Prediction Of Stroke Based on Real-Time Patient Data

DIC Project Phase 3

Team:

Name	UB ID
Venkata Rohil Wardhan Kancharla	vkanchar
Mahendra Nallabothu	mnallabo

Phase 3 comprises of creating a UI to display whether the patient is at risk of stroke or not based on the given set of input parameters. Visualizations to display various parameters for ur dataset has also been displayed at the bottom of the UI.

Final Model:
Our final model comprises of classification models mainly -

- Logistic Regression
- Random Forest
- Ridge Classifier
- Ridge Regression

We determine whether a patient is at risk of stroke or note based on their patient data. Since this is a prediction model, we use classification which is most desired for prediction. We have used a regression model to show how it performs.

- **Logistic Regression -** We used logistic regression as it predicts a dependent variable from a set of independent variables. Since the output can only be a discrete or categorical value, which is the same as shown in the 'x_test" and 'y_test' of the dataset.
  Justification - We have achieved an accuracy of 75% and a precision of 0.12 and recall of 0.78. A precision of 0.12 means that the model is correct only 12% of the time. 0.78 of recall indicates a certainty of 78% of all strokes occurring for that particular person. Although accuracy is low, this model is more reliable..

- **Random forest -** Since random forest builds decision trees based on different samples and takes a majority vote, this is useful in stroke prediction as the output is either 0 or 1, and not in between values required like for other datasets. Also,

since random forest performs better for classification, we can achieve higher accuracy.
Justification - We have achieved an accuracy of 96% and a precision score 0 of and recall of 0. This indicates that the the model is predicting everything as opposite to what it should predict.

- **Ridge Classifier -** Ridge classifier converts the target variable to +1 or -1 just like the dataset. This helps in better classification of data. Even though it performs as a regression task after converting the target variables, the dataset is suitable for regression after converting the target variables.
  Justification - We have achieved an accuracy of 96% with a precision of 0 and recall of 0. This indicates that the the model is predicting everything as opposite to what it should predict.

Overall Output Parameters for all the models -

	Accuracy	**Precision**	**Recall**
Logistic Regression	0.73	0.12	0.78
Random Forest	0.96	0.0	0.0
Ridge Classifier	0.96	0.0	0.0

- **Ridge Regression -** For ridge regression, the accuracy, MSE, RMSE, R2 Score and MAE are displayed.

UI for Web Application -

We have used ‘Streamlit’ to build the UI. The UI consists of a sidebar to input different parameters to obtain the prediction based on the input parameters. The sidebar also shows the 4 different models to choose from, in order to obtain the evaluation matrix of which accuracy, precision and recall are displayed for the classification models. Confusion Matrix, ROC curve and Precision-Recall Curve are the visualizations displayed for all the classification models.

- **Confusion Matrix -** Confusion matrix comprises 4 values displayed in the form of a 2x2 matrix. The (1,1) value is the ‘true positive’, (2,1) is ‘False Negatives’, (1,2) is ‘False Positive’ and (2,2) is ‘True Negatives’. The higher number of false positives shows the positive or correct output obtained.

- **ROC Curve -** ROC curve shows the classification of the model. The x-label is 'False Positive Rate' and the y-label is 'True Positive Rate'. For a perfect classifier, the true positive rate is at 1 and False positive is at 0.
- **Precision-Recall Curve -** A precision-recall curve plots the output between precision and recall. If the area under the curve is high, that means the precision and recall of the model is high. High precision means low false positive rate and high recall means low false negative rate.

