Predictive Analysis in Medical Care through tools and techniques of Machine Learning

Kanchan Naithani,
Asst. Professor, Department of Comp. Sc. & Info. Tech., Graphic Era Hill University,
Dehradun, Uttarakhand India 248002,

Abstract:
Medical sector is the area where vast amount of data is needed to store with privacy. In conventional ways all this is done manually, but now all the things are digitized and which gives more flexibility to maintain and store the data. Also accessing and analyzing the data is more easy and accurate rather than manual way. This article shows the medical care data analysis by using the machine learning approaches. Basically we use three approaches to manage and result calculation & prediction of the data i.e. RF, DT & SVM.

Keywords: Data Mining, health care, Machine learning, SVM, RF, DT.
DOI Number: 10.48047/NQ.2022.20.4.NQ22339 **NeuroQuantology2022;20(4): 1143-1148**

1. Introduction

Digital information and technology are growing rapidly in all sectors. It is essential to step out in the medical or health care sectors, by analyzing the huge amount of digital medical records. Machine learning tools and techniques process huge amount of health care data and provide good insights which can be the difference between life and death for some patients. By using the digital technology in the health care it increase the speed of the providing health records and gives more priority about the patient healthy status.

Machine Learning algorithms are of two types: Supervised Learning and Unsupervised Learning. In Supervised Learning the machine learns to predict by analyzing examples of data which have been already classified. Here humans provide the machine with guidance by providing it examples which are already processed and the machine try to imitate the results .This is also why it's called supervised learning as humans supervise the machine. Supervised Learning is also called predictive models as it tries to guess the target value based of the other values. In supervised learning the machine is given clear instruction on what they need to learn and how they are supposed to learn it. Some of the examples of machine learning algorithms are Random Forest, Support Vector Machine, and Decision tree etc.

Unsupervised Learning on the other hand doesn't get any help from humans. It is generally used for gaining unknown insights which are over looked by the humans. Unlike the supervised learning where the machine has knowledge of the results, the goal of unsupervised learning is not to get any result but to group similar data or summarize the data. Some of the examples are K – Means, BIRCHetc.

2. RelatedWorks

Machine Learning has played a huge part in our lives in recent years as the computer science develops more sophisticated and accurate machine learning algorithms are created. Machine learning allows machine to think like us and find insights from huge amounts of data which doesn't make sense to humans. Many multinational companies like Google, Amazon, Accenture and many others are using machine learning to improve their products and services. Google has used machine learning for Google translate, spotify

for music recommendation system, tesla for auto pilot for its electrical cars. Machine learning is used in wide range of applications.
Machine Learning is effective in finding insights in huge amounts of data. In medical application machine learning algorithms can produce better results in predicting or detecting diseases and can also be a key factor in preventing diseases.

3. Datasets:

Heart attack possibility:
Heart attack is a heart related disease which occurs when the heart cannot pump blood. This causes death if not treated early so detecting it is very important. This dataset contains 13 features and 1target value.

Cardio Vascular Diseases:
Cardiovascular disease is a major disease resulting in millions of death worldwide. This disease is related to heart and most of the time leads to death if not diagnosed properly. This dataset contains11 features and70000 records.

Diabetes Health Indicators Dataset:
Diabetes is one of major diseases in United States which affects millions of Americans each year. The body is unable to produce insulin to maintain normal glucose level in blood and lead to problems and also decrease life expectancy. This dataset contains22 features.

4. Methodology

We applied different machine learning on medical datafor prediction of target value.

