Predicting Survival chances for stroke patients 80 years and above

By: Bhavya Balasubramanya

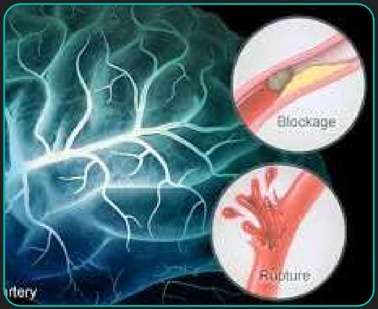
May 2019

For complete project with code, please visit my repository

Outline

- Introduction
- Understanding the Dataset
- Data Cleaning
- Exploratory Data Analysis
- Unsupervised Modelling
- Supervised modelling for patients 80 years and above
- Best model for patients 80 years and above
- Conclusions
- Next Steps
- Credits
- Thank You

Introduction

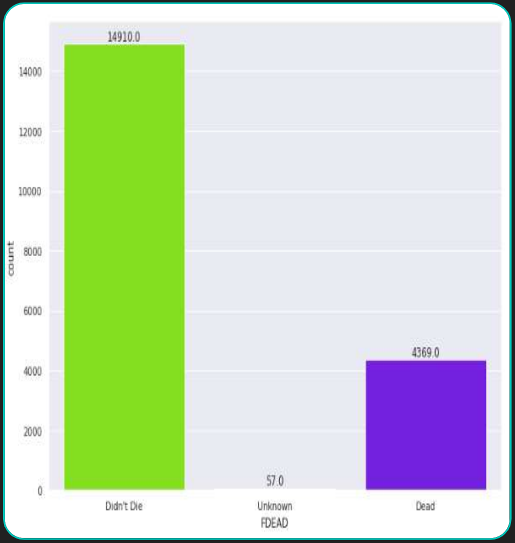


Face drooping
Arm weakness
Speech difficult
Time to call 91
911

According to WHO, 15 million people suffer from stroke every year. Of these 5 million die and other 5 million are permanently disabled. In, U.S., stroke is the 3rd leading cause of death. More than 140,000 people die due to stroke and it is also a leading cause of serious and long-disability in US. On an average someone in US has a stroke every 40 seconds. Risk of having stroke more than doubles each decade after age of 55. It can occur at any age and nearly 1/4th occur in people under age of 65. Risk factors increase with preconditions like diabetes, high blood pressure, high cholesterol, Atrial fibrillation. However, out of these, high blood pressure is the most common factor for stroke. In developed countries, incidence of stroke is declining, largely due to efforts to lower blood pressure and reduced smoking. However, overall rate of stroke remains high due to aging of the population.

Understanding the dataset

- This dataset has been taken from the International Stroke trial database (version 2) which was one of the biggest randomized trials in actual stroke.
- The data was collected worldwide in 43 different countries.
- There were a total of 19,432 records and 112 attributes.
- The data was recorded from 1991 to 1996.
- Data from 1991 to 1993 was considered pilot phase. We haven't considered this phase in our capstone since there was lot of missing data during this period. So for future reference our data is only from the period March 1993 – April 1996
- Target variable chosen for modelling purposes is 'FDEAD' which indicates if the person is alive or not at sixth month follow up
- The initial distribution of target variable is as shown in Fig A.
- Though there appears to be a bias in data, we will be using this data by removing just the unknowns and not really balancing the data as our predictions actually compares survival rate of patients aged 80 and above to those below.



Data Cleaning

10

Remove Unwanted columns

There were a lot of columns in our data which was not necessary. Since this data was collected worldwide, the dates didn't make sense as it was not uniform and didn't tell anything important

19

Remove columns and rows that don't have much data

There were some columns for comments which didn't have significant data to use imputation, so they were removed. Also most of the data was missing during pilot phase, so pilot phase rows were removed

53

Convert categorical data to indicator variables

We used dummy data for Stroke types. While doing this, we also replaced most common categorical data like Can't access (C) to 0, No (N) to 1, Yes (Y) to 2 and Unknown (U) to 3

