GROUP DATA MINING PROJECT

TITLE**: BRAIN STROKE PREDICTION**

TROOP ZERO TEAM

TEAM MEMBERS:

1. Sai Nath Vanacharla
2. Swetha Maddyala
3. Sai Ramya Krishna Maddukuri
4. Srichandu Sai Battala
5. Ravali Venkatayogi

Introduction:

In this project we have a dataset which is analyzed for predicting brain stroke by considering various input features from Kaggle. As per the World Health Organization (WHO), stroke is the second most leading cause of death worldwide, accounts for approximately 11% of all fatalities. Approximately 750,000 people decease due to brain stroke in USA annually it is also #1 leading cause of disability according to WHO. Where 25% of population with initial stroke decease within 1 year and 50-75% will be functionally dependent where 25% will live with permanent disability. A risk factor is a condition or behavior that is more frequent in individuals with an illness or who have a higher chance of developing it than in those without it. Stroke risk factors do not imply that you will have a stroke. However, not posing a risk factor does not mean that you will not have a stroke. However, as the number and severity of risk factors rises, so does your risk of stroke. Some factors for stroke can't be modified by medical treatment or lifestyle changes are age, gender, race, and family history of stroke. But we do have some of the most important treatable risk factors for stroke. They are High blood pressure, or hypertension, Cigarette smoking, Heart disease, Warning signs or history of TIA or stroke, Diabetes, Cholesterol imbalance, Physical inactivity, and obesity.

This dataset contains input parameters like gender, age, various diseases, and smoking status. This dataset is used to predict whether a patient is prone to get a stroke. The data's rows each provide crucial information about the patient.

Problem Specification:

The main aim is to predict the brain stroke of the person based on the input parameters of the that person. The problem is creating the model in such a way that the designed model should predict the true positives and true negatives properly and have very few false negatives.

Project Objective:

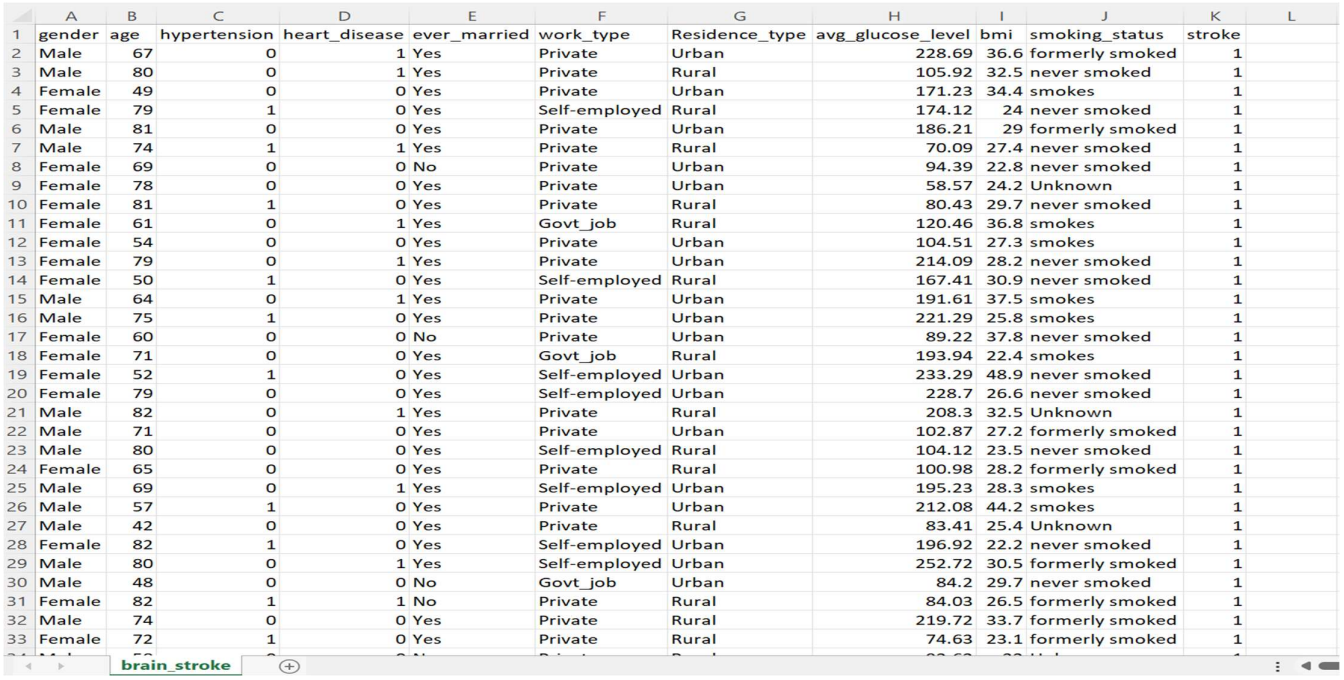
The objective of the project is to build CNN, Decision Tree, SVM, Naive-Bayes models that would predict if a subject had a chance of brain stroke or not and compare the accuracies and recall factors. We will be using accuracy and recall as metrics to justify the model's performance.