Visualizations for the Models -

Logistic Regression -

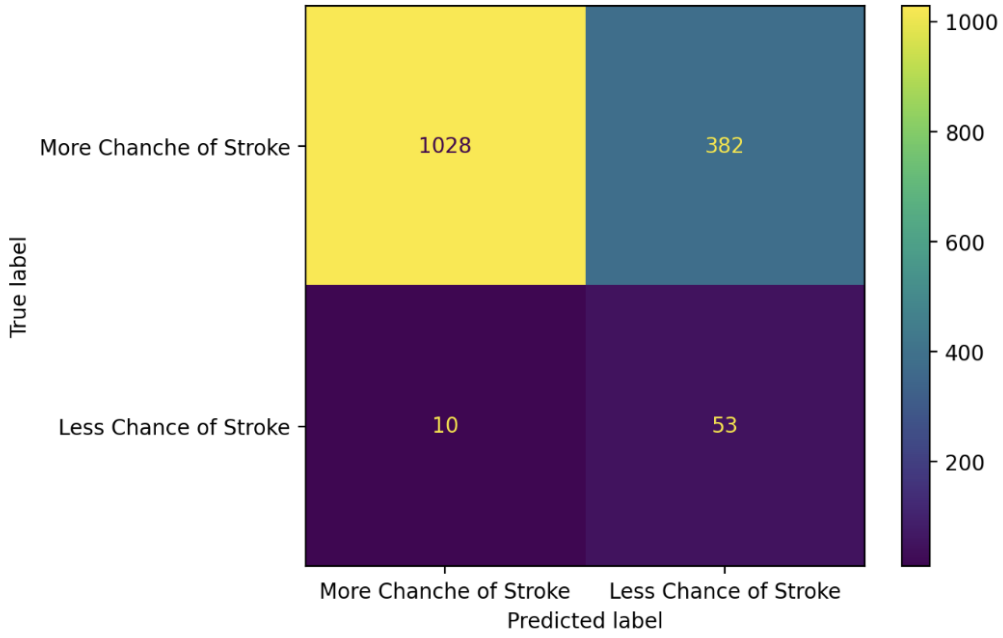
Logistic Regression

Accuracy `0.73`

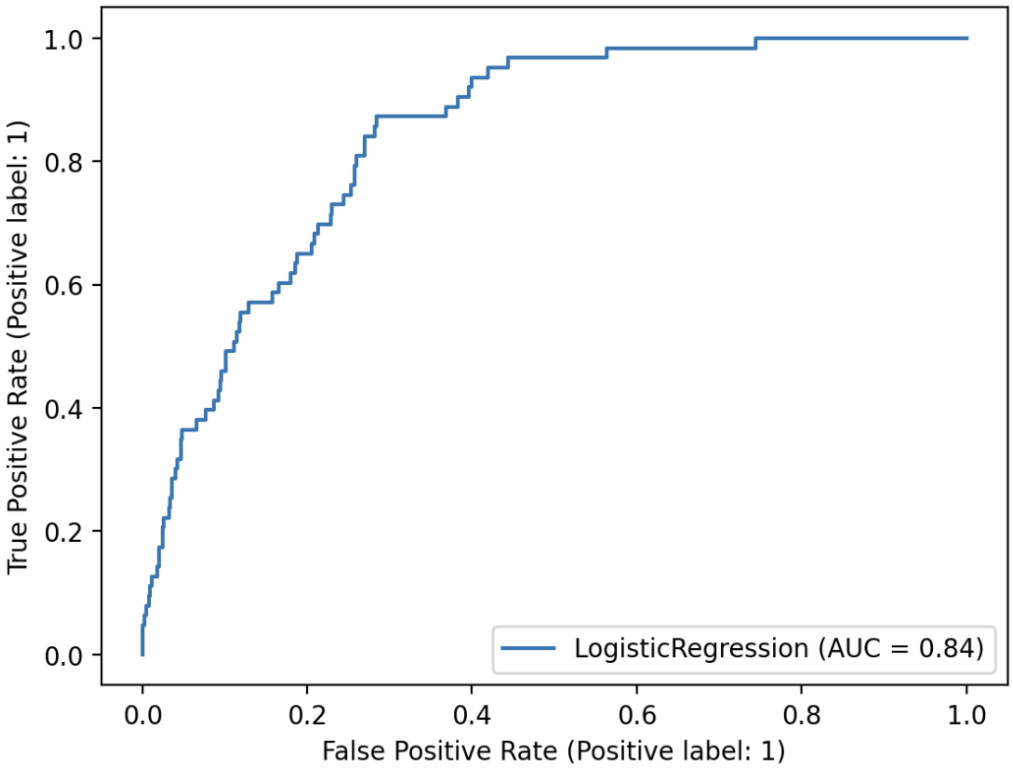
Precision: `0.12`

Recall: `0.84`

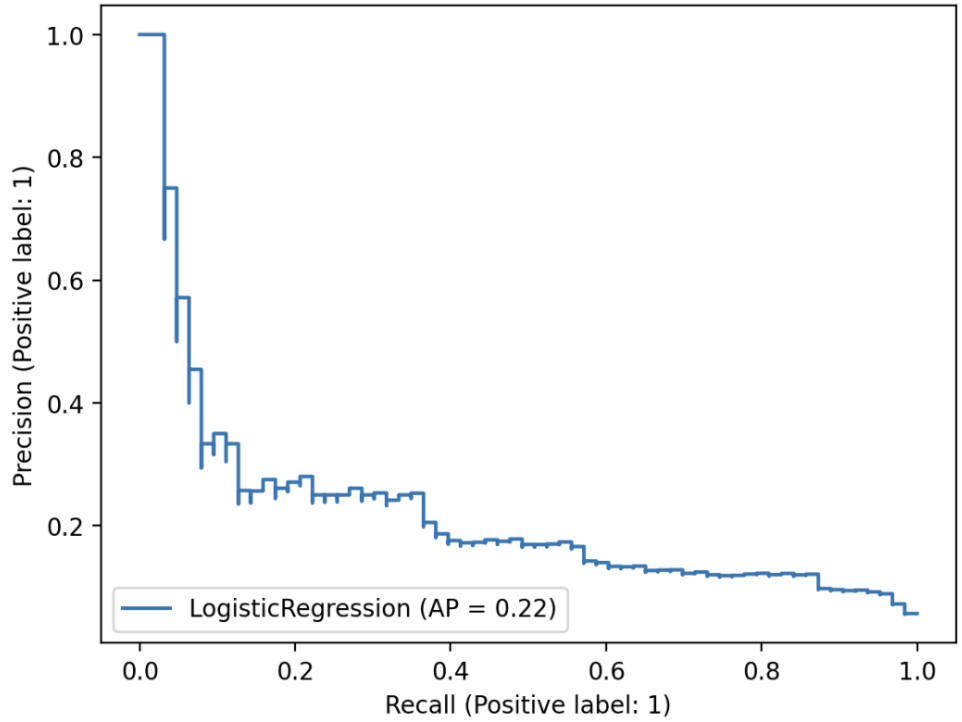
Confusion Matrix -



ROC Curve -



Precision-Recall Curve -



Random Tree -

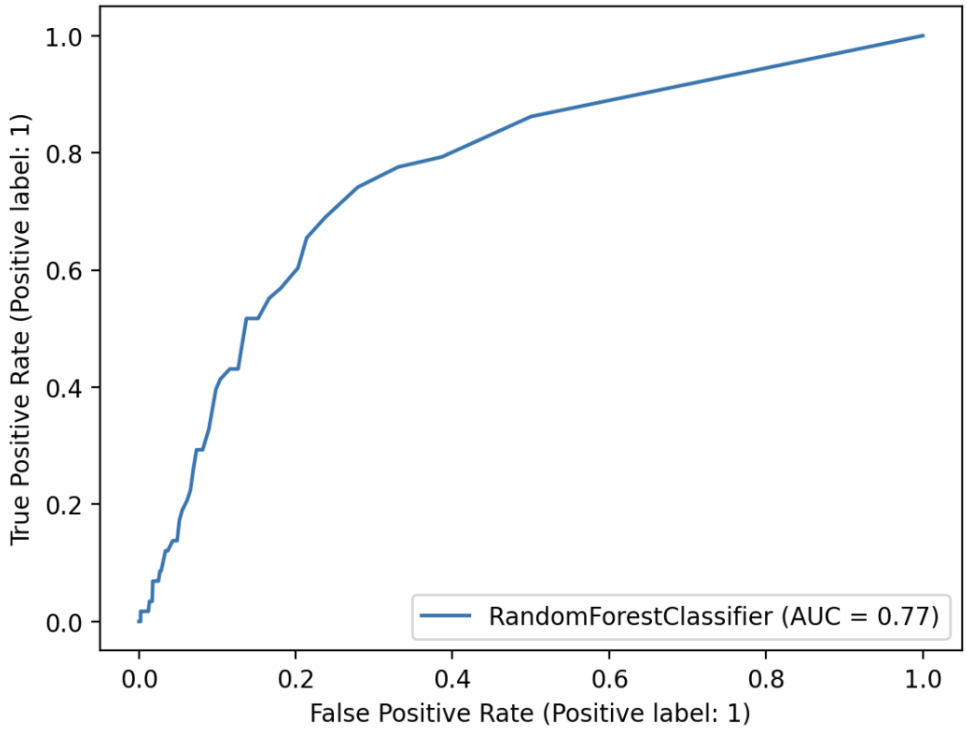
Accuracy `0.96`

Precision: `0.0`

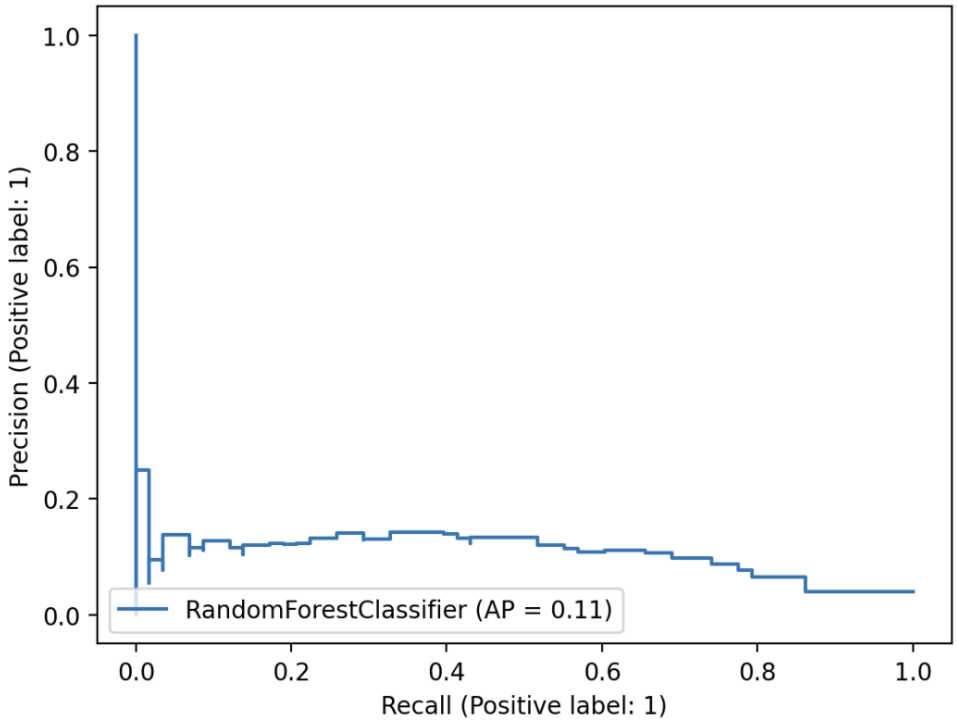
Recall: `0.0`

Confusion Matrix -

ROC Curve-



Precision-Recall Curve-



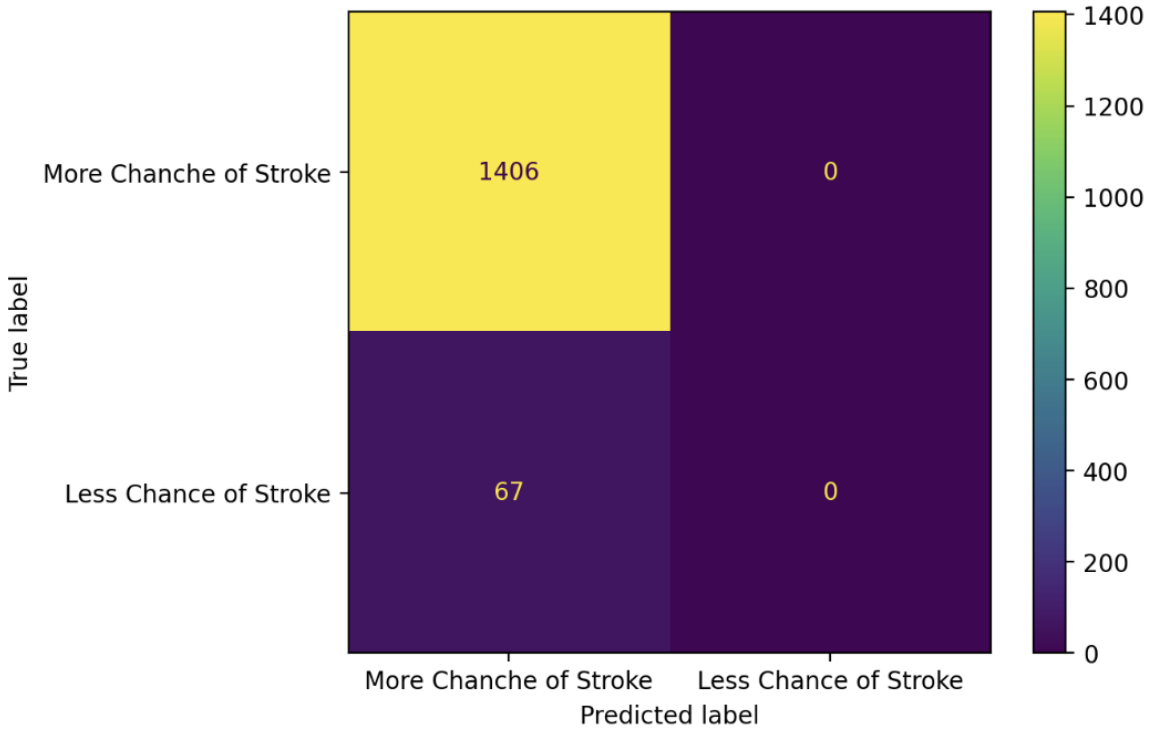
Ridge Classifier -

Accuracy `0.95`

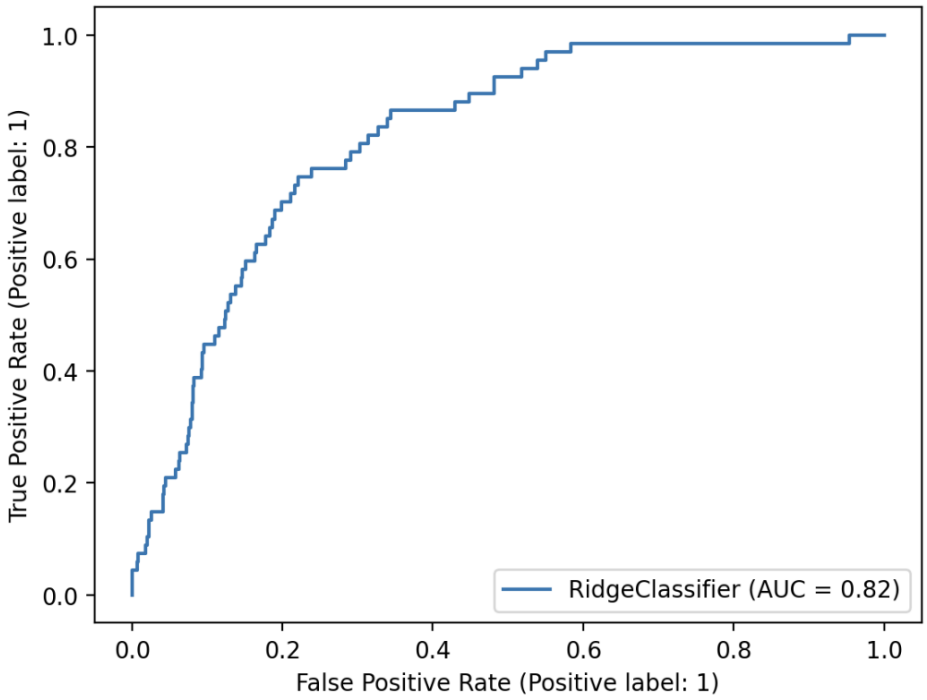