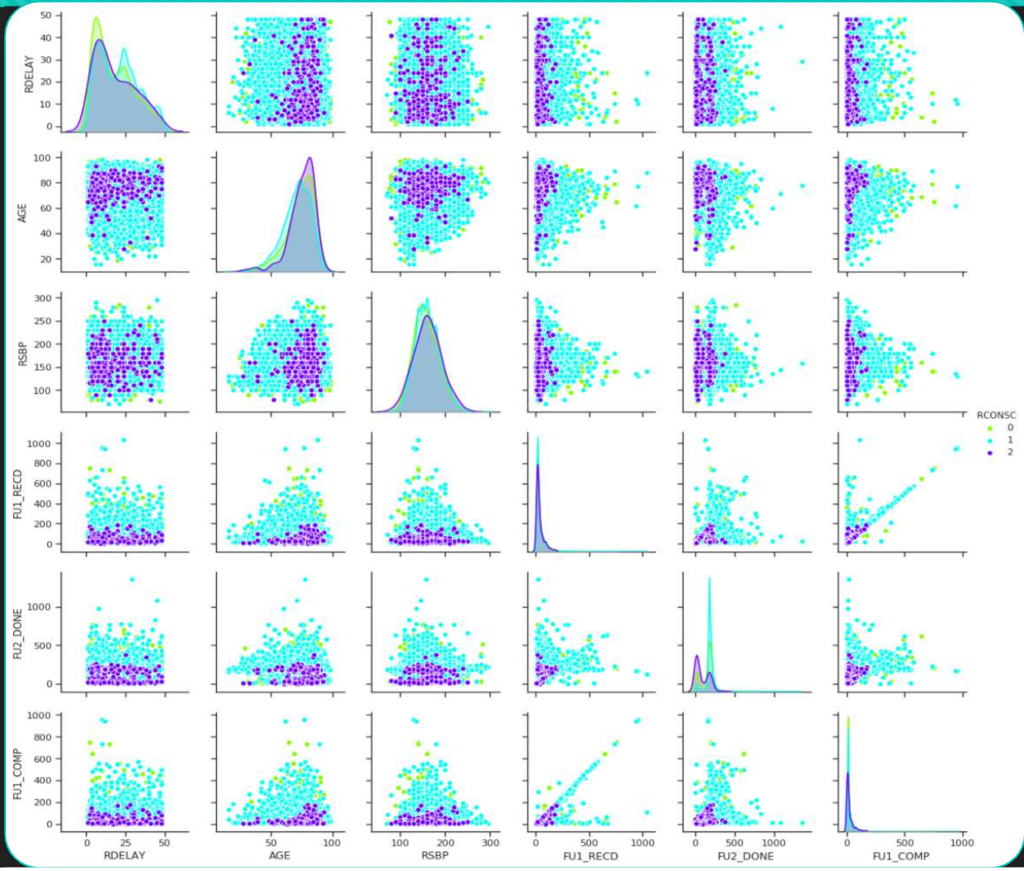
1

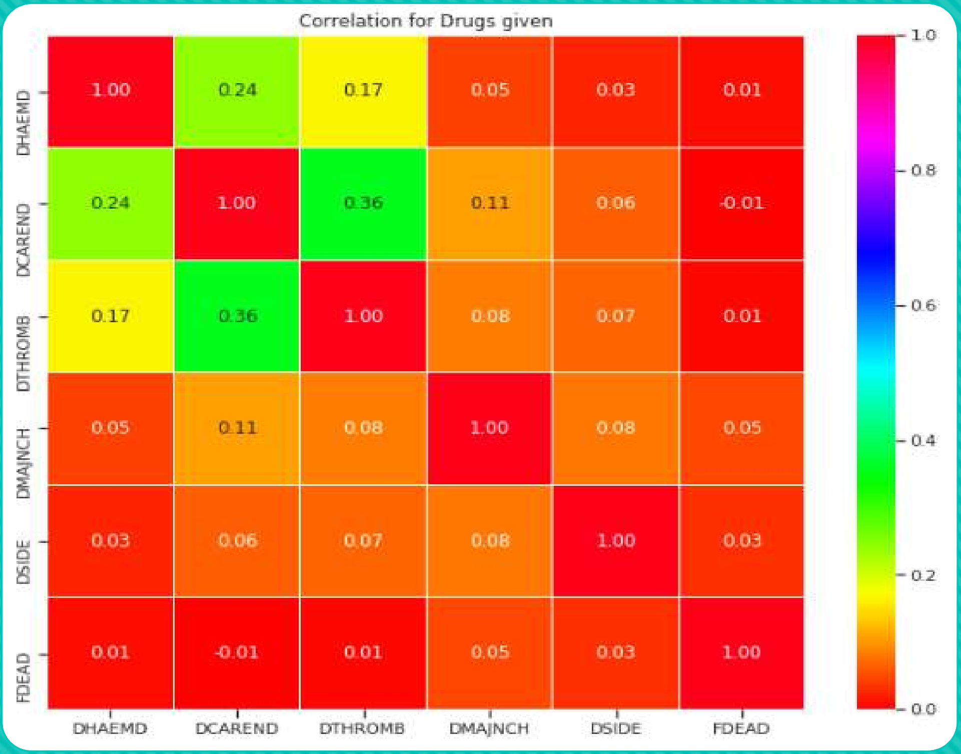
Replace unreadable data to readable data

Most of the months in our data had Polish names and English names. Since Polish names are not understandable by the majority, we have converted them to English

Exploratory Data Analysis

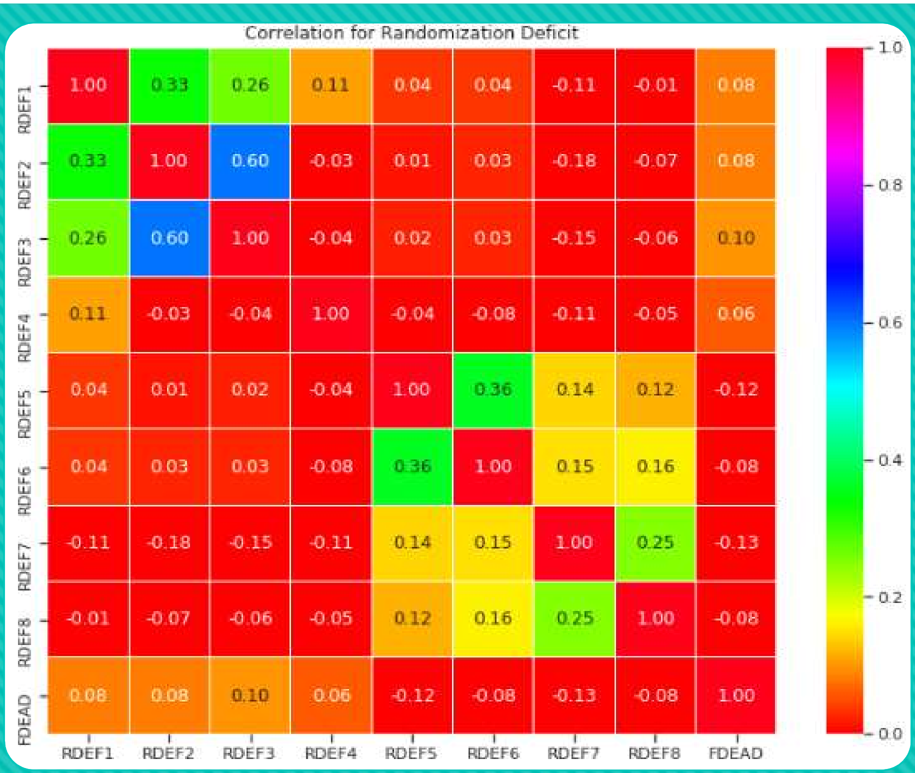
Let us try to explore our data and see how the attributes are related to each other and with our target.





Correlation for Drugs given

	DHAEMD	DCAREND	DTHROMB	DMAJNCH	DSIDE	FDEAD
DHAEMD	1.00	0.24	0.17	0.05	0.03	0.01
DCAREND	0.24	1.00	0.36	0.11	0.06	-0.01
DTHROMB	0.17	0.36	1.00	0.08	0.07	0.01
DMAJNCH	0.05	0.11	0.08	1.00	0.08	0.05
DSIDE	0.03	0.06	0.07	0.08	1.00	0.03
FDEAD	0.01	-0.01	0.01	0.05	0.03	1.00

