as cross-validation, which is a variant of rolling forecasts. By using this approach, the algorithm is able to test the performance of the model on different periods of the test set while using an expanding set of training data as more data become available. This helps to provide a more accurate assessment of the model performance and allows for the selection of the best prediction model. After organizing the data, we built the XGBoost, Ada Boost, KNN and logistic regression for Diabet Disease and Heart Disease datasets. The evaluation of these algorithms is made based on accuracy scored. The accuracy is calculated using the accuracy_score function from, which returns the count (normalise = False) or the fraction (default) of accurate predictions. The subset accuracy is 1.0 if all of the projected labels for a sample precisely match the true labels; if not, it is 0.0. If $\hat{y_i}$ is the predicted value of the i-th sample and y_i is the corresponding true value, then the fraction of correct predictions over nsamples is defined as where 1(x) is the indicator function. (Scikit-Learn developers in 2021)

$$accuracy(y, \hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} 1(x)(\hat{y_i} = y_i)$$

Results

First we studied Heart disease dataset using four different algorithms, XGBoost, Ada Boost, K-Nearest Neighbors and Logistic Regression. For a number of reasons, dividing the data into training and testing sets is crucial. First of all, it enables us to evaluate our machine learning model's performance on hypothetical data, giving us a more accurate picture of how well it works in actual situations. A part of the data, in this example 20%, can be set aside for testing so that we can confirm the model's generalizability to fresh data and identify any possible overfitting problems. Moreover, the model can better identify patterns and relationships in the data by being trained on a bigger subset of the sample (usually 80%). This makes it easier to create a solid, accurate model that can produce trustworthy forecasts. After spliting the data, 80% for model training, and 20% for model prediction, we notice that Ada Boost has the accuracy score 0.89. This algorithm is used to classify data by combining multiple weak or base learners (e.g., decision trees) into a strong learner. The combination of Adaboost and SVM-based component classifier has been noted for its deviation from the standard Boosting principle due to the complexities associated with SVM training and the challenge of balancing diversity and accuracy in comparison to basic SVM classifiers. In the study, the Adaboost classifier was trained with SVM as the base classifier, incorporating a dynamic alteration of the kernel function parameter σ value, which gradually decreased in correspondence with the fluctuation of the weight value of the training sample. The performance of the proposed classifier was evaluated through experimentation on human subjects, focusing on the classification of left- and right-hand motor imagery tasks. The testing phase revealed an impressive average classification accuracy of 90.2% on the test data, showcasing a significant improvement compared to SVM classifiers that do not integrate Adaboost and the commonly used Fisher Linear Discriminant classifier. The findings provide substantial evidence that the integration of Adaboost with SVM classifiers has the potential to enhance accuracy in the classification of motor imagery tasks, thus offering promising prospects for enhancing the performance of brain-computer interface (BCI) systems." (Scikit-Learn developers, n.d.)

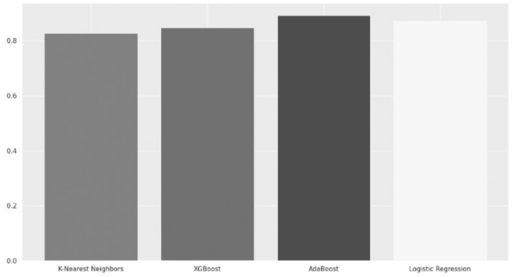

Figure 1: Accuracy Score for Heart disease dataset

The accuracy ratings of several machine learning algorithms used to forecast cardiac disease are shown in Table 1. The accuracy score, which is given as a percentage, indicates how effectively the model identifies cases properly. The table can be interpreted as follows:

1. XGBoost: This algorithm accurately predicts the condition of heart disease in about 84.7% of cases, with an accuracy score of 0.847.
2. Ada Boost: With an accuracy score of 0.89, it outperforms other algorithms and suggests a high degree of accuracy in heart disease prediction. About 89% of the time, this method classifies cases accurately.
3. KNN (K-Nearest Neighbours): Displays an accuracy value of 0.75, meaning that in 75% of cases, it accurately predicts heart disease. It shows decent predictive performance, however not as good as the other algorithms.
4. Logistic Regression: Gets an accuracy score of 0.82, meaning that 82% of the time, it accurately predicts heart disease. In terms of accuracy, this algorithm lies in between XGBoost and KNN.

To sum up, out of all the algorithms evaluated in this table, Ada Boost seems to be the most accurate. Nonetheless, a number of variables, including the dataset's properties, interpretability, and processing efficiency, influence the best algorithm selection. For a thorough assessment of the model's performance, various metrics such as precision, recall, and F1-score must be taken into account.

