HEALTHCARE ANALYTICS

PREDICTION OF STROKE DISEASE

NAME: PRAVEEN KUMAR MARICHAMY

STUDENT NO: 0663980

PROGRAM: AMOD (BIG-DATA ANALYTICS)

Problem Statement:

Healthcare Analytics plays an important role in everyone's life and its being recognized as an upcoming trend in the field of applying analytics with an estimated revenue crossing a billion dollars in the future. Our application on analytics in healthcare will be based on predicting whether a patient will have stroke or not with the data available in the health care industry. The data will have various parameters or variables along with several rows that are relevant to the patient's history. This information will give us a better understanding of the patients along with their past medical health or history. In simple words, various prediction algorithms are to be used on top of any healthcare dataset specially to predict the stroke and the possibilities of getting the disease.

Background research:

There are tons of sites and forums available in the internet providing quality dataset especially for the healthcare domain. In our research, we have contacted several healthcare websites in providing the access to the data for our research. As this

information are related to the patient's personal information and being classified some rules and regulations were to be followed in getting those data. First and foremost, I have decided to narrow down my field of interest in predicting the possibility of getting stroke in patients. There are also several possibilities of predicting other types of diseases like lung cancer detection, heart disease prediction etc.,

1. **Stroke prediction using electronic health records:**

Throughout my research work I have studied and read several articles, various publications, past research works on the same topic. Among those publications one publication which was based on the clinical records of Japanese people. According to the [1] author Douglas Teoh in his article Towards stroke prediction using electronic health records in Japan one of the major causes for the death is stroke as of 2014. The models that were developed according to him were part of a national decision support system. He made his approach in such a way that it will predict the possibility of stroke in a patient who may or may not have stroke in the past. The data were collected in a population with a span of 32 years in total.

2. **Involvement of Artificial Intelligence:**

In several articles the prediction of stroke in patients were determined by the application of artificial intelligence and various machine learning techniques. These AI systems were trained according to the clinical or medical records, healthcare devices, laboratory studies etc., These will help the researchers in finding the hidden patterns and information's. The application of technologies in prediction is not limited to the data set in the excel or csv files. The advancement of technologies in the real world will help us to find the patterns

with the help of the images also. These images include the CT (computerized tomography) scan image, MRI (Magnetic resonance imaging) Image. These images were accessed by any artificial intelligence related predication algorithm. The most important advantage of using these types of technologies will help the people to understand the prediction pattern without the knowledge of the clinical records or the data.

3. Medical Analytics Startup's:

Neuroview, a medical technology Startup is already involved in predicting the stroke in a person. There are in the verge of creating and implementing new technologies in the medical research with the help of big data and data science technologies. The main component in showing the symptom of stock in a person is by their facial expression. Their team is building a standard solution for taking quality pictures by using standard video instead of 3D camera.

iRapid, this particular Startup is specialized in finding the neurovascular conditions in patients. Their platform was used by more than 1200 hospitals and clinics around the world. The application is made in such a way that it will help the doctors to make any medical decisions, building a gap between the hospitals and the specialists, patient transfers to the proper hospital on time.

4. Stroke Statistics:

[2] Stroke kills more than 140,000 Americans each year on which 87 percentage of strokes were based on ischemic stocks which is based on block in the blood flow. It costs more than 34 Million dollars which includes health care services, treatment for the stroke, medicines for the treatment of the stroke and several other factors as well. Some statistics

revealed that this deadly disease will be based on the race also. Risk of having stroke for black race is twice for the blacks when compared with the white race.

5. **Functional limitations and survival following stroke:**

According to the article [3] on the limitations by Johnston M, Pollard B, Morrison V, MacWalter R on Functional limitations and survival following stroke: psychological and clinical predictors of 3-year outcome. Some of the variable's that were taken into account in the study were demographic related, clinical based and psychological based. The multiple regression on their data explained 16 to 40 percent variations on clinical and demographic in terms of recovery.

6. **Application of Deep Learning:**

As discussed earlier the medical images can give better perception and analyses of the data set presented. Sometimes these images were difficult to interpret because of the various background features that includes different colors, patterns, values and most importantly shapes. According to the paper predication of [4] cardiovascular risk factors from retinal fundus photograph via deep learning written by Michael v. McConnell, Greg S. Corrado, Lily, Avinash V. Varadharajan. They trained the model with the help of 284,335 patients in total. They trained the model with the help of deep-learning techniques that used anatomical features such as optic disc, blood vessels to make the prediction.