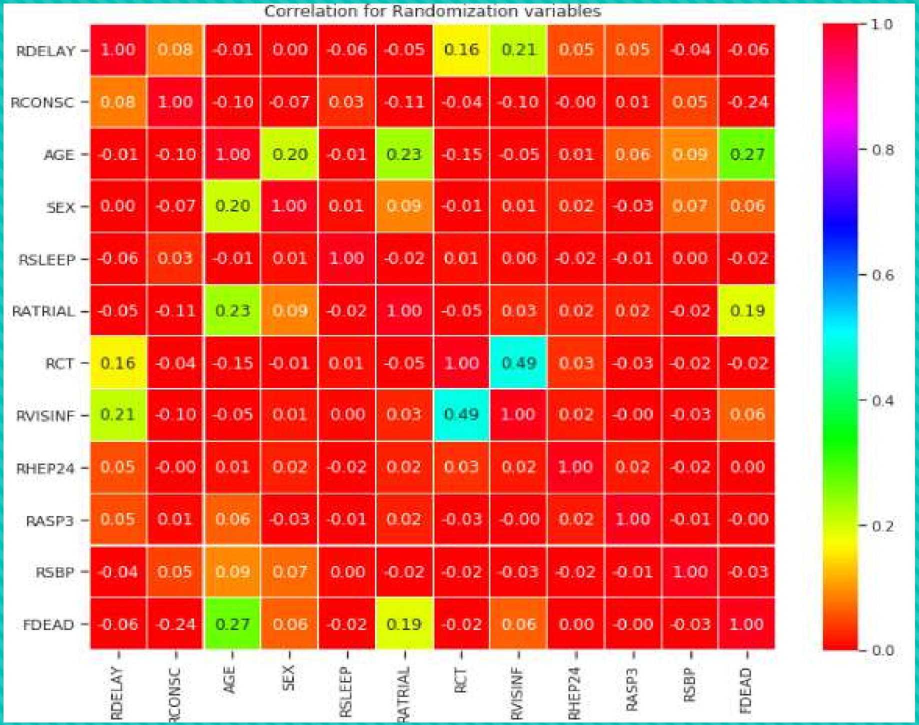


Correlation for Randomization Deficit

	RDEF1	RDEF2	RDEF3	RDEF4	RDEF5	RDEF6	RDEF7	RDEF8	FDEAD
RDEF1	1.00	0.33	0.26	0.11	0.04	0.04	-0.11	-0.01	0.08
RDEF2	0.33	1.00	0.60	-0.03	0.01	0.03	-0.18	-0.07	0.08
RDEF3	0.26	0.60	1.00	-0.04	0.02	0.03	-0.15	-0.06	0.10
RDEF4	0.11	-0.03	-0.04	1.00	-0.04	-0.08	-0.11	-0.05	0.06
RDEF5	0.04	0.01	0.02	-0.04	1.00	0.36	0.14	0.12	-0.12
RDEF6	0.04	0.03	0.03	-0.08	0.36	1.00	0.15	0.16	-0.08
RDEF7	-0.11	-0.18	-0.15	-0.11	0.14	0.15	1.00	0.25	-0.13
RDEF8	-0.01	-0.07	-0.06	-0.05	0.12	0.16	0.25	1.00	-0.08
FDEAD	0.08	0.08	0.10	0.06	-0.12	-0.08	-0.13	-0.08	1.00

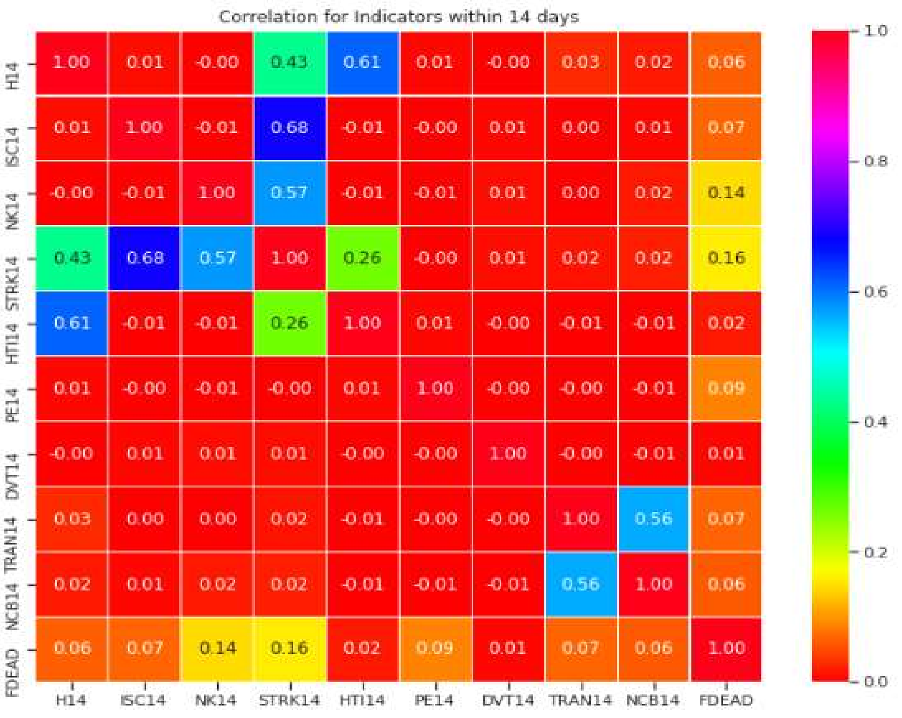
Correlations

Correlations Continued..