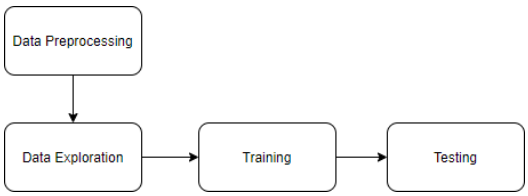

Figure 1: steps of medical data prediction

We followed these four steps:

1. Data Preprocessing: Data preprocessing is an essential step in preparing data for machine learning algorithms. It involves various techniques to clean and transform the data into a format that can be easily understood by machines. The process includes the following steps:
1.**Data Splitting:** The first step is to split the data into two parts: the training data and the test data. The training data is used to build the machine learning model, while the test data is used to evaluate the model's performance.
2. **Handling Missing Data:** In this step, we identify any missing data points in the dataset. These missing values can be either removed from the dataset or replaced using statistical methods such as mean, median, or mode imputation. This ensures that the dataset is complete and ready for analysis.
3. **Encoding Categorical Labels:** Many machine learning algorithms require numerical inputs. Hence, categorical variables need to be converted into numerical representations. This can be achieved by assigning unique numerical codes to each category (label encoding) or by creating dummy variables for each category (one-hot encoding).
4. **Data Normalization:** It is crucial to normalize the data to bring all features to a similar scale. Different features may have different ranges or units, which can negatively impact the performance of certain algorithms. Common normalization techniques include min-max scaling or z-score standardization.
5. **Handling Unbalanced Data:** Sometimes, the dataset may have an imbalanced distribution of classes, where one class has significantly more samples than the others. This can lead to biased predictions. To address this issue, techniques like SMOTE (Synthetic Minority Over-sampling Technique) can be used to generate synthetic samples of the minority class, thus balancing the dataset.
By performing these preprocessing steps, we ensure that the data is in a suitable format for machine learning algorithms, improving the accuracy and reliability of the models' predictions.

2.Data Exploration: Data exploration is a crucial

step in the data analysis process, where we aim to uncover meaningful patterns, characteristics, and features within the dataset. It involves utilizing various statistical techniques to gain insights and understand the data better. The following steps are typically involved in data exploration:

1. Identifying Initial Patterns and Characteristics: In this step, we examine the dataset to identify any initial patterns or characteristics. This can involve visualizing the data through plots, charts, or summary statistics to gain a preliminary understanding of the data's distribution, central tendencies, and variations.

2. Feature Selection: The goal of feature selection is to identify the most relevant features that contribute to predicting the target variable. Statistical measures such as skewness can help identify features with skewed distributions. If necessary, transformations can be applied to reduce skewness and make the data more suitable for analysis. Additionally, correlation analysis can be performed to identify features that are highly correlated with the target variable or with each other.

3. Dealing with Multicollinearity: Multicollinearity occurs when two or more independent variables in a dataset are highly correlated. This can create challenges in accurately estimating the effect of individual variables on the target variable. To address multicollinearity, statistical techniques such as variance inflation factor (VIF) analysis can be used to identify and mitigate the presence of highly correlated variables.

By conducting thorough data exploration, we gain insights into the dataset's characteristics, uncover relevant features, and identify potential issues such as multicollinearity. This process enhances our understanding of the data, facilitates effective feature selection, and enables more accurate predictions and analysis.

3.Training: During the training phase, various machine learning algorithms such as Random Forest, Support Vector Machine (SVM), and Decision Tree are employed to process the data. These algorithms are used to build models that can make predictions or classifications based on the provided dataset. The training process involves the following steps:

1. **Algorithm Selection:** Different machine learning algorithms are considered based on the nature of the problem and the characteristics of the dataset. Random Forest, SVM, and Decision Tree are some examples of algorithms that can be utilized. The selection of the algorithm depends on factors such as the complexity of the problem, the amount of available data, and the desired output.

2**. Hyperparameter Tuning:** Each machine learning algorithm has certain parameters, known as hyperparameters, which need to be set before the training process. To achieve optimal performance, these hyperparameters are tuned. This can be done through a hit and trial approach, where different combinations of hyperparameter values are tested to find the best configuration. Alternatively, hyperparameter tuning techniques such as grid search or random search can be employed to systematically explore the hyperparameter space and find the optimal values. The aim of hyperparameter tuning is to fine-tune the algorithm's settings to achieve the best possible performance on the given dataset. This process enhances the model's ability to generalize well to new, unseen data.

By leveraging different machine learning algorithms and tuning their hyperparameters, we optimize the model's performance and increase its accuracy in making predictions or classifications.

