***Brain Stroke Prediction using Models on the Algorithms***

***Abstract:***

Stroke has become a severe problem among the world. Many researches have been operated in hope of finding a cure for this problem. A lot of developed work has been published regarding this as well as their success isn't negligible. But all those hard worker's contribution isn’t satisfactory enough to nail the goal of identifying this disease. So we did serious research over the nature of these diseases and find out the unique symptoms and went through a lot of testing about the algorithms which can be implemented for bringing out the desired result. And hopefully our work is paid off. We have achieved 98 up percentage accuracy in total and in some specific Algorithm our accuracy is near 100% where others were stick to 90 percentage around.

Rakib Hasan

rakib15-823@diu.edu.bd

[monad.wizard.r@gmail.com](mailto:monad.wizard.r@gmail.com)

***1.Introduction:***

Stroke has been in the top most cause for the death of humankind and created long-time handicap among the adult citizen that reached the number like almost 15 million people every year (**[WHO](http://www.strokecenter.org/patients/about-stroke/stroke-statistics/" \l ":~:targetText=U.S. Statistics&targetText=More than 140%2C000 people die,and 185%2C000 are recurrent attacks.)**). As the population is rising , this number will enhance. According to its nature and visualizing the reasons behind it , strokes are being categorized **into three kinds : Hemorrhagic Stroke, Ischemic Stroke,** TIA(transient ischemic attack). Generally all of this are connected to each other. Survey says the 80% of the stroke patient are from the ischemic victim. Merely It needed to be said the current technology we have and use to follow to facilitate the process to find out the Stroke Disease, is complicated and found out to be a delayed process. But by the blessing of the recent data scientists and programmer , we have come to be familiarized with the way of detecting strokes with their research some of them reached a tremendous position of success. So we have head up to reaching the goal of detecting the strokes with 98 percent up accuracy in order to benefit the human kind.

What we did to reach this , in order to realize that we need to have a clear understanding of the problems. Let’s talk about the nature of strokes, strokes or lesion in brain area created because of the unusual blood circulation. When a specific area lacks of blood , the cells in those areas dies creating the symptoms of ischemic stroke. Lacking occurs due to the block in the blood vessel by the lipid proteins which produces excessively in our body , that’s why most the patient have unusual excessive cholesterol in their body. On the other hand hemorrhagic stroke are occurred for different reasons like when a patent have hypertension or if there is an excessive pressure in the blood vessel of the brain, some time these vessels become sensitive and tore down resulting the blood dropped to the brain which will make a [clot](https://www.healthline.com/health/how-to-tell-if-you-have-a-blood-clot). So the normal activity of brain cell will be interfered as the brain’s every domain have some special tasks to perform. So the concerning region’s activity cant be executed resulting in many problems. So these two problems have identical symptoms we have tracked down those symptoms which signifies the significant domains of the affected portion of the brain. So we use the [classification and regression](https://www.geeksforgeeks.org/ml-classification-vs-regression/) to filter the symptoms of stroke from the other health issue. Our dataset is efficient enough to distinguish among the types of strokes depending over the symptoms.

