ISSN 2278-3091

Volume 9, No.4, July – August 2020

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse298942020.pdf

https://doi.org/10.30534/ijatcse/2020/298942020

Hypertension Prediction using Machine Learning Technique

Youngkeun Choi[1], Jae Won Choi[2]

[1]Division of Business Administration, College of Business, Sangmyung University Seoul, Korea
penking1@smu.ac.kr

[2]Department of Computer Science, Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas, Richardson, TX, USA. jxc190057@utdallas.edu

ABSTRACT

Machine learning technology is used in advanced data analysis and optimization approaches for different kinds of medical problems. Hypertension is complicated, and every year it causes a lot of many severe illnesses such as stroke and heart disease. This study essentially had two primary goals. Firstly, this paper intends to understand the role of variables in hypertension modeling better. Secondly, the study seeks to evaluate the predictive performance of the decision trees. Based on these results, first, age, bmi, and average glucose level influence hypertension significantly, while other variables have on influence. Second, for the full model, the accuracy rate is 0.905, which implies that the error rate is 0.095. Among the patients who were predicted not to have hypertension, the accuracy that would not have hypertension was 90.51%, and the accuracy that had strike was 30.77% among the patients who were predicted to have hypertension.

Key words : Artificial intelligence, Data mining, Decision tree, Hypertension, Machine learning

1. INTRODUCTION

Machine learning and artificial intelligence (AI) are the first choices for data mining and big data [1]. It offers a variety of applications, including technologies such as neural network modeling, simulation models, DNA computing and quantum computing. AI in the bioscience field of 's life sciences greatly reduces random problems when processing this type of data [2]. Many technological advances have helped AI technology evolve in a way that can efficiently and conveniently handle such data. This means that machine learning and AI models are also used in advanced data analysis and optimization approaches, such as drug design and analysis, medical imaging, biological inspiration learning and adaptation to analysis.

A brief description of the function of the machine learning algorithm is learned from previously diagnosed patient cases [3]. To save lives, you need to diagnose heart problems quickly, efficiently and accurately. Researchers have become interested in predicting the risk of disease and creating a variety of health care risk forecasting systems using a variety of machine learning techniques. Missing and outlier data in training sets often degrade model performance and inaccurate predictions. Therefore, it is important to process missing and specific values before making a prediction.

Hypertension is a serious disease which can result in a lot of negative situation [4]. According to the World Health Organization (WHO), high blood pressure causes one in every eight deaths and therefore Hypertension is considered the third leading killer in the world. They resulted in 31% of deaths and currently hypertension affects around 26% of the adult population in the region [5]. The purpose of intervention is to control the heart disease and reduce blood pressure. Early prediction of hypertension is useful for prevention or early treatment intervention. Machine learning is a form of artificial intelligence aimed at building computers with human thinking abilities. The goal of machine learning is to enable the computer to perform certain tasks that rely on patterns and interference without using clear instructions.

The purpose of this study is to find and analyze the causes of hypertension so that doctors and scientists can use this study to formulate possible solutions to problems. The pre-access and modeling methodology used in this white paper can be seen as a roadmap for readers to follow the steps taken in this study and apply procedures to identify the causes of many other medical problems.

2. RELATED STUDY

Medical classification is one of the most important and widely used decision making tools. Many modern technologies have been introduced to accurately and accurately predict hypertension; some work related to this area is briefly described as follows: Previous literature in the field of hypertension prediction in patient populations include statistical models, neural networks and fuzzy models [6].

Echouffo-Tcheugui et al. [7] provides a comparative study and performance summary of various statistical approaches that present a model of hypertension risk. The paper presents 15 high blood pressure prediction risk models obtained from 11 studies reporting on the development, verification and impact analysis of high blood pressure risk prediction models. Major comparison criteria include design and characteristics, predictors, model identification, calibration and reclassification capabilities, verification and impact analysis. Common predictors used in most models are age, gender, BMI, diabetes status and blood pressure. Some of the other variables used are smoking, family history and lack of physical activity; most risk models have acceptable discriminatory abilities (C-statistics 0.7 to 0.8). Some of these tasks focus on specific gender, age, or racial groups and develop models using different data sources. However, none of these models apply a neuronetwork approach that has proven promising in many data domains. To better relate to the ANN approach presented in this white paper, we describe some of the existing ANN approaches below.