Algorithms	XGBoost	Ada Boost	KNN	Logistic Regression
Accuracy score	0.847	0.89	0.75	0.82

Table 1: Accuracy score for Heart Disease

Using key variables like age, sex, exercise-induced angina (exang), ST depression induced by exercise relative to rest (oldpeak), slope of the peak exercise ST segment (slope), number of major vessels coloured by fluoroscopy (ca), thalassemia (thal), resting electrocardiographic results (restecg), maximum heart rate achieved (thalach), and exercise-induced angina (exang), the heatmap reveals insightful patterns within the heart disease dataset. Brighter colours in the heatmap imply higher correlations, while the heatmap's colour intensity represents the strength of the correlations. Interestingly, there are strong connections between the "thal" variable—which is thought to indicate the existence of heart defects—and the variables linked to chest pain. This implies that these elements are critical to comprehending heart disease. Subsequent investigation may show that specific kinds of chest discomfort and the particular traits indicated by "thal" may be important indicators of heart disease. As a result, in the context of cardiac disease, these characteristics might be given priority for risk assessment and predictive modelling.

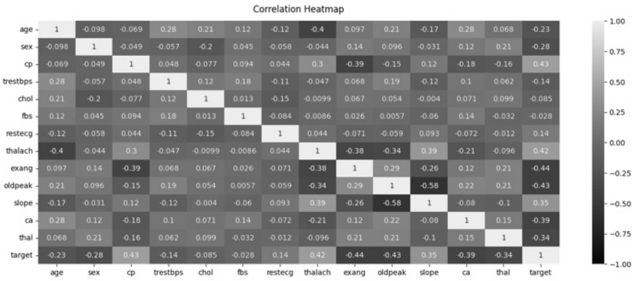

Figure 2: Heat map for Heart disease dataset

The heatmap indicates that the brighter the colors the higher the correlation and vice versa. We can see that glucose is highly correlated to the dependent varaiable, which invariables means that the above medical facts is shown to be true in the given data. Also, Insulin is the next correlated independent variable in the given data, but obviously does not correlate with outcome which medically is true because Insulin levels are used to predict the type of diabetes. Based on these facts, the following conclusion is made: That Glucose is the major predictor of diabetes Insulin is the major indicator of the type of daibetes. For the above reasons, only the glucose and insulin will be used to predict diabtes and type of diabetes.

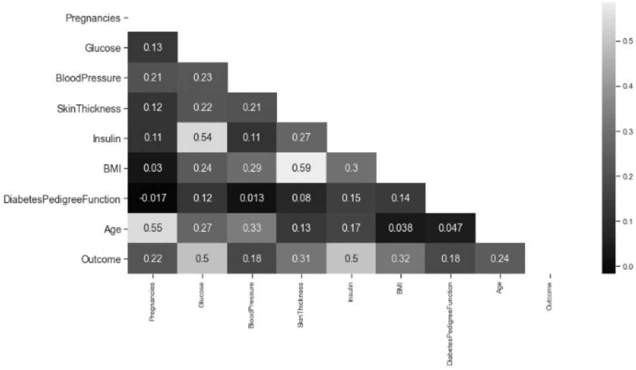

Figure 3: Heatmap for Diabetes Disease

The above plot shows the training and test set accuracy on the y-axis against the setting of n_neighbors on the x-axis. Considering if we choose one single nearest neighbor, the prediction on the training set is perfect. But when more neighbors are considered, the training accuracy drops, indicating that using the single nearest neighbor leads to a model that is too complex. The best performance is somewhere around 13 neighbors. The accuracy score of each algorithm is given on the table below. Here's a succinct explanation: XGBoost: Achieved a 0.75 accuracy rating. Ada Boost: Achieved a 0.66 accuracy rating. With KNN (K-Nearest Neighbours), an accuracy score of 0.75 was attained. With a score of 0.79, logistic regression produced the best accuracy result of all the described

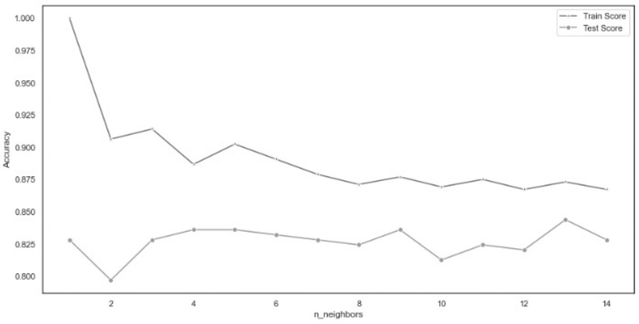

Figure 4: Test vs Training for KNN Algorithm

Algorithms	XGBoost	Ada Boost	KNN	Logistic Regression
Accuracy score	0.847	0.89	0.75	0.82

Table 2: Accuracy score for Diabetes Disease

techniques. Accuracy is a gauge of how successfully a model forecasts the right results. The better the model is in making accurate predictions on the given task, the higher the accuracy score. Thus, based on the scores given, it appears that the most accurate algorithm for this problem is Logistic Regression, which is followed by XGBoost and KNN. Ada Boost did slightly worse in terms of accuracy.