***2.Literature Review:***

|  |  |  |
| --- | --- | --- |
|  | **Research Summary** | **Our Research** |
| Chantamit-o-pas et al.[1] | They claimed that they developed a system which is more efficient in predicting heart disease than the Medical scoring system which is currently providing services to the patient. Their Analytic research includes such algorithm to perform that. | Our research includes such algorithm which has brought about accuracy more than 95 , which is tremendous and hoping it will also be more efficient than medical scoring system. |
| Review on Heart Disease Prediction System using Data Mining Techniques | Data mining techniques deals with large data set and this feature can cast out result depending over the data s that has been stored, so they collected the algorithm that uses NN to identity the heart issues form a research. They tend to use this and mines the data into system to diagnosis the heart problem | It shares very common things in respect of our research.  We have plans to fuse NLP with NN for detecting symptoms basing over the input from the patient. |
| Asadi et al[2] | stroke is a major threat that disables human-body and encounter deaths. Report says ischemic stroke takes lead among the others, they have researched and gathered datas regarding ischemic stroke. They used SPSSH, MATLABH, and RapidminerH, classical statistics as well as artificial neural network and support vector algorithms to nail their goals identifying stroke paitents. Their effort succeeded to gain 80% accuracy. | Regarding that our dataset consists with real-time data then we will pre process the data , we will use NLP , CNN  To process the patient data and screen out the result. |
| Feng et al.[3] | This report signifies that using DNN with the neurological images or Neuro image dataset can provide a tremendous result in diagnosing the stroke issues. Convolutional Neural network and deep learning techniques has the high advantage in the medical sector | Our work will be more efficient because they are confined in digital Image technology for nailing their goal. We are not only depending over Image data but also over the patient’s physical condition |
| Hung et al[4] | Their work is quiet similar in respect of other'work but the one that devines them from other ia their unique data set of EMCs., where 800,000 paitents data is stored. This Dataset has been processed using ML'S NLP and CNN to bring out a outcome having 90 up accuracy in percentage | This work is unique , need to confess that but we have common unique features with unique algorithms, their dataset is strong so do ours. They are also using the same technique that we are using, but hence their process orders is not organized like ours, we specially maintain a certain flow for processing the data obtained from user which will enables us to detect disease in any cases and might not create bug. |