Samant and Rao [8] developed the Levenberg-Marquadt inverse propagation neural network, which consists of 13 input nodes and one output node to predict hypertension in Matlab. The input factors consisted of: blood pressure, serum protein, albumin, hematocrit, cholesterol, triglyceride, and hemorrhagic parameters. The author also evaluated performance differences based on hidden nodes and hierarchical numbers to determine optimal performance. They concluded that the deep network with 20 nodes in the first hidden layer and 5 nodes in the second hidden layer brings the best accuracy. The author reported that he achieved 92.85 % accuracy in an approach using a rich data set collected over a decade at the Institute of Hemorrhagic Studies at the Indian Bombay (IITB) Institute of Technology (IITB) Hospital in Mumbai, India. The data set consisted of 13 clinical, biochemical and hemorrhagic data metrics for patients with hypertension and non-hypertension.

Tour et al. [9] compares the performance of three decision tree models, four statistical algorithm models and two ANN models, all predicting the risk of this fetal hypertension disease. Predictors used in the model include: age, gender, family history, smoking habits, lipoproteins, triglycerides, uric acid, cholesterol, and BMI. Based on the sensitivity and specificity analysis of the model, this study infers that the metric used is a good predictor for hypertension diagnosis and that ANN is the best model with gradual learning capabilities to complement existing statistical models.

3. METHODOLOGY

3.1 Dataset

The corresponding variable information is drawn from a third-party website, international challenge on the popular internet platform Kaggle (www.kaggle.com), which provides data in the title of ‘healthcare data’ that was uploaded by Saumya Agarwal. It contains the data of 43,400 patients and 9 attributes, including the predicted attribute. To help with algorithmic development, the organizers provided the types of a data stream for a large set of individual factors. These variables are listed and defined in Table 1.

Table 1: The variables in each category

Variables	Definition
Id	Patient ID
Gender	Gender of Patient
Age	Age of patient
Hypertension	0 - no hypertension, 1 - suffering from hypertension
Ever_married	Yes/No
Work_type	Type of occupation
Residence_type	Area type of residence (Urban/Rural)
Avg_gluscose_level	Average of Glucose level (measured after meal)
Bmi	Body mass index
Smoking_status	Patient’s smoking status

3.2 Dataset

Among the various analytical techniques, decision tree (DT) is a powerful and widely used machine learning algorithm to predict and classify medical data to date [10]. Used for both classification and regression issues. Now you may have questions about why you want to use DT classifiers rather than other classifiers. There are two reasons to answer that question. One is that because decision trees often try to imitate the way the human brain thinks, understanding data and making good conclusions or interpretations is very simple. The second reason is that the decision tree allows you to see the logic that the data interpret, not the black box algorithms such as SVMs and NNs. It has simple and clear expertise and has become one of the favorites among the programmers of this generation. Now we've looked at why we can look at the decision tree in more detail at what the decision tree categorizer is. The decision tree start is a tree with multiple nodes, each node represents a function (properties), each link represents a decision called a rule, and each leaf in the tree represents a different known result. The idea of category type or duration is to create a tree for the entire data and get results from all riffs. Now we know a little bit more about the decision tree. We will continue to discuss how to create a decision tree separator. A decision tree can be built with two algorithms. One is Cart (classification and regression tree) and the other is ID3.

For ID3, first use the x and y values in the column. This value remains in the last position in the column and only has a "YES" or "NO" value. The above chart has x values (view, temperature, humidity, wind) and only two options 'YES' or 'NO' are at the end of the column or are y values. You must now map x and y. As you can see, this is a binary classification problem, so I'm going to use the ID3 algorithm to build the tree. To create a tree, we must first select the root node to be the root node. A general rule of thumb is to first select the function that most affects the y-value as the root node. Select the next most influential function as the next node. Here we're going to use the entropy concept. Entropy concepts measure the degree of uncertainty in a data set. Entropy must be calculated for all category type values of binary classification problems. In summary, the entropy of the data set should be calculated first. For all properties / functions, first calculate the entropy for all category type values, then import the average value information entropy for the current property to calculate the amount obtained for the current property. Then you must select the highest gain property and repeat it until you get the tree you want. This is the process of ID3.