4.Testing: Once the model is trained, it is necessary to evaluate its performance using test data. The testing phase involves the following steps to assess the accuracy and effectiveness of different algorithms for analyzing healthcare data:

1. Model Evaluation: The trained model is tested using a separate set of test data. The predictions made by the model are compared with the known true values in the test dataset. The accuracy of the model is determined by calculating the percentage of correct predictions.

2. Algorithm Comparison: Different algorithms used during the training phase are compared based on their accuracy scores. The accuracy of each algorithm is computed, and the algorithm with the highest accuracy is considered the most suitable for analyzing healthcare data. This comparison allows us to identify the algorithm that performs the best on the given dataset.

3. Evaluation Metrics: In addition to accuracy, other evaluation metrics such as precision, recall, and F1 score are commonly used to assess the model's performance. Precision is calculated by dividing the number of true positives by the sum of true positives and true negatives. Recall, on the other hand, is calculated by dividing the number of true positives by the sum of true positives and false negatives. These metrics provide insights into the model's ability to correctly identify positive cases (precision) and its ability to capture all positive cases (recall).

By conducting rigorous testing and comparing different algorithms using various evaluation metrics, we can determine the best algorithms suited for analyzing healthcare data. This helps in selecting the most accurate and effective model for further analysis and decision-making in the healthcare domain.

5. Machine Learning Algorithms

Random Forest:

Random Forest is the result of combination of two concepts Decision Tree and Ensemble learning. Random Forest as the name suggests make many decision tree on subsets of given datasets and with the help of ensemble learning combine the results to improve the accuracy of the dataset.

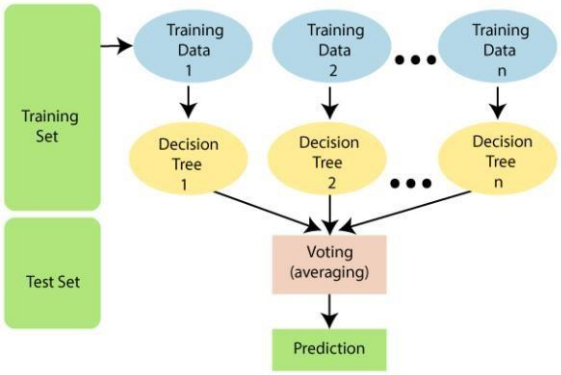

Figure 2: random Forest

Random Forest works in two steps. First it creates a lot of decision trees from subsets of the dataset and then tries to give prediction for each of the trees. For new data points it tries to find predictions for each decision tree and take the majority vote to decide the class of the new point.

Decision Tree

Decision Tree as name suggest is a tree like structure where each node is a condition applied on the feature and each branch signifies the outcome and the leaves of decision tree denotes the classes.

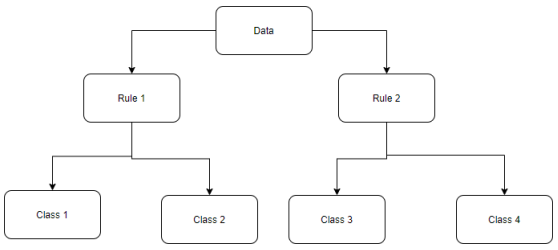

Figure 3: Decision tree

Support Vector Machine(SVM)

Support vector machine is part of supervised learning which try to predict the target variable by analyzing the training data it was provided. In support vector machine a data point is viewed as ndimensional vector and we try to separate the data by using best n-1 dimensional hyper plane which divides the data points in such a way that the distance from it to the nearest data point on each side

is maximized.

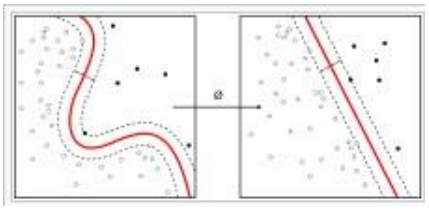

Figure 4:SVM

Conclusion:

Using the proposed mechanism we are able to predict or detect the presence of any disease and provide timely treatment to patient. This mechanism can prove to be the difference in patient's life and death. This system benefits from the evolution of data science and will get better and better as the digital data grows.