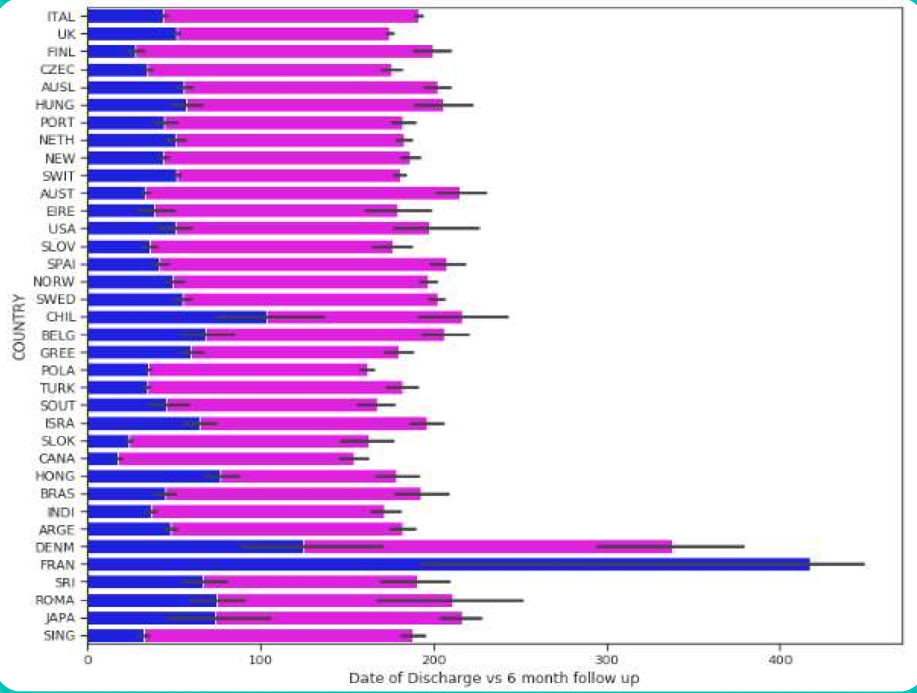
Correlation for Randomization variables

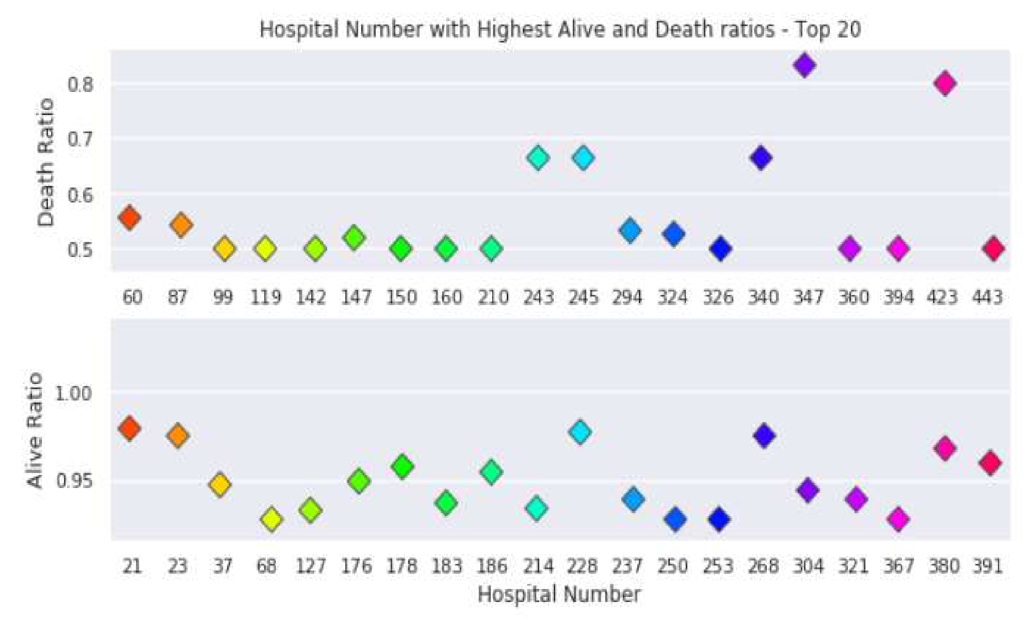
	RDELAY	RCONSC	AGE	SEX	RSLEEP	RATRIAL	RCT	RVISINF	RHEP24	RASP3	RSBP	FDEAD
RDELAY	1.00	0.08	-0.01	0.00	-0.06	-0.05	0.16	0.21	0.05	0.05	-0.04	-0.06
RCONSC	0.08	1.00	-0.10	-0.07	0.03	-0.11	-0.04	-0.10	-0.00	0.01	0.05	-0.24
AGE	-0.01	-0.10	1.00	0.20	-0.01	0.23	-0.15	-0.05	0.01	0.06	0.09	0.27
SEX	0.00	-0.07	0.20	1.00	0.01	0.09	-0.01	0.01	0.02	-0.03	0.07	0.06
RSLEEP	-0.06	0.03	-0.01	0.01	1.00	-0.02	0.01	0.00	-0.02	-0.01	0.00	-0.02
RATRIAL	-0.05	-0.11	0.23	0.09	-0.02	1.00	-0.05	0.03	0.02	0.02	-0.02	0.19
RCT	0.16	-0.04	-0.15	-0.01	0.01	-0.05	1.00	0.49	0.03	-0.03	-0.02	-0.02
RVISINF	0.21	-0.10	-0.05	0.01	0.00	0.03	0.49	1.00	0.02	-0.00	-0.03	0.06
RHEP24	0.05	-0.00	0.01	0.02	-0.02	0.02	0.03	0.02	1.00	0.02	-0.02	0.00
RASP3	0.05	0.01	0.06	-0.03	-0.01	0.02	-0.03	-0.00	0.02	1.00	-0.01	-0.00
RSBP	-0.04	0.05	0.09	0.07	0.00	-0.02	-0.02	-0.03	-0.02	-0.01	1.00	-0.03
FDEAD	-0.06	-0.24	0.27	0.06	-0.02	0.19	-0.02	0.06	0.00	-0.00	-0.03	1.00



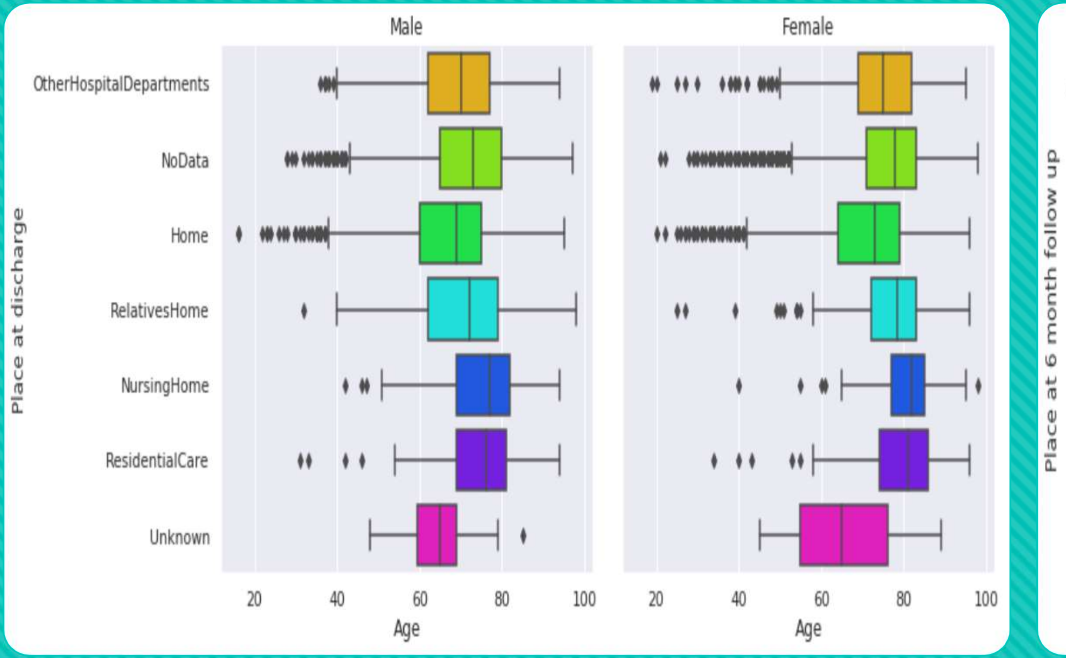
Correlation for Indicators within 14 days

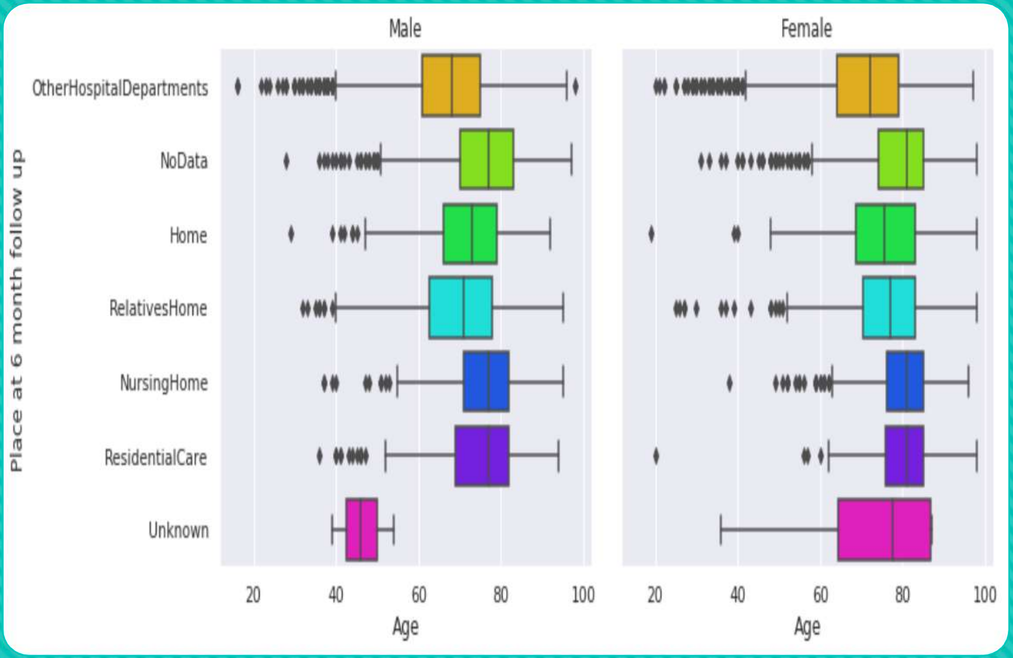
	H14	ISC14	NK14	STRK14	HTI14	PE14	DVT14	TRAN14	NCB14	FDEAD
H14	1.00	0.01	-0.00	0.43	0.61	0.01	-0.00	0.03	0.02	0.06
ISC14	0.01	1.00	-0.01	0.68	-0.01	-0.00	0.01	0.00	0.01	0.07
NK14	-0.00	-0.01	1.00	0.57	-0.01	-0.01	0.01	0.00	0.02	0.14
STRK14	0.43	0.68	0.57	1.00	0.26	-0.00	0.01	0.02	0.02	0.16
HTI14	0.61	-0.01	-0.01	0.26	1.00	0.01	-0.00	-0.01	-0.01	0.02
PE14	0.01	-0.00	-0.01	-0.00	0.01	1.00	-0.00	-0.00	-0.01	0.09
DVT14	-0.00	0.01	0.01	0.01	-0.00	-0.00	1.00	-0.00	-0.01	0.01
TRAN14	0.03	0.00	0.00	0.02	-0.01	-0.00	-0.00	1.00	0.56	0.07
NCB14	0.02	0.01	0.02	0.02	-0.01	-0.01	-0.01	0.56	1.00	0.06
FDEAD	0.06	0.07	0.14	0.16	0.02	0.09	0.01	0.07	0.06	1.00