As discussed above, decision tree classifiers were written by a different algorithm called Cart that represents classification and regression trees. This algorithm uses Gini Index as a cost function used to evaluate segmented anger in a data set. Here, the target variable is actually a binary variable, so you use two values (Yes and No). As we all know, there can be four combinations. Now you need to understand the Gini score to get a good idea of how to divide the data. If the Gini score is 0, the worst-case scenario is 50/50 split, but it is a complete separation. The problem is now how to calculate the Gini index value.

The Gini index is similar even if the target variable is a category type variable at a different level. Therefore, the step in this method is the first calculation of the Gini index for a data set. Then you must calculate the Gini index for all category type values for all functions, and then import the average information entropy for the current property and eventually calculate the Gini gain. When you complete this task, you can select the best Gini gain properties and repeat them until you get the tree you want. This is how the decision tree algorithm works.

DT classification methods include creating a tree model that consists of a set of predictors [11]. These predictors (properties) within the training set are divided repeatedly until a pure subset is obtained. This repetitive, personal split process is affected by the characteristics of a particular entity (e.g., customer). The basic structure of the DT consists of a leaf node and a decision node. The leaf node represents the predictor and the point at which the binary division occurs. Leaf nodes are also known as internal nodes. A crystal node, also known as a terminal node, represents an output variable (a binary result variable) and is graphically displayed at the end of the branch. In most cases, terminal nodes based on the Exit Forecast Report category. According to existing literature, 1) Classification and Regression Tree (CART) 2) C4.5 3) Chi squared Automatic Interaction Detection (CHAID) and 4) C5.0 are commonly used. DT is the basis of other tree methods, such as random forest and ensemble forest, and by default multiple decision trees must be aggregated.

The process of binarying a property must select the correct attribute to split. The correct property selection depends on either entropy measurement (C4.5) calculation or GART reference (CART) selection [12] depending on the type of DT algorithm. DT analysis is very famous for its simplicity, graphical layout and ease of interpretation. DT provides an appropriate circuit diagram to model quantitative and positive sexual decision problems without the need to create dummy variables or transformations. DT can also monitor and calculate nonlinear gender (Höper et al., 2017). However, there are several disadvantages to DT. DT results are not always as predictable as in other ways. Moreover, some changes in data sets can lead to unexpected predictions. However, this classification technique has often been used to model deviations [13].

3.3 Data mining models

In order to survive in a competitive market, many companies use data mining technology for decision-making predictive analysis. Building an effective and accurate decision-predictive model is more important than managing your customers effectively. Statistical and data mining techniques were used to construct decision forecasting models. Data mining technology can be used to predict or classify behavior by discovering interesting patterns or relationships in data and fitting models based on available data [13]. If the learning data set and the test data set are separated for machine learning, the test data set must meet the following requirements. It first needs to create a learning data set and a test data set in the same format. Second, the test data set should not be included in the training data set. Third, the learning data set and the test data set must be consistent with the data. However, it is very difficult to create a test data set that meets these requirements. Various verification frame works have been developed in data mining using one data set to address this problem. This study supports using the Split Validation operator provided by RapidMiner. The operator divides the input data set into training and test data sets to support performance evaluation. This study selects relative segmentation among the segmentation method parameters of this operator and uses 70 percent of the input data as learning data.

3.4 Performance evaluation

Performance assessments determine how well a model created using learning data works. Performance measurements can be divided into technical performance measurements and heuristic measurements. The technical performance measurements used in this study produce a model from training data, process the test data as a model, and compare the class labels of the original verification case to the predicted class labels to show performance results. Technical performance measurements can be divided into learning and learning and non-study. The learning used in this study is also classified and reversed. All data used for this learning and testing will have the original class values. Obtain performance by comparing and analyzing the original class values with the predicted results.

Classification issues are the most common data analysis issues. Various metrics have been developed to measure the performance of classification models. Classification problems of category types are often characterized by accuracy, precision, recall, and f measurements. RapidMiner includes performance (classification) that measures performance metrics for common classification problems and performance (differential classification) that provides performance metrics for binary classification problems. Table 2 shows how these metrics are calculated.