***3.Methodology:***

3.1. Working flowchart :

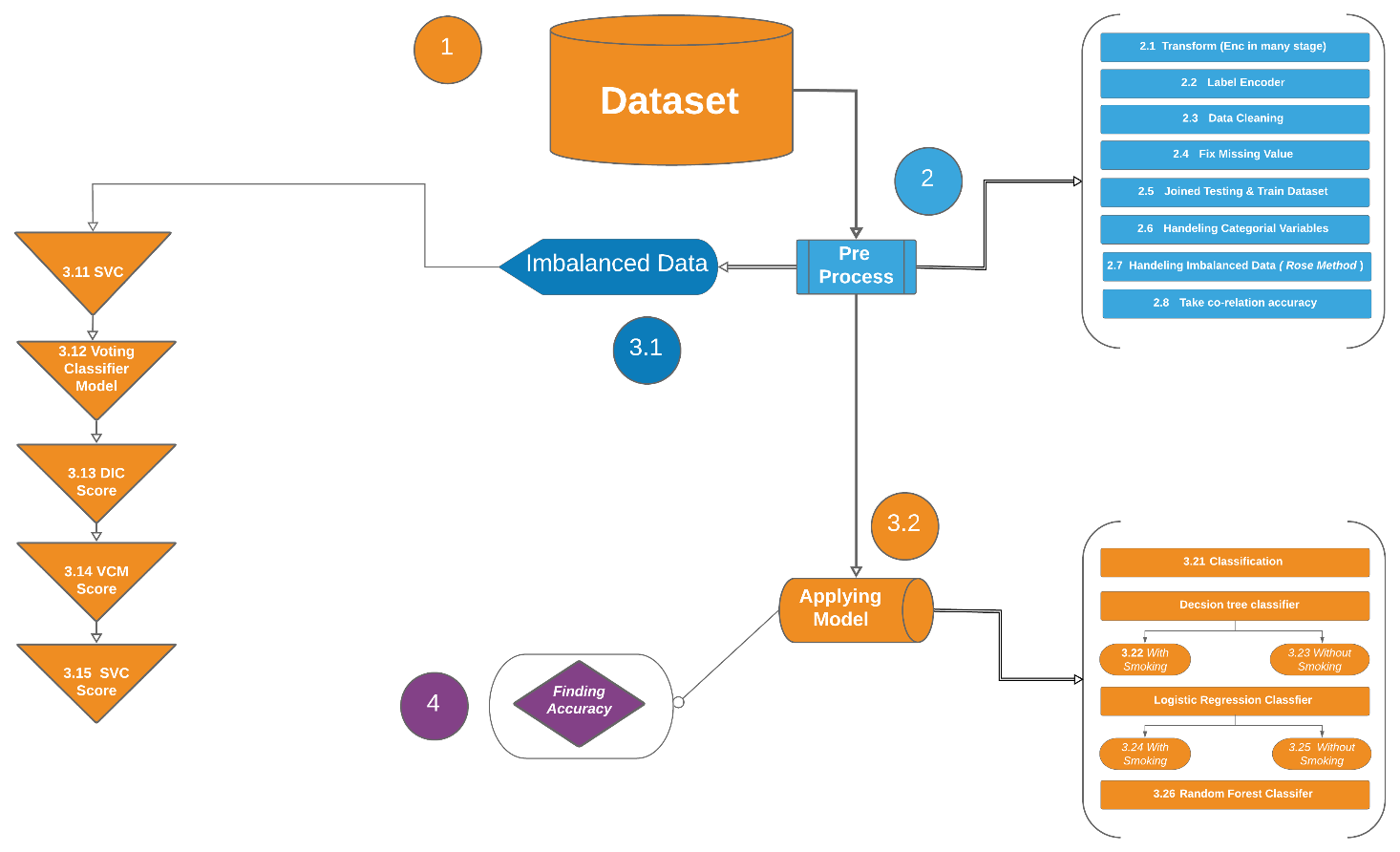


Fig: 3.1 workflow diagram

In the acceding workflow

***3.2. Implementation & Result:***

***3.2.1 Libraries:***

We are work in python utilizing following library :

Pandas, Numpy, Matplotlib, Sklearn.processing(Imputer, LabelEncode),

Sklearn.model\_selection(train\_test\_split), Sklearn.tree(DecisionTreeClassifier), Sklearn.Linear\_mdel(LogisticRegression), Sklearn.Metricis(oc\_curve, Classification\_report , Accuracy\_score, confusion\_matrix, auc, roc\_auc\_score, precision\_score, recall\_scoreensemble, RandomForestClassifier), Imblearn. over\_sampling (RandomOverSample, SMOTE, ADASY)

We use *pandas* for take data and make data frame to process our imbalanced dataset for prediction.

For advance computing we use ***numpy*** for efficient multi-dimensional container of generic data. **Matplotlib** use for visualization our accuracy in graph. **Sklearn** is a python library to work with regression and classification in supervised and unsupervised learning. **processing** is a sub library of Sklearn.  preprocessing bundle gives a few normal utility capacities and transformer classes to change crude element vectors into a portrayal that is progressively reasonable for the downstream estimators. The **Imputer** class provides basic strategies for imputing missing values. And convert to constant value. ***LabelEncode*** used for target labels with value between 0 and n\_class-1. ***Model\_selection*** is another sklearn library which use for model\_selection. Split clusters or networks into arbitrary train and test subsets. Snappy utility that wraps input approval and next(ShuffleSplit().split(X, y)) and application to include information into a solitary call for parting (and alternatively subsampling) information in an one liner. ***DecisionTreeClassifier*** Choice tree fabricates order or relapse models as a tree structure. It separates a dataset into littler and littler subsets while simultaneously a related choice tree is steadily created. ***LogisticRegression*** In insights, the strategic model (or logit model) is utilized to demonstrate the likelihood of a specific class or occasion existing, for example, pass/fall flat, win/lose, alive/dead or sound/wiped out. The ***sklearn.metrics*** module executes a few misfortune, score, and utility capacities to quantify arrangement execution.

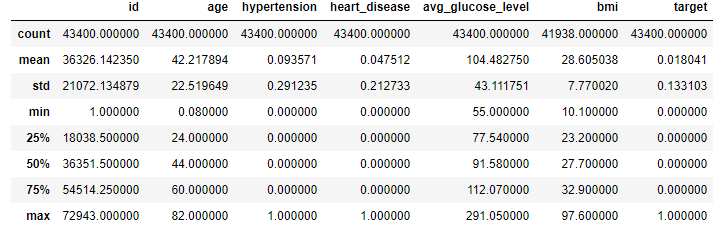
sklearn.metrics.**classification\_report**(y\_true, y\_pred, labels=None, target\_names=None, sample\_weight=None, digits=2, output\_dict=False, zero\_division='warn')

***imblearn*** over\_sampling . Destroyed. ... Examining data to resample the informational index. At the point when glide , it relates to the ideal proportion of the quantity of tests in the minority class over the quantity of tests in the greater part class subsequent to resampling.