Questions Answered:

We aim to reduce the predicted false negatives of the model (False negatives: Model predicting that person does not have brain stroke based on input features, where person has brain stroke) as in the health care analytics it is crucial for a model to predict the correct outcome when the actual result is positive. We utilised this dataset to apply the concepts which we have learned in class, and we have also applied few novelty concepts of cleaning the missing data and scaling the data for better modelling.

So, in this problem, we are analysing the dataset on predicting whether the person had a brain stroke by reducing the false negatives of the model as far as we can.

Data Characteristics:



	A	B	C	D	E	F	G	H	I	J	K	L
1	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke	
2	Male	67	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1	
3	Male	80	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1	
4	Female	49	0	0	Yes	Private	Urban	171.23	34.4	smokes	1	
5	Female	79	1	0	Yes	Self-employed	Rural	174.12	24	never smoked	1	
6	Male	81	0	0	Yes	Private	Urban	186.21	29	formerly smoked	1	
7	Male	74	1	1	Yes	Private	Rural	70.09	27.4	never smoked	1	
8	Female	69	0	0	No	Private	Urban	94.39	22.8	never smoked	1	
9	Female	78	0	0	Yes	Private	Urban	58.57	24.2	Unknown	1	
10	Female	81	1	0	Yes	Private	Rural	80.43	29.7	never smoked	1	
11	Female	61	0	1	Yes	Govt_job	Rural	120.46	36.8	smokes	1	
12	Female	54	0	0	Yes	Private	Urban	104.51	27.3	smokes	1	
13	Female	79	0	1	Yes	Private	Urban	214.09	28.2	never smoked	1	
14	Female	50	1	0	Yes	Self-employed	Rural	167.41	30.9	never smoked	1	
15	Male	64	0	1	Yes	Private	Urban	191.61	37.5	smokes	1	
16	Male	75	1	0	Yes	Private	Urban	221.29	25.8	smokes	1	
17	Female	60	0	0	No	Private	Urban	89.22	37.8	never smoked	1	
18	Female	71	0	0	Yes	Govt_job	Rural	193.94	22.4	smokes	1	
19	Female	52	1	0	Yes	Self-employed	Urban	233.29	48.9	never smoked	1	
20	Female	79	0	0	Yes	Self-employed	Urban	228.7	26.6	never smoked	1	
21	Male	82	0	1	Yes	Private	Rural	208.3	32.5	Unknown	1	
22	Male	71	0	0	Yes	Private	Urban	102.87	27.2	formerly smoked	1	
23	Male	80	0	0	Yes	Self-employed	Rural	104.12	23.5	never smoked	1	
24	Female	65	0	0	Yes	Private	Rural	100.98	28.2	formerly smoked	1	
25	Male	69	0	1	Yes	Self-employed	Urban	195.23	28.3	smokes	1	
26	Male	57	1	0	Yes	Private	Urban	212.08	44.2	smokes	1	
27	Male	42	0	0	Yes	Private	Rural	83.41	25.4	Unknown	1	
28	Female	82	1	0	Yes	Self-employed	Urban	196.92	22.2	never smoked	1	
29	Male	80	0	1	Yes	Self-employed	Urban	252.72	30.5	formerly smoked	1	
30	Male	48	0	0	No	Govt_job	Urban	84.2	29.7	never smoked	1	
31	Female	82	1	1	No	Private	Rural	84.03	26.5	formerly smoked	1	
32	Male	74	0	0	Yes	Private	Rural	219.72	33.7	formerly smoked	1	
33	Female	72	1	0	Yes	Private	Rural	74.63	23.1	formerly smoked	1	