Table 2: Key performance indicators

		Actual class (as determined by Gold Standard)	
		True	False
Predicted class	Positive	True Positive	False Positive(Type □ error)
	Negative	False Negative(Type □ error)	True Negative

Precision = TP/(TP+FP), Recall = TP/(TP+FN), True negative rate = TN/(TN+FP), Accuracy = (TP+TN)/(TP+TN+FP+FN), F-measure = 2·((precision·recall)/(precision + recall))

4. RESULTS

4.1 Decision tree

Figure 1 shows the classification tree for the full model after pruning the tree using cross-validation to avoid overfitting. The key variables in the full model analysis consist of 9 ones, as shown below, based on the criterion established with each of these variables. In other words, the classifier has identified four potential questions along each of these variables and specific criteria as defined below to aid in the classification of unknown patients. Age, bmi, and average glucose level influence hypertension significantly, while other variables have on influence.

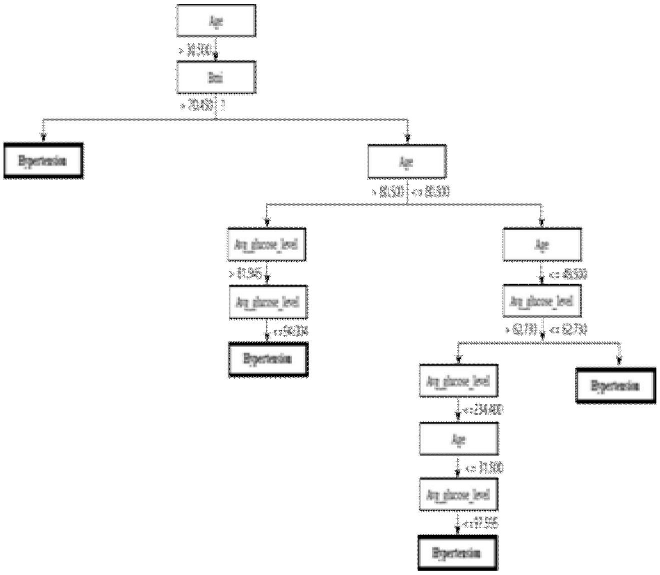

Figure 1: Classification Tree for the Full Model

Tables 3 illustrate each of the confusion matrix measures. For the full model, the accuracy rate is 0.905, which implies that the error rate is 0.095. Among the patients who were predicted not to have hypertension, the accuracy that would not have hypertension was 90.51%, and the accuracy that had strike was 30.77% among the patients who were predicted to have hypertension.

Table 3: Performance evaluation

	True 0	True 1	Class precision
Pred. 0	11,772	1,235	90.51%
Pred. 1	9	4	30.77%
Class recall	99.92%	0.32%	

5. CONCLUSIONS

The results of this study provide additional information about the patient's profile. Doctors should predict hypertensions for the classification methods used. For example, first, age, bmi, and average glucose level influence hypertension significantly, while other variables have on influence. Second, for the full model, the accuracy rate is 0.905, which implies that the error rate is 0.095. Among the patients who were predicted not to have hypertension, the accuracy that would not have hypertension was 90.51%, and the accuracy that had strike was 30.77% among the patients who were predicted to have hypertension.

This study provides some research contributions and actual contributions. First, the study expands existing literature by experimenting with the combined effects of variables on hypertension modeling. Hypertension has a great effect on the patient. A lot of research has been done on hypertension, but no one can say that we can create a universal human tool to predict hypertension. Hypertension is so complex and associated with so many factors that researchers tend to use fewer elements and ignore the effects of others. Patient demographics are often changed and monitored continuously, which can cause hospital problems and damage to personal information. Some studies examined age, gender, and geographic location. But researchers still cannot express cultural and behavioral factors that can affect hypertension. This study contributes to the literature on hypertensions by providing a global model that summarizes the hypertension determinants of individual factors in patients. Second, the methodology used in this white paper can be seen as a roadmap to follow the steps taken by the reader in this case study and apply a day-long procedure to identify the causes of many other problems. This paper proposes the best performance model for predicting hypertension based on a limited set of functions, including patient factors. Achieve the best results in terms of accuracy using machine learning techniques and functional importance analysis, including decision trees and neural networks. In this way, the study identified hypertension patterns that could predict a patient's hypertension.