***3.2.2 Dataset:***

We use (row = 43400, column = 12) data as train data. And ( row = 18601, column = 11) data as test set. It’s a imbalanced dataset.

Fig:3.2.2 describe numerical data columns .



*Fig:3.2.2*

3.2.3**. Data Preprocessing** :

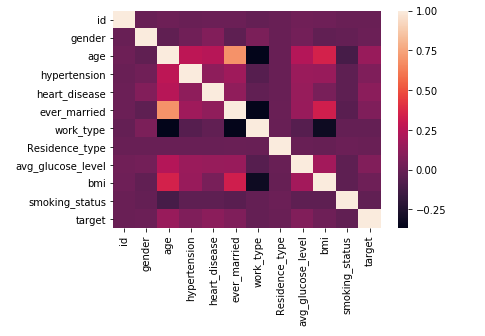
We ***Clean Data*** by 1st fix missing value with some condition and calculation.

If any data is equal to null then make sum of all data then it divided by it’s total length then make multiply by 100. Ex : *if (dataset==null)* then *sum(allData)/length(allData)\*100*

3.2.4. Handling Categorical Data :

We handle categorical Data by **Encode**. Here we use labelEncoder. Label Encoding refers to converting the labels into numeric form to convert it into the machine-readable form. LabelEncoder encode labels with a value between 0 and n\_classes-1 where n is the number of distinct labels. If a label repeats it assigns the same value to as assigned earlier.

*After categorical Data Correlation heatmap is:*



**3.2.5 Decision tree:**

For predict Decision tree

1st we fit our train and test data in DecisionTreeClassifier(). Which is pre processed.

2nd define test data to dtree.predict(). Which predict our tree

3rd  define accuracy by (TP + TN)/(TP+TN+FP+FN).

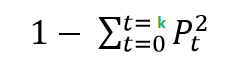
4th define recall by (TP)/(TP+FN)

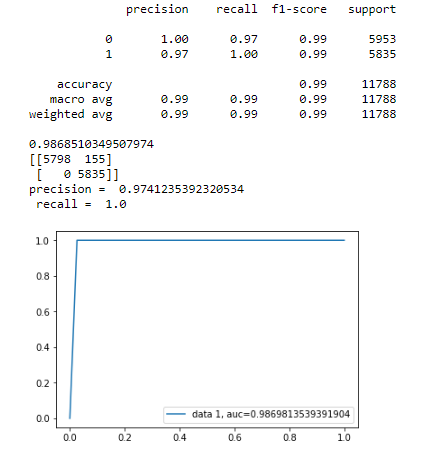
5th define precision by (TP)/(TP+FP)

Here TP = True Positive , TN = True Negative, FP = False Positive, FN = False Negative

**3.2.5.1 : *Decision tree with smoke status:***

After complete 3.2.5 we define columns amount , we use all columns to compute prediction . target= 1 — P^2(Target=0) — P^2(Target=1)

here k = 10



***Fig:3.2.5.1***

***Decision tree with smoke***

{10=smoking\_status} 10 is the last columns. so we need to define all columns .