brain_stroke

The figure shown above is the snippet of the dataset having various input features and dependent variable. The first column is gender, it is having two categories: Male or Female. The second column is age, we observed that this is the crucial input parameter in our analysis. The third and fourth columns are hyper_tension and heart_disease these come under most important treatable risk factors for stroke. We can see that these two parameters have the binary data. The fifth column is ever_married which is a Boolean data. The sixth and seventh columns are work_type with four categories and Residence_type with two categories. The eighth and ninth columns are important factors of a person in predicting the brain stroke, which are avg_glucose_level and BMI (Body Mass Index). The tenth column is smoking_status of the person. The eleventh column is the dependent variable which is Stroke. This has binary data as the output is simple, 0 or 1. 0 means the person does not have brain stroke and 1 means the person has brain stroke.

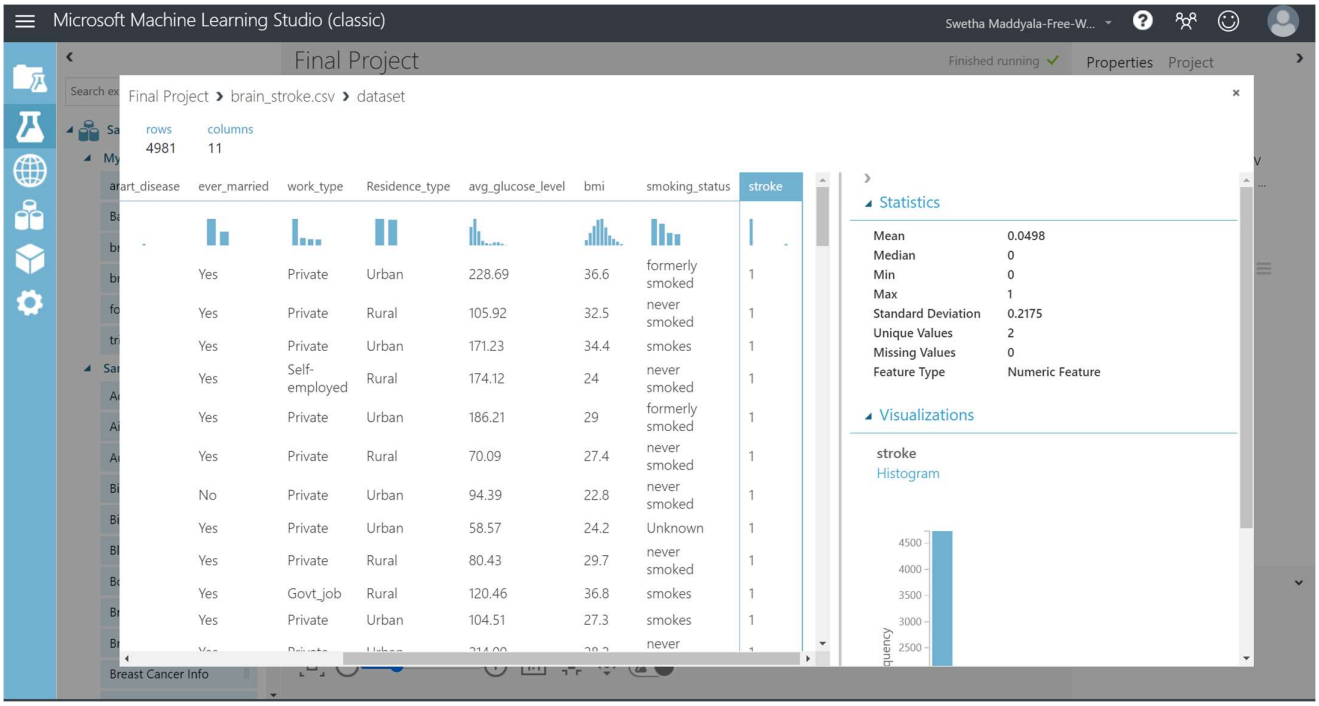
It is predominant that first we clear all the missing data, normalise, and scale the data to get accurate model in predicting the brain stroke of a person based on various input features provided in the dataset.