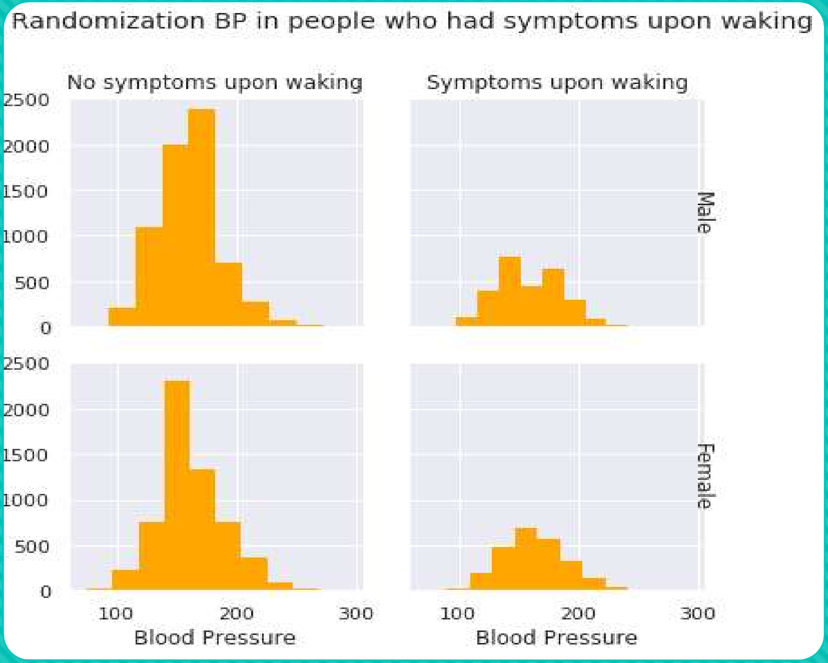


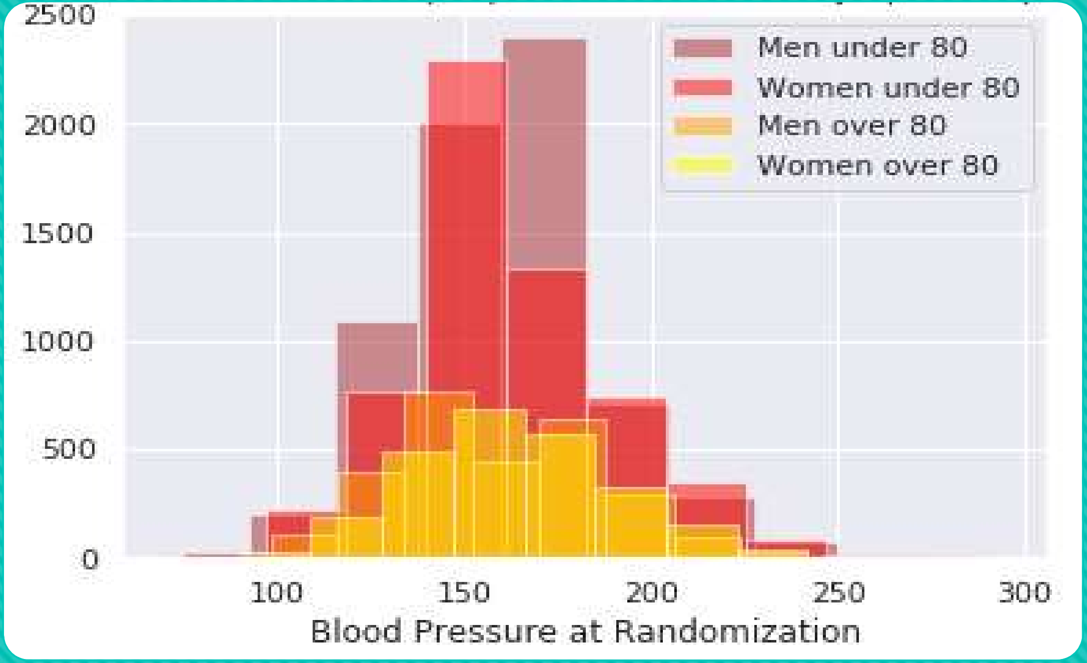
Geographical Data



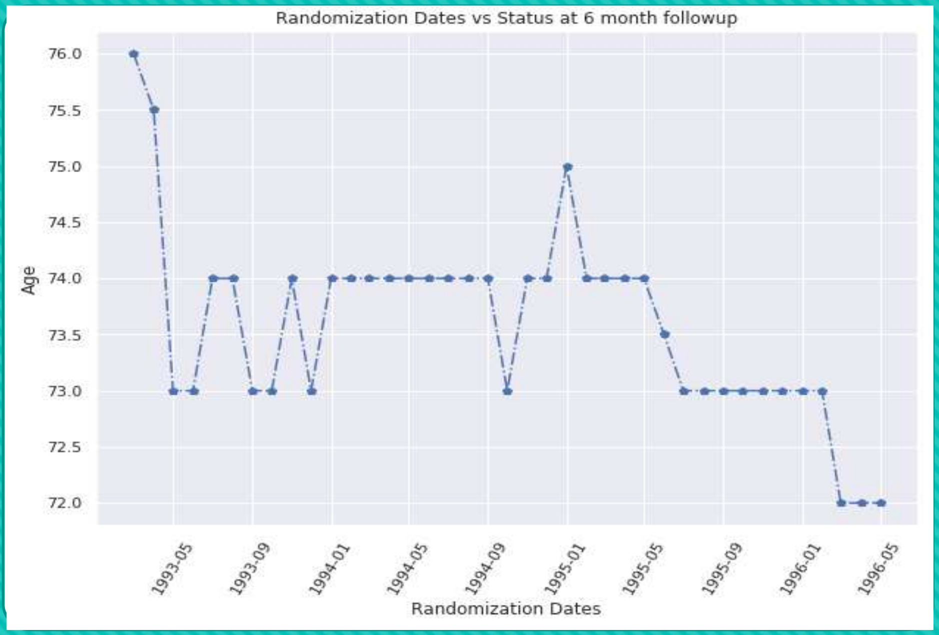


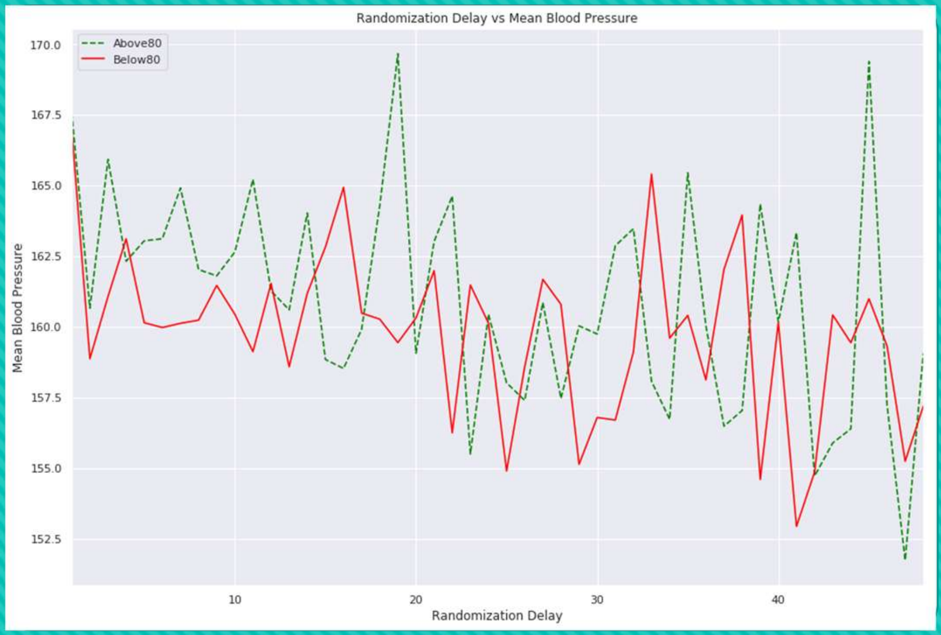
Where do they end up?



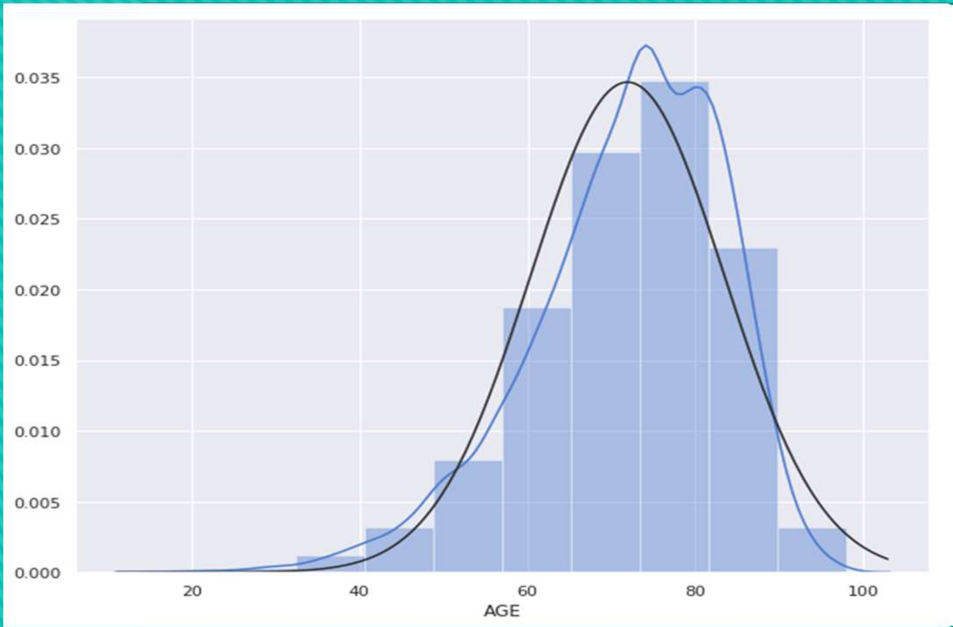


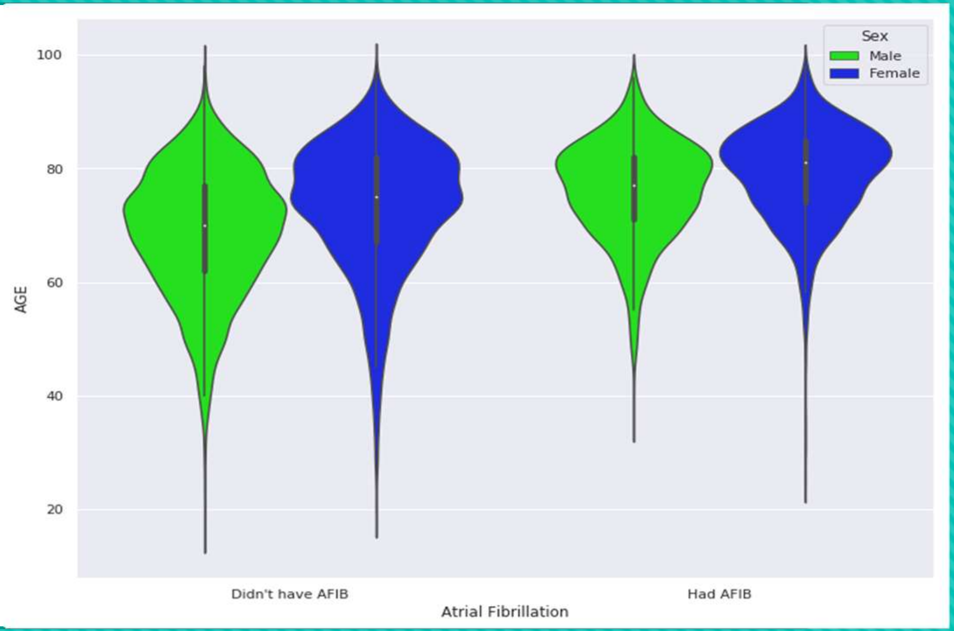