Here we see: accuracy 0.98686

Precision 0.9741

Recall 1.0

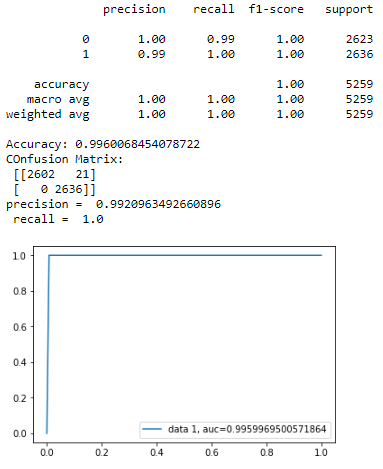
*Confusion matrix:*

[[5798 155]

[ 0 5835]]

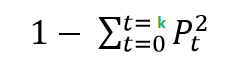
**3.2.5.2 : *Decision tree without smoke status:***

After complete 3.2.5 we define columns amount , we use all columns to compute prediction . target= 1 — P^2(Target=0) — P^2(Target=1)



***Fig:3.2.5.1***

***Decision tree without smoke:***

here k = 9

Here we see: accuracy 0.9961

Precision 0.9920

Recall 1.0

{10=smoking\_status} 10 is the last columns. so we need to define all columns

*Confusion matrix:*

[[2554 33]

[ 0 2672]]

**3.2.6: Logistic Regression** :

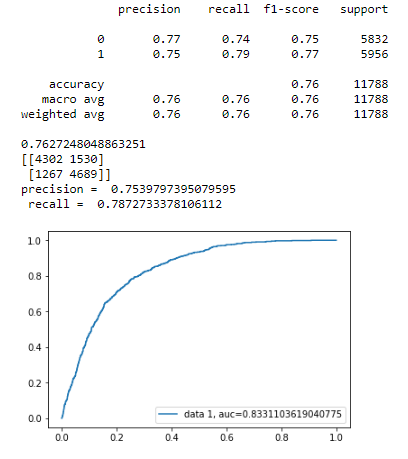


Fig:3.2.6.1

***Logistic Regression with smoke:***

**Logistic regression** is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome.

Where p is the probability of presence of the characteristic

In our implement we done logistic regression with smoke status in **fig:3.2.6.1**

Here we see: accuracy 0.762725

Precision 0.7539

Recall 0.78727

*Confusion matrix:*

[[4299 1654]

[1247 4588]]

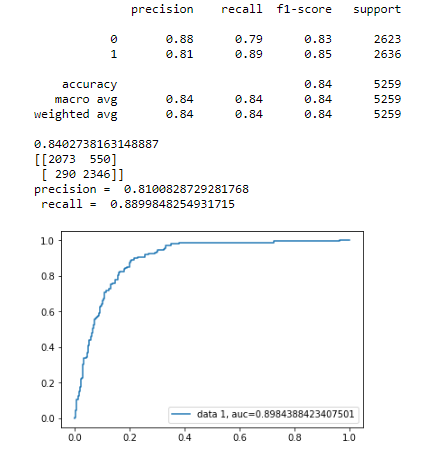


Fig:3.2.6.2

***Logistic Regression without smoke:***

In our implement we done logistic regression with smoke status in **fig:3.2.6.2**

Here we see: accuracy 0.762725

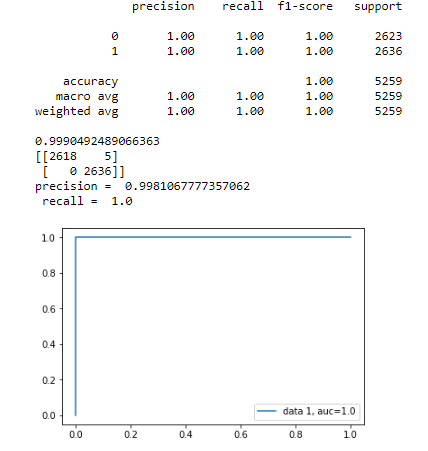
Precision 0.7539

Recall 0.78727

*Confusion matrix:*

[[2043 544]

[ 291 2381]]



***Fig:3.2.7***

***Random Forest Classification With Smoke:***

***3.2.7 Random Forest Classification:***

*In the time of implementing in Random Forest Classification we got maximum accuracy.*

Here we see: accuracy 0.99980

Precision 0.99962

Recall 1.0

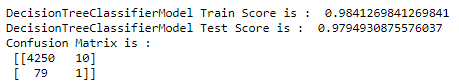
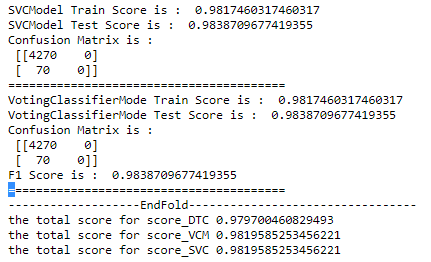
*Confusion matrix* :