Independent and Dependent variables:

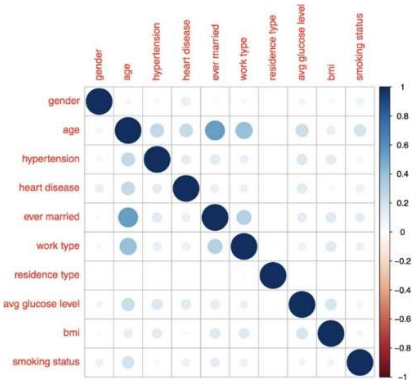
Independent variables: gender, age, hyper_tension, heart_disease, ever_married, work_type, residence_type, avg_glucose_level, bmi, smoking_status

Dependent variable: stroke.

Visualizing the Dataset:



The above figure is the snippet from the Azure ML Studio which shows the data characteristics. It shows that we have 4981 entries and 11 features where stroke being the dependent variable.



The above figure tells us that none of the features are highly correlated with each other. Thus, each feature might have their individual contribution towards stroke prediction.

Splitting Data:

Here, we split the data into two sets. One is Training dataset, and one is Test dataset.

Training Dataset: It is a subset to train the model.

Testing Dataset: It is the subset to test the trained model.

For classification tasks, a supervised learning algorithm looks at the training data set to determine, or learn, the optimal combinations of variables that will generate a good predictive model. The goal is to produce a trained (fitted) model that generalizes well to new, unknown data. So, we have approached this problem with

a 70-30 split. This means that 70% of the data we have will be used for training the model and the remaining 30% will be used for testing the model against the predicted value and the actual value.

We wanted to make sure that the model is trained against all the different cases to make sure that no poor results with a different set of training data.

Data Mining Model Construction:

Data Pre-Processing:

Data pre-processing is nothing but organizing the data and getting in a clear and applicable format for the current scenario.

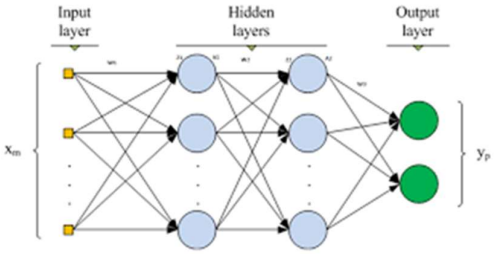
We have scaled, normalized, and then selected the appropriate columns and cleansed the inappropriate values and summarized the data for better understanding.

Model Building:

We have used the following regression algorithms on the data set to see which model best fits for the scenario by evaluating the results.

1. **Two Class Neural Network**

- A neural network is a set of interconnected layers. The inputs are the first layer and are connected to an output layer by an acyclic graph comprised of weighted edges and nodes.
- A two-class neural network contains two output nodes, and all inputs will map to one of the two nodes in the output layer.

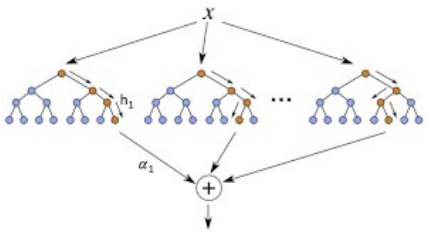


2. **Two Class Bayes Point Machine**

- The Bayes Point Machine is a Bayesian approach to linear classification. It efficiently approximates the theoretically optimal Bayesian average of linear classifiers by choosing one "average" classifier, the Bayes Point.
- This implementation improves on the original algorithm in several ways.
    a. It uses the expectation propagation message-passing algorithm. For more information, see A family of algorithms for approximate Bayesian inference.
    b. It does not require parameter sweeping.
    c. It does not require data to be normalized.