Precision: `0.0`

Recall: `0.0`

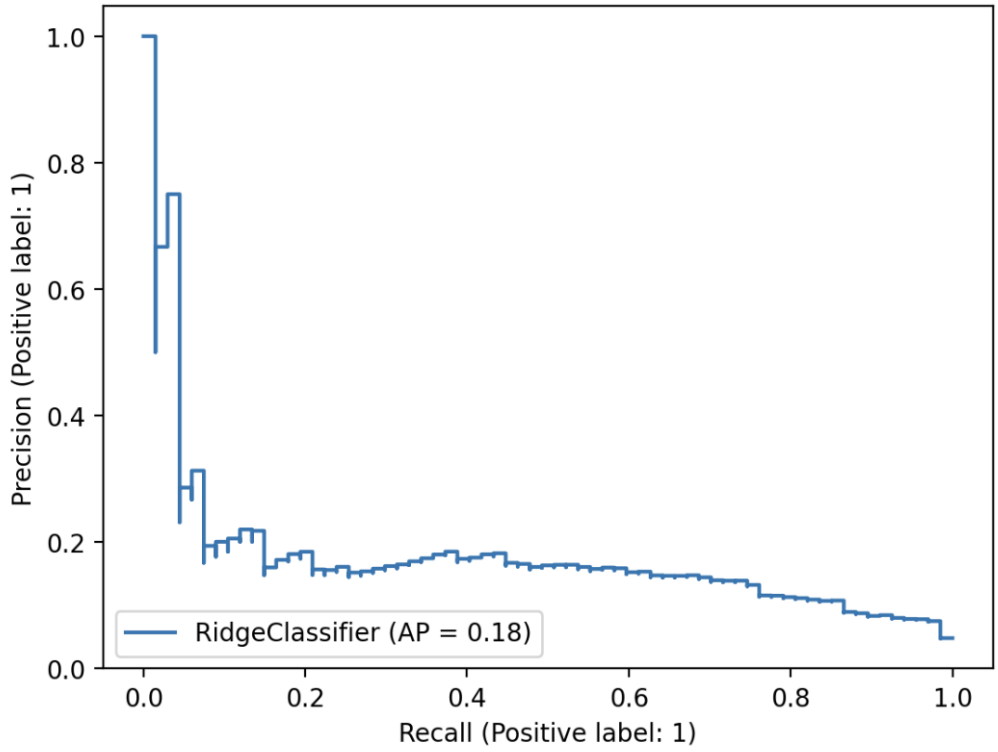
Confusion Matrix -



ROC Curve -



Precision-Recall Curve -



Ridge Regression -

Accuracy `0.06`

MSE: `0.035525919328235386`

R2 Score: `0.060782046812077106`

RMSE Score: `0.18848320701918087`

Mean Absolute Error: `0.08356571141414128`

Instructions to launch the Web application -

1. Open Terminal
2. conda create -n py39 python =3.9.0 (use python 3.9 to avoid any issues)
3. conda activate py39
4. Once activated, !pip install required packages. Required packages present in Readme.txt file.
5. After packages installed, >streamlit run app.py