In fact, this application helps doctors manage patient health records and speed up treatment if you already have patient reports. Quick treatment saves lives. This application helps patients track their health records. Therefore, it is helpful to take care of your health regularly. Analyst reports help doctors easily and easily predict hypertensions. The proposed system has a database that stores patient records, and as the number of patients increases, more data is generated and storage is a problem. Therefore, in future releases, we will provide the cloud capability to store all records in the cloud. So if you have the right to protect and access your data, you can search everywhere. Smart devices are synchronized with applications in future releases. This monitors the patient's real-time health and alerts you in case of an emergency. This reduces risk.

In the future, the machine learning model will use a larger training data set that uses more than one million different data points maintained in electronic health recording systems. While calculations and software sophistication can be a big leap forward, a system that works with artificial intelligence allows doctors to provide the best care to the patient as soon as possible. Software APIs can be developed for free access to health websites and apps to patients. Probability predictions are made regardless of whether processing is delayed or not.

REFERENCES

1. B. Karthikeyan, G. Sujith, V. S. Harsha, K. G. Pavan, Y. M. Sai. **Breast Cancer Detection Using Machine Learning**, *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 2, pp. 981-984, 2020. https://doi.org/10.30534/ijatcse/2020/12922020
2. L. J. Spina, and J. D. Spina. ***Looking Ahead: Artificial Intelligence (AI) and Developing Talent, Harnessing Change to Develop Talent and Beat the Competition***, Emerald Publishing Limited, 2020, pp. 159-162.
3. S. Kaparthi, and D. Bumblauskas. **Designing predictive maintenance systems using decision tree-based machine learning techniques**, *International Journal of Quality & Reliability Management*, vol. 37, no. 4, pp. 659-686, 2020. https://doi.org/10.1108/IJQRM-04-2019-0131
4. D. Zhao, Q. Zhang, and F. Ma. **Communication that changes lives: an exploratory research on a Chinese online hypertension community**, *Library Hi Tech*, Vol. ahead-of-print No. ahead-of-print. https://doi.org/10.1108/LHT-08-2019-0172, 2020.
5. H. C. Kung, and J. Xu. **Hypertension-related Mortality in the United States, 2000-2013,** *NCHS data brief*. Vol. 193, pp. 1-8, 2015.
6. L. Nannia, S. Ghidoni, and S. Brahnam. **Ensemble of convolutional neural networks for bioimage classification**, *Applied Computing and Informatics*, Vol. ahead-of-print No. ahead-of-print. https://doi.org/10. 1016/j.aci.2018.06.002, 2020.
7. J. Echouffo-Tcheugui, G. Batty, M, Kivima, and A. Kengne. **Risk Models to Predict Hypertension: A Systematic Review,** *PLoS ONE*, vol. 8, no. 7, pp. 1-10, 2013. https://doi.org/10.1371/journal.pone.0067370
8. R. Samant, and S. Rao. **Evaluation of Artificial Neural Networks in Prediction of Essential Hypertension**, *International Journal of Computer Applications*, vol. 81, no. 12, pp. 34-38, 2013. https://doi.org/10.5120/14067-2331
9. M.Ture, I. Kurt, A. T. Kurum, and K. Ozdamar. **Comparing classification techniques for predicting essential hypertension**. *Expert Systems with Applications*, vol. 29, no. 3, pp. 583-585, 2005.
10. S. Mishra and D. Bag. **Deriving successful venture capital deal profile through decision tree analysis in Indian context, World Journal of Entrepreneurship**, *Management and Sustainable Development*, vol. 16, no. 2, pp. 97-108, 2020. https://doi.org/10.1108/WJEMSD-03-2018-0031
11. S. Abba, and A. Girsang. **Optimization of Debtor Credit Quality Determining Prediction using Decision Tree**, *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 2, pp. 1076-1081, 2020. https://doi.org/10.30534/ijatcse/2020/27922020
12. W. Verbeke, K. Dejaeger, D. Martens, J. Hur, and B. Baesens. **New insights into churn prediction in the telecommunication sector: A profit driven data**

mining approach, *European Journal of Operational Research*, vol. 218, no. 1, pp. 211-229, 2012.

13. A. De Caigny, K. Coussement, and K. W. De Bock. **A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees**, *European Journal of Operational Research*, vol. 269, no. 2, pp. 760-772, 2018.