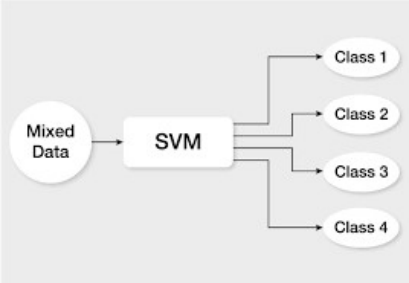
3. **Two Class Boosted Decision Tree**

- A boosted decision tree is an ensemble learning method in which the second tree corrects for the errors of the first tree, the third tree corrects for the errors of the first and second trees, and so forth. Predictions are based on the entire ensemble of trees together that makes the prediction.
- A two-class boosted decision classifies to one of the two outputs.



4. **Two Class Super Vector Machine**

- This SVM model is a supervised learning model that requires labeled data. In the training process, the algorithm analyzes input data and recognizes patterns in a multi-dimensional feature space called the hyperplane.
- All input examples are represented as points in this space and are mapped to output categories in such a way that categories are divided by as wide and clear a gap as possible.

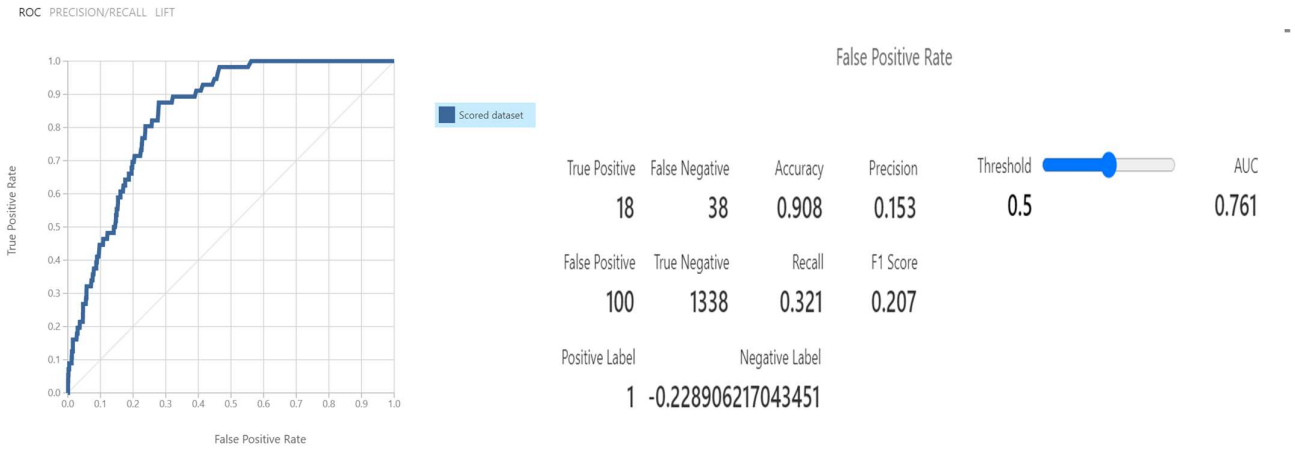


Metrics for Evaluation:

Below are the metrics used for comparing the above-mentioned models and their performances:

True Positive	Positive Label	Accuracy
True Negative	Negative Label	Recall
False Positive	Precision	Threshold
False Negative	AUC	F1 Score