Blood pressure and Symptoms





Randomization data

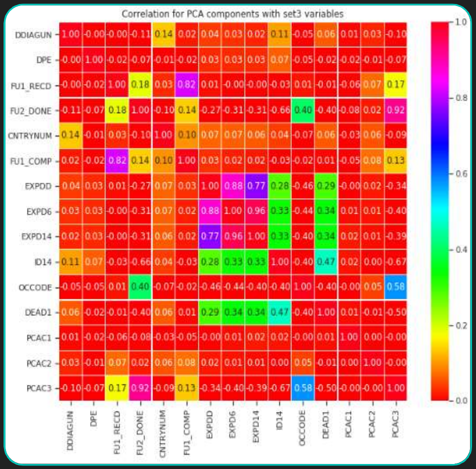




Age distribution and Atrial Fibrillation

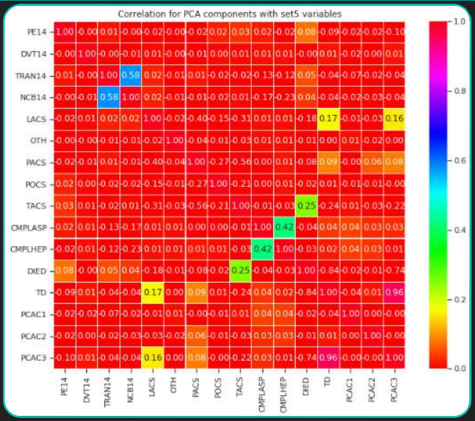
Unsupervised Modelling - PCA

Correlations of PCA components with other components in dataset containing patients 80 years or older.