6. The webpage is opened automatically in browser
7. On the sidebar of the webpage, input desired parameters by selecting the options provided. Some parameters are given in the form of sliders.
8. On the main page, the input parameters specified are shown and the prediction of stroke is shown.
9. If prediction is 1, the patient is risk of stroke. If prediction is 0, the patient is not at a risk of stroke.
10. On the bottom of the webpage, select the type of ML Model to view visualization for that model.
11. For example, select ‘Logistic Regression’ and click on display to display the output metrics along with vizualization graphs.
12. Visuaization Graphs are -
    a. Confusion Matrix
    b. ROC Curve
    c. Precision-Recall Curve

13. Click on ‘Show Raw Data’ to view the data taken for that model. The raw data refreshers randomly everytime the model viewed is changed.

Changes Made From the Previous Phases -

1. Removed the id column in the dataframe.
2. Downloaded a csv file after removing null values and unwanted columns (after preprocessing).
3. Data changed to dataframe from list format for inputs.
4. New models have been trained. Logistic regression and Ridge Classifier have ben designed for visualization.

Analysis and Future Work -

- This model can be used in all hospitals to check the prediction of stroke.
- This helps in deducting the possibility of stroke for the doctor so they can focus on other parameters.
- More different parameters of the patient data can be added for better results in the future.

References -

https://scikit-learn.org/stable/
https://www.analyticsvidhya.com/blog/2021/03/everything-you-need-to-know-about-machine-learning/

https://scikit-learn.org/stable/supervised_learning.html
https://scikit-learn.org/stable/unsupervised_learning.html
https://matplotlib.org/stable/tutorials/introductory/pyplot.html
https://docs.streamlit.io/