References

1. FranckOhlhorst,January2013'BigDataAnalytics:TurningBigDataintoBigMoney',ISBN:978-1-118-14759-7, pp176
2. SamsonOluwaseun,F.,Serdar,S.,andVanduhe,V.,(2014),"Advancingbigdataforhumanitarianneeds", Procedia Engineering, vol.78,N., pp 88-95
3. Amir,G.,Murtaza,H.,(2015),"Beyondthehype: Bigdataconcepts,methods,andanalytics»,Internationaljournal of Information Management, vol .,pp 137-144.
4. H.,Chen,H.L.,Chiang,C.,Storey,(2012),_BUSINESSINTELLIGENCEANDANALYTICS: FROM BIG DATA TO BIGIMPACT', MISQuarterly,Vol.36,No.4, pp.1165-1188.
5. Jonathan Northover, Brian McVeigh,SharatKrishnagiri. Healthcare in the cloud: the opportunity and the challenge. MLD. Available athttp://www.sunquestinfo.com/images/uploads/CMS/445/mlo_02-12014_healthcare_in_the_cloud.pdf
6. Gabriel I. Barbas, Sherry A. Glied, (2010),'' New Technology and Health Care Costs — The Caseof Robot-Assisted Surgery''; the new England journal of medicine , N°, 363, pp 707-704 . Availableathttp://www.nejm.org/doi/full/10.1056/NEJMp1006602
7. MarianthiTheoharidou,NikosTsalis,_'SmartHomeSolutionsforHealthcare:PrivacyinUbiquitous Computing Infrastructures''. Available online athttp://www.cis.aueb.gr/Publications/Smart%20Home%20-%20Site%20TR.pdf
8. SteveG.Peters,JamesD.Buntrock,(2014),''BigDataandtheElectronicHealthRecord'',
9. AmbulatoryCareManage , Vol.37,No.3, pp.206–210
10. R. Weil, (2014),'' Big Data In Health: A New Era For Research And Patient Care Alan R. Weil'',HealthAffair, Vol. 33, N° 7, pp 1110.
11. Peter Groves; Basel Kayyali, (2013),'' The _big data' revolution in healthcare'', McKinsy andCompany.CenterforUSHealthSystemReformBusinessTechnologyOffice.Availableathttp
12. ://digitalstrategy.nl/wp-content/uploads/E2-2013.04-The-big-data-revolution-in-US-health- careAccelerating-value-and-innovation.pdf.
13. T.,Huang,L.,Lan,(2015),_'Promises and Challenges of Big Data Computing in Health Sciences'', Big Data Research vol.2,pp2-11availableathttp://dx.doi.org/10.1016/j.bdr.2015.02.002 [12] Khurshid R., G., Kai, Z., John T., W., and CharlesP., F., (2014), _'Harnessing Big Data for Health Care and Research Are Urologists Ready? ", Journal of European Urology, vol.N.,pp1-3[13] Wullianallur Raghupathi, VijuRaghupathi,(2014),''Bigdata analytics in health care: promise and Potential'', Health Information Science and Systems.
14. Availableathttp://www.biomedcentral.com/content/pdf/2047-2501-2-3.pdf
15. Rashedur M. Rahman, Fazle Rabbi Md. Hasan _Using and comparing different decision treeclassification techniques for mining ICDDR, B Hospital Surveillance data'', Elsevier, Vol. 38, pp11421–11436
16. AndrewKusiak,BradleyDixonb,ShitalShaha,(2005),''Predictingsurvivaltimeforkidneydialysispatients:adataminingapproach'',ElsevierPublication,ComputersinBiologyandMedicine,Vol.35,pp311–327[16]Abhishek,GourSundarMitraThakur,D

eISSN1303-5150

ollyGupta,(2012)

17. ‐Proposing Efficient Neural Network Training Model for Kidney Stone Diagnosis‖, International Journalof ComputerScienceandInformationTechnologies,Vol.3(3), pp3900-3904