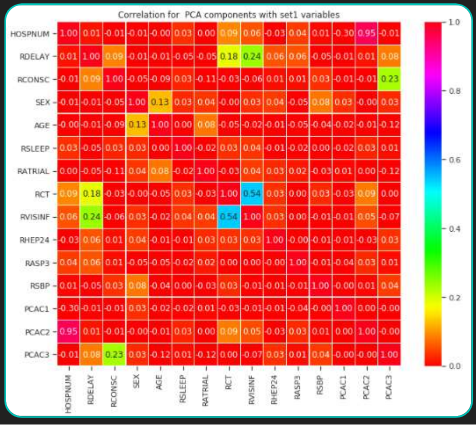
Correlation for PCA components with set3 variables

	DDIAGUN	DPE	FU1_RECD	FU2_DONE	CNTRYNUM	FU1_COMP	EXPDD	EXPD6	EXPD14	ID14	OCCODE	DEAD1	PCAC1	PCAC2	PCAC3
DDIAGUN	1.00	-0.00	-0.00	-0.11	0.14	0.02	0.04	0.03	0.02	0.11	-0.05	0.06	0.01	0.03	-0.10
DPE	-0.00	1.00	-0.02	-0.07	-0.01	-0.02	0.03	0.03	0.03	0.07	-0.05	-0.02	-0.02	-0.01	-0.07
FU1_RECD	-0.00	-0.02	1.00	0.18	0.03	0.82	0.01	-0.00	-0.00	-0.03	0.01	-0.01	-0.06	0.07	0.17
FU2_DONE	-0.11	-0.07	0.18	1.00	-0.10	0.14	-0.27	-0.31	-0.31	-0.66	0.40	-0.40	-0.08	0.02	0.92
CNTRYNUM	0.14	-0.01	0.03	-0.10	1.00	0.10	0.07	0.07	0.06	0.04	-0.07	0.06	-0.03	0.06	-0.09
FU1_COMP	0.02	-0.02	0.82	0.14	0.10	1.00	0.03	0.02	0.02	-0.03	-0.02	0.01	-0.05	0.08	0.13
EXPDD	0.04	0.03	0.01	-0.27	0.07	0.03	1.00	0.88	0.77	0.28	-0.46	0.29	-0.00	0.02	-0.34
EXPD6	0.03	0.03	-0.00	-0.31	0.07	0.02	0.88	1.00	0.96	0.33	-0.44	0.34	0.01	0.01	-0.40
EXPD14	0.02	0.03	-0.00	-0.31	0.06	0.02	0.77	0.96	1.00	0.33	-0.40	0.34	0.02	0.01	-0.39
ID14	0.11	0.07	-0.03	-0.66	0.04	-0.03	0.28	0.33	0.33	1.00	-0.40	0.47	0.02	0.00	-0.67
OCCODE	-0.05	-0.05	0.01	0.40	-0.07	-0.02	-0.46	-0.44	-0.40	-0.40	1.00	-0.40	-0.00	0.05	0.58
DEAD1	0.06	-0.02	-0.01	-0.40	0.06	0.01	0.29	0.34	0.34	0.47	-0.40	1.00	0.01	-0.01	-0.50
PCAC1	0.01	-0.02	-0.06	-0.08	-0.03	-0.05	-0.00	0.01	0.02	0.02	-0.00	0.01	1.00	0.00	-0.00
PCAC2	0.03	-0.01	0.07	0.02	0.06	0.08	0.02	0.01	0.01	0.00	0.05	-0.01	0.00	1.00	-0.00
PCAC3	-0.10	-0.07	0.17	0.92	-0.09	0.13	-0.34	-0.40	-0.39	-0.67	0.58	-0.50	-0.00	-0.00	1.00



Correlation for PCA components with set5 variables

	PE14	DVT14	TRAN14	NCB14	LACS	OTH	PACS	POCS	TACS	CMPLASP	CMPLHEP	DIED	TD	PCAC1	PCAC2	PCAC3
PE14	1.00	-0.00	0.01	-0.00	-0.02	-0.00	-0.02	0.02	0.03	0.02	-0.02	0.08	-0.09	-0.02	-0.02	-0.10
DVT14	-0.00	1.00	-0.00	-0.01	0.01	-0.00	-0.01	0.00	0.01	0.01	0.01	-0.00	0.01	-0.02	0.00	0.01
TRAN14	0.01	-0.00	1.00	0.58	0.02	-0.01	0.01	-0.02	-0.02	-0.13	-0.12	0.05	-0.04	-0.07	-0.02	-0.04
NCB14	-0.00	-0.01	0.58	1.00	0.02	-0.01	-0.01	-0.02	0.01	-0.17	-0.23	0.04	-0.04	-0.02	-0.03	-0.04
LACS	-0.02	0.01	0.02	0.02	1.00	-0.02	-0.40	-0.15	-0.31	0.01	0.01	-0.18	0.17	-0.01	-0.03	0.16
OTH	-0.00	-0.00	-0.01	-0.01	-0.02	1.00	-0.04	-0.01	-0.03	0.01	0.01	-0.01	0.00	0.01	-0.02	0.00
PACS	-0.02	-0.01	0.01	-0.01	-0.40	-0.04	1.00	-0.27	-0.56	0.00	0.01	-0.08	0.09	-0.00	0.06	0.08
POCS	0.02	0.00	-0.02	-0.02	-0.15	-0.01	-0.27	1.00	-0.21	0.00	0.01	-0.02	0.01	-0.01	-0.01	-0.00
TACS	0.03	0.01	-0.02	0.01	-0.31	-0.03	-0.56	-0.21	1.00	-0.01	-0.03	0.25	-0.24	0.01	-0.03	-0.22
CMPLASP	0.02	0.01	-0.13	-0.17	0.01	0.01	0.00	0.00	-0.01	1.00	0.42	-0.04	0.04	0.04	0.03	0.03
CMPLHEP	-0.02	0.01	-0.12	-0.23	0.01	0.01	0.01	0.01	-0.03	0.42	1.00	-0.03	0.02	0.04	0.03	0.01
DIED	0.08	-0.00	0.05	0.04	-0.18	-0.01	-0.08	-0.02	0.25	-0.04	-0.03	1.00	-0.84	-0.02	-0.01	-0.74
TD	-0.09	0.01	-0.04	-0.04	0.17	0.00	0.09	0.01	-0.24	0.04	0.02	-0.84	1.00	-0.04	0.01	0.96
PCAC1	-0.02	-0.02	-0.07	-0.02	-0.01	0.01	-0.00	-0.01	0.01	0.04	0.04	-0.02	-0.04	1.00	0.00	-0.00
PCAC2	-0.02	0.00	-0.02	-0.03	-0.03	-0.02	0.06	-0.01	-0.03	0.03	0.03	-0.01	0.01	0.00	1.00	-0.00
PCAC3	-0.10	0.01	-0.04	-0.04	0.16	0.00	0.08	-0.00	-0.22	0.03	0.01	-0.74	0.96	-0.00	-0.00	1.00