Conclusions

In this study, we conducted a thorough evaluation of four machine learning algorithms on two important healthcare datasets. The goal was to identify the algorithm that had the best predicted performance across all datasets. We have come to relevant conclusions about the effectiveness of these algorithms through a thorough evaluation using performance measurements. AdaBoost stood out as the most efficient algorithm for heart disease prediction, delivering the highest accuracy, precision, recall, and F1-score. Its capacity as an effective tool in healthcare was demonstrated by its capacity to merge poor learners into a robust model, which proved essential for spotting prospective heart disease cases. On the other hand, Logistic Regression excelled in predicting diabetes. Its usefulness in forecasting health outcomes was underlined by its simplicity, clarity, and strong performance in this particular application. It is impossible to overestimate the importance of using machine learning in healthcare, particularly in countries like Albania with little resources. By providing precise and timely disease forecasts, which is essential in contexts with limited resources, machine learning algorithms have the potential to change healthcare. These algorithms can help healthcare personnel make better decisions, allocate resources more efficiently, and ultimately improve patient care by leveraging the power of data. Furthermore, the ability to anticipate diseases, as this study's results show, not only improves the effectiveness of healthcare systems but also has the potential to help people and governments spend less money. The early diagnosis and prevention of diseases by machine learning can have a significant influence on public health, especially in economically underdeveloped areas like Albania where access to cutting-edge medical facilities may be restricted. Machine learning can contribute to cost-effective healthcare solutions by detecting high-risk patients and advising preventive steps, thereby enhancing the quality of life for people and communities. The research's conclusions highlight the potential of machine learning algorithms in healthcare applications. We can use the predictive power of machine learning to address health concerns in a targeted and efficient manner by customizing algorithms to particular datasets and health difficulties. This study emphasizes the revolutionary potential of machine learning in the field of medicine and serves as an important first step in improving healthcare in resource-limited areas.

Reference List

xgboost. (s.d.). Tratto da https://github.com/dmlc/xgboost

T. Chen, C. G. (2016). XGBoost: a scalable tree boosting system. *International Conference on Knowledge Discovery and Data Mining*, (p. 785-794). New York, USA.

Schapire, R.E. (2013). Empirical Inference. Springer; Berlin/Heidelberg, Germany. In "*Explaining AdaBoost*" (pp. 37–52).

Fix, E., & Hodges, J.L. (1989). Discriminatory analysis. Nonparametric discrimination: Consistency properties. *International Statistical Review/Revue Internationale de Statistique*, 57, 238–247. doi: 10.2307/1403797.

Scikit-Learn developers. (2021). Supervised Learning—Scikit-Learn 1.1.11 Documentation. Available online: https://scikit-learn.org/stable/supervised_learning.html.

Scikit-Learn developers. (n.d.). Model Evaluation: Quantifying the quality of predictions. Scikit-Learn Documentation. Available online: https://scikit-learn.org/stable/modules/model_evaluation.html#accuracy-score.

Pfurtscheller, G., Neuper, C., Guger, C., Harkam, W., Ramoser, R., Schlögl, A., Obermaier, B., Pregenzer, M. (2000). Current trends in Graz brain–computer interface (BCI) research. IEEE Trans. Rehab. Eng., 8, 216–219.

Brownlee, J. (2016). *Machine Learning Mastery with Python*. Machine Learning Mastery Pty Ltd., 527, 100–120.

Lu J., Hu H., Bai Y. (2015). Generalized radial basis function neural network based on an improved dynamic particle swarm optimization and AdaBoost algorithm. *Neurocomputing*, 152, 305–315. doi: 10.1016/j.neucom.2014.10.065.

Freund Y., Schapire R.E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1), 119–139. doi: 10.1006/jcss.1997.1504.

Schapire R.E. (2003). The boosting approach to machine learning: An overview. Nonlinear Estimation and Classification *Lecture Notes in Statistics*, 149–171. doi: 10.1007/978-0-387-21579-2_9.

Fix, E. and Hodges, J.L. (1951). *Discriminatory Analysis, Nonparametric Discrimination: Consistency Properties*. Technical Report 4, USAF School of Aviation Medicine, Randolph Field.
Hand, D., Mannila, H., & Smyth, P. (2001). Principles of Data Mining. The MIT Press.

Weinberger KQ, Saul LK. (2009). Distance metric learning for large margin nearest neighbor classification. The Journal of *Machine Learning Research*, 10, 207–244.

Cost S, Salzberg S. (1993). A weighted nearest neighbor algorithm for learning with symbolic features. *Machine Learning*, 10, 57–78. doi: 10.1007/BF00993481.

Breiman L. (2001). Random forests. *Machine Learning*, 45, 5–32. doi: 10.1023/A:1010933404324.