[[2586 1]

[ 0 2672]]

*We use all columns to predict accuracy.*

***After Bringing few changes our system will provide least 96 percent accuracy finding out the stroke. This changes includes the implementing the SVC ,DTC,VCM model over the Decision Tree Algorithm.***



***Future Work :***

This is just the beginning of our journey. Our goal is to construct a complete AI Embedded Software that will be able to detect any disease of human body. To attain in that position our path need to acquire some short time goal.

* Building a heath care assistant using NLP and voice recognition
* Ability to interact with patient via voice communication
* Will collect input symptoms from the voice
* Predict disease probability
* Enhancing the capability of our assistant by implementing image processing & CNN (Convolutional Neural Network) adding to the previous facilitation
* Providing the 100 percent accuracy
* Suggesting medicine for the detected health issues
* And also providing the health advisory ex: food diet
* Linking hardware technologies with our assistant so that our machine can optimize patients suffering to undergo clinical tests. The idea is to facilitate the system with ability to take MR & CT scan images and casting a result consisting the patients problem and it’s solution

***Conclusion:***

So we believe in welfare of humanity and this kind of work is an asset for the human race and for those who wanted to work over it in the near future. we have used a number of algorithms which is presented in methodology. And we have also improved or we can say that we have customized Decision tree algorithm using some models which is our tremendous achievement because the accuracy is near 100 % so it can be possible to detect any health problem with this by the help of a valuable dataset.

**References :**

1. Chantamit-o-pas, P., & Goyal, M. (2017). *Prediction of Stroke Using Deep Learning Model. Lecture Notes in Computer Science, 774–781.* doi:10.1007/978-3-319-70139-4\_78
2. Asadi, H., Dowling, R., Yan, B., & Mitchell, P. (2014). *Machine Learning for Outcome Prediction of Acute Ischemic Stroke Post Intra-Arterial Therapy. PLoS ONE, 9(2), e88225.* doi:10.1371/journal.pone.0088225
3. Feng, R., Badgeley, M., Mocco, J., & Oermann, E. K. (2017). *Deep learning guided stroke management: a review of clinical applications. Journal of NeuroInterventional Surgery, 10(4), 358–362.* doi:10.1136/neurintsurg-2017-013355
4. Hung, C.-Y., Chen, W.-C., Lai, P.-T., Lin, C.-H., & Lee, C.-C. (2017). *Comparing deep neural network and other machine learning algorithms for stroke prediction in a large-scale population-based electronic medical claims database. 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC).* doi:10.1109/embc.2017.8037515
5. King, A., & Markus, H. S. (2009). *Doppler Embolic Signals in Cerebrovascular Disease and Prediction of Stroke Risk: A Systematic Review and Meta-Analysis \* Supplemental Online References. Stroke, 40(12), 3711–3717.* doi:10.1161/strokeaha.109.563056
6. Khanna, S. (2013). *Artificial intelligence in health – the three big challenges. Australasian Medical Journal, 6(5), 315–317.* doi:10.4066/amj.2013.1758
7. Lambrou, A., Papadopoulos, H., Kyriacou, E., Pattichis, C. S., Pattichis, M. S., Gammerman, A., & Nicolaides, A. (2010). *Assessment of Stroke Risk Based on Morphological Ultrasound Image Analysis with Conformal Prediction. Artificial Intelligence Applications and Innovations, 146–153.* doi:10.1007/978-3-642-16239-8\_21
8. Sedghi, E., Weber, J. H., Thomo, A., Bibok, M., & Penn, A. M. W. (2015). *Mining clinical text for stroke prediction. Network Modeling Analysis in Health Informatics and Bioinformatics, 4(1).* doi:10.1007/s13721-015-0090-5