[5] Three people Songhee Cheon 1, Jungyoon Kim 2 and Jihye Lim 3 devised a method for the use of deep learning to predict the patient mortatlity. This paper is based on the population of Korean people. The paper indicated that the major cause for the strock

related deaths were happened due to the economic burden in the society. This method uses medical imaging to analyse the patients health conditions. They have collected the data of 15,099 patients across the nation with stroke. Principal component analysis was done to take the components of the image accordingly from the clinical records or from the image got from the CT scan , MRI image. The model accuracy at the end was given 83 percentage. This was more than sufficient to tell the patients will have a stroke or not and the model was accepted by the patients and the doctors.

As Korean population is rapidly ageing, and several Koreans are at an age more than 60 stroke problem is predominant in their population. It was published that stroke is responsible for the second leading cause of death. The patients have to stay longer in the hospitals and therefore the cost for the treatment is also high and not affordable for many people. For eliminating the problem, the introduction of artificial intelligence and machine learning came into picture. The patient's records were stored in the form of electronic health record (EHR) databases and utilized to predict the disease.

Proposed approach:

While modelling the data most of prediction problem were solved using the logistic regression technique. The method of solving the problem with the regression was proposed when the data is in either csv format or the excel format. In many cases as discussed in the earlier approach the data was given in the form of images where the application of image segmentation, component analysis was done to extract the information from the data. In our research we are going to implement with a better or another prediction method to solve the same problem. This includes the application of

Random Forest algorithm method in predicting the probability of getting stroke disease. We can assure that the accuracy rate will get increased as well the prediction rate.

This helps us to eliminate the flaws in the older method of predicting the disease and testing. In terms of missing values, the older method eliminating or deleting the entire rows was used. In our research we are focussing on using various regression techniques in filling the missing values and also taken care of outliers in the data. In some cases, we can fill the data with the help of various statistical methods also. This helps us to build a model with higher accuracy rates than the previous ones.

Dataset Description:

There are lot of websites and forums providing the data set to explore and to make useful analyses. Some data set might be lacking in the proper values in other words we can say it as missing values. So, finding a proper dataset will save our time and the model accuracy will be better. In our research we have used Kaggle.com for the stroke prediction dataset. It was originally based on [6] **Mckinsey data set of healthcare.** The number of variables i.e. the number of factors is 12 in number that includes id, gender, age, hypertension 0, hypertension 1, ever married, work type, residence type, avg_glucose_level, bmi, smoking status, stroke. It contains 43400 patients records respectively and the possibility of prediction for approximately 18601 patients. Some of the BMI and smoking _status was missing for few patients. In terms of percentage it was 3.31 in BMI and in smoking status it was 30.71 percentage.

Proposed analyses/evaluation:

The dataset is perfectly suitable for logistic regression in predicting the stroke in a patient and in addition to that it is well suitable for other models as well. First of all, we need to ensure the data is in good quality and suitable for further analyses. First and foremost, technique is to clean and balancing the data in terms of missing value. Data mining is the technique which is used to clean the data before the training data enters in to modelling. The data which we have collected is of raw data and it should be processed. Once the data is processed through various data cleansing techniques the data is ready for EDA (Exploratory Data Analysis). In exploratory data analysis we can plot the variables and factors in terms of graphs, plots etc., Various graphs used in this EDA are Bar charts, pie charts, doughnut charts, line charts, area charts, tree map charts, bridge charts, scatter plots, histogram plots. The better the visualization the better the interpretation from the variables plotted. With the help of the above plots we can find the relationship between the variables whether it is positively correlated or negatively correlated. This makes us to come to a conclusion what are all the variables is going to impact the model.

Skills to be learned:

Lot of tools are readily available in the world to explore the data. The commonly used tools are R and Python. For doing the statistical analysis R is very much useful in finding the strength and the relationship between the variables. This platform can also be useful in plotting various graphs and most commonly used tool for data visualization. The second most important tool is python. When deploying the model and for continuous monitoring of the system deployed with the real-world data python plays a major role in most of the industries. Both the kernels are nowadays available in the cloud platform too.

Cited references:

[1] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6278134/

[2] [https://www.cdc.gov/stroke/facts.htm

[3] https://www.ncbi.nlm.nih.gov/pubmed/15657018

[4]https://www.researchgate.net/publication/319415855_Predicting_Cardiovascular_Risk_Factors_from_Retinal_Fundus_Photographs_using_Deep_Learning

[5] https://www.sciencedirect.com/science/article/pii/S0169260716314705

https://www.mdpi.com/journal/ijerph (The Use of Deep Learning to Predict Stroke Patient Mortality)

[6] https://www.kaggle.com/asaumya/healthcare-problem-prediction-stroke-patients