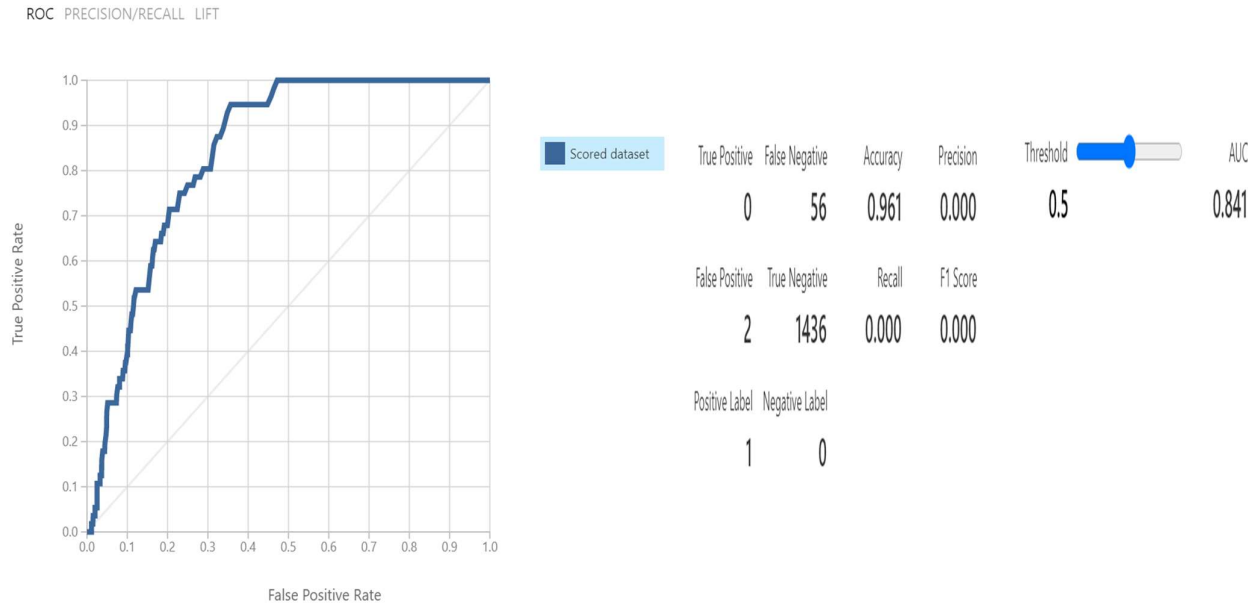
Visualization of Metrics:



Two Class Neural Networks Result

Two Class Bayes Point Machine Result

Two Class Boosted Decision Tree Result

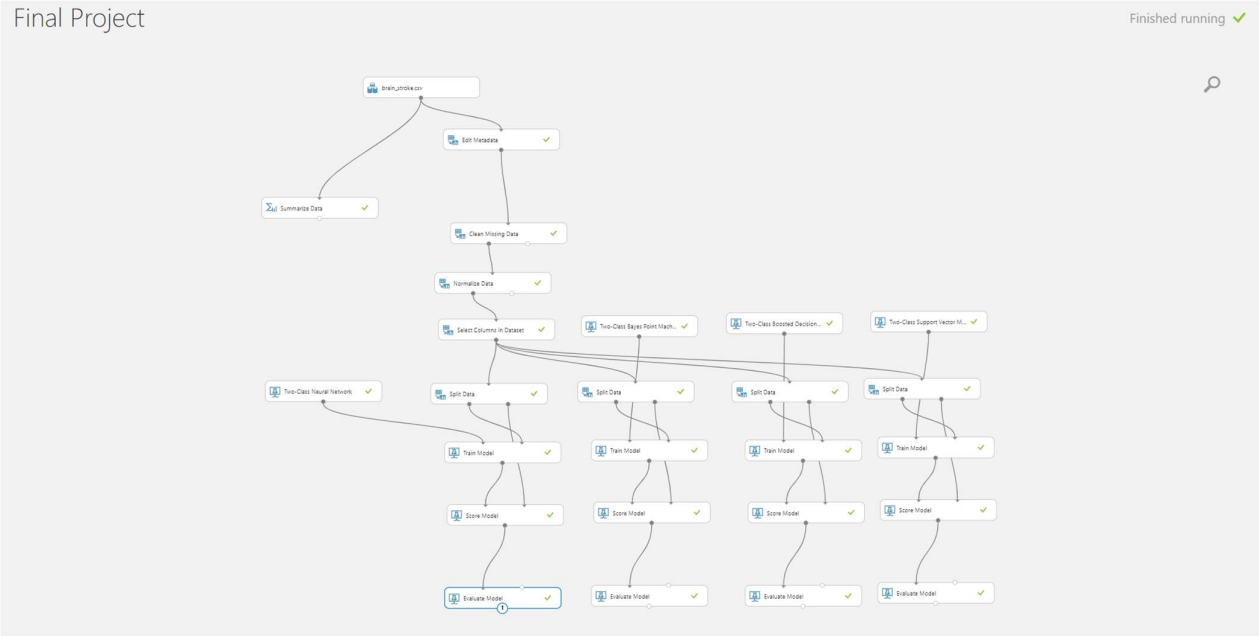


Super Vector Machine Result

Interpretation of Metrics:

- The accuracy is higher for the Two class Bayes Point machine model. But we observe that recall value is completely 0.
- The accuracy of the Two class Neural network is 0.908 but has recall value of 0.321 which is greater than any other model used here for predicting the brain stroke.
- This shows that our model is working fine.

Snippet of Data Mining Models in Azure ML studio:



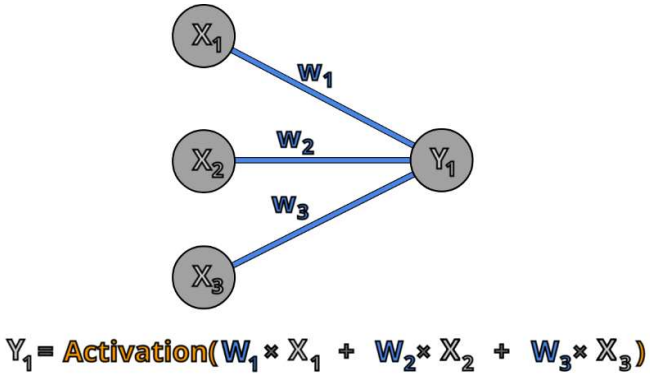
Conclusion:

- In this project, we proposed a system that enables the early detection and prediction of stroke disease based on machine learning techniques.
- This is an important experimental result indicating that the early detection of stroke disease can be accurately predicted using **Two Class Neural Networks** as it is having accuracy of 0.908 and recall factor value of **0.321** which is higher compared to any other models.
- In Health care analytics it is better to have few false negatives. And this model has less false negatives compared to any other models used in the project.

Further Improvement:

Further research and develop systems that provide more reliable and clinically interpretable stroke prediction results by conducting multi modal studies can be built by combining electronic medical recording (EMR) data, such as individual health checkups, with CT or MRI scans and interpretating the information.

Interpretability of Two class Neural Network:



- In the above figure, we can see the formula for Two Class Neural Network model showing the weights and features used for predicting the dependent variable.
- Y is dependent variable,
- W1, W2, W3 are the weights against the features and X1, X2, X3 are the features.

Business Value of this Analytics:

We have the following business value with all the above analytics and techniques that would be beneficial to or appealing to a business stakeholder:

- The predominant value would be predicting whether the person will get affected by brain stroke or not by analyzing various input features incorporated in the design.
- The other important value would be greater recall value with fewer false negatives as predicting a person will not get affected with brain stroke when person gets affected by brain stroke is not acceptable in health care analytics.

References:

1. Choi, Y.-A; Park, S.-J.; Jun, J.-A.; Pyo, C.-S.; Cho, K.-H.; Lee, H.-S.; Yu, J.-H. Deep Learning-Based Stroke Disease Prediction System Using Real-Time Bio Signals. Sensors 2021, 21, 4269. https://doi.org/10.3390/ s21134269
2. G. R. Kumar, P. Vyshnavi, S. Prasanna, T. H. Reddy, C. Charanya, and P. Chandrababu, “Brain Stroke Detection Using Machine Learning”, *IJRESM*, vol. 5, no. 3, pp. 34–36, Mar. 2022.
3. https://www.sciencedirect.com/science/article/pii/S2772442522000090
4. Sivapalan G., Nundy K., Dev S., Cardiff B., Deepu J.ANNet: A lightweight neural network for ECG anomaly detection in IoT edge sensors

   IEEE Transactions on Biomedical Circuits and Systems (2) (2022)