Correlation for PCA components with set1 variables

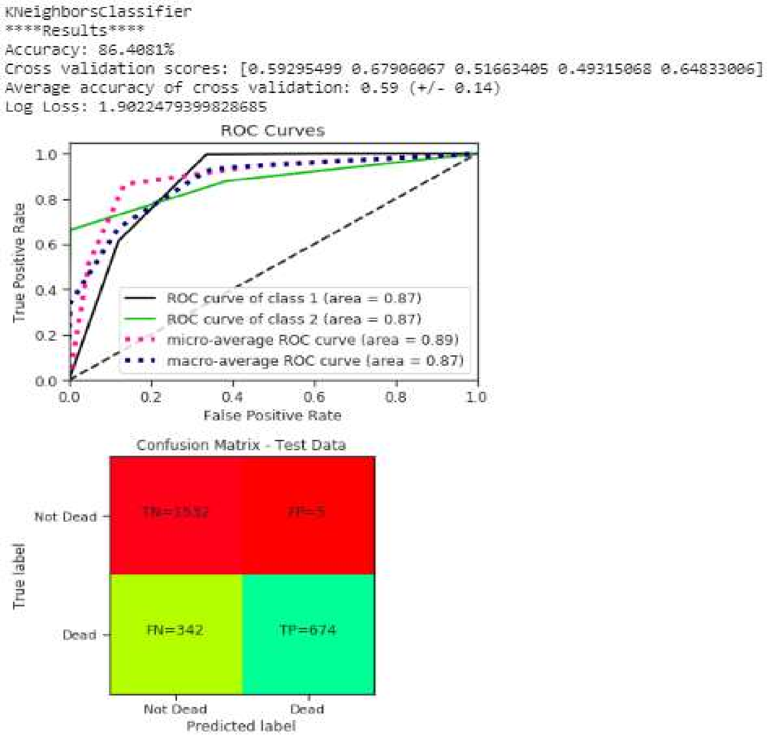
	HOSPNUM	RDELAY	RCONSC	SEX	AGE	RSLEEP	RATRIAL	RCT	RVISINF	RHEP24	RASP3	RSBP	PCAC1	PCAC2	PCAC3
HOSPNUM	1.00	0.01	-0.01	-0.01	-0.00	0.03	0.00	0.09	0.06	-0.03	0.04	0.01	-0.30	0.95	-0.01
RDELAY	0.01	1.00	0.09	-0.01	-0.01	-0.05	-0.05	0.18	0.24	0.06	0.06	-0.05	-0.01	0.01	0.08
RCONSC	-0.01	0.09	1.00	-0.05	-0.09	0.03	-0.11	-0.03	-0.06	0.01	0.01	0.03	-0.01	-0.01	0.23
SEX	-0.01	-0.01	-0.05	1.00	0.13	0.03	0.04	-0.00	0.03	0.04	-0.05	0.08	0.03	-0.00	0.03
AGE	-0.00	-0.01	-0.09	0.13	1.00	0.00	0.08	-0.05	-0.02	-0.01	-0.05	-0.04	-0.02	-0.01	-0.12
RSLEEP	0.03	-0.05	0.03	0.03	0.00	1.00	-0.02	0.03	0.04	-0.01	-0.02	0.00	-0.02	0.03	0.01
RATRIAL	0.00	-0.05	-0.11	0.04	0.08	-0.02	1.00	-0.03	0.04	0.03	0.02	-0.03	0.01	0.00	-0.12
RCT	0.09	0.18	-0.03	-0.00	-0.05	0.03	-0.03	1.00	0.54	0.03	0.00	0.03	-0.03	0.09	0.00
RVISINF	0.06	0.24	-0.06	0.03	-0.02	0.04	0.04	0.54	1.00	0.03	0.00	-0.01	-0.01	0.05	-0.07
RHEP24	-0.03	0.06	0.01	0.04	-0.01	-0.01	0.03	0.03	0.03	1.00	-0.00	-0.01	-0.01	-0.03	0.03
RASP3	0.04	0.06	0.01	-0.05	-0.05	-0.02	0.02	0.00	0.00	-0.00	1.00	-0.01	-0.04	0.03	0.01
RSBP	0.01	-0.05	0.03	0.08	-0.04	0.00	-0.03	0.03	-0.01	-0.01	-0.01	1.00	-0.00	0.01	0.04
PCAC1	-0.30	-0.01	-0.01	0.03	-0.02	-0.02	0.01	-0.03	-0.01	-0.01	-0.04	-0.00	1.00	0.00	-0.00
PCAC2	0.95	0.01	-0.01	-0.00	-0.01	0.03	0.00	0.09	0.05	-0.03	0.03	0.01	0.00	1.00	-0.00
PCAC3	-0.01	0.08	0.23	0.03	-0.12	0.01	-0.12	0.00	-0.07	0.03	0.01	0.04	-0.00	-0.00	1.00

Supervised modelling

Let us try some of the supervised modelling techniques on patients 80 years and above and hence predict survival chances. Older people have been substantially under represented in stroke trials to date, so we hope the number of patients aged over 80 in this data set could also facilitate planning of trials in the 'older old'.

Supervised Modelling Continued..

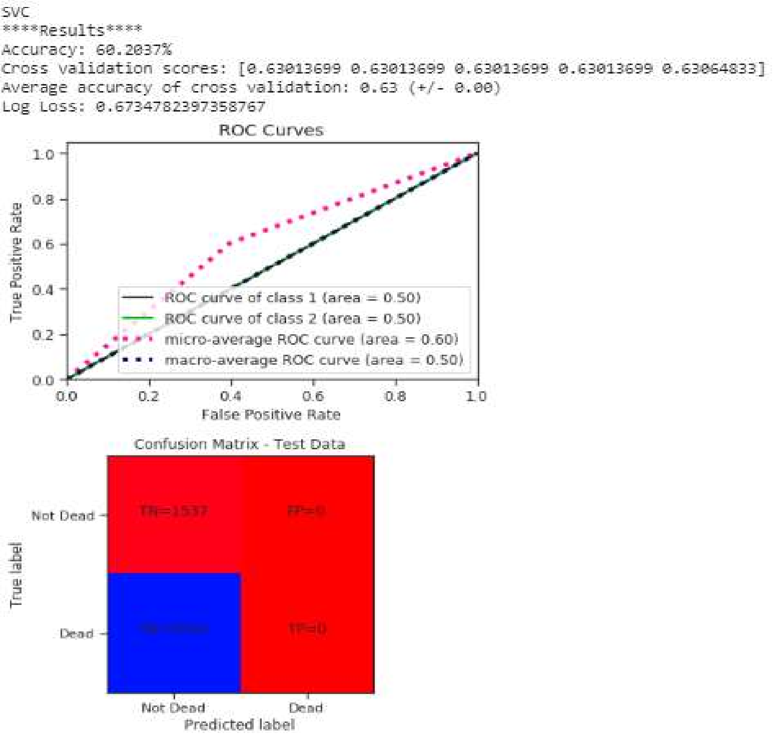
KNN Classifier Model



```
KNeighborsClassifier
****Results****
Accuracy: 86.4081%
Cross validation scores: [0.59295499 0.67906067 0.51663405 0.49315068 0.64833006]
Average accuracy of cross validation: 0.59 (+/- 0.14)
Log Loss: 1.9022479399828685
```

ROC Curves
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve of class 1 (area = 0.87)
ROC curve of class 2 (area = 0.87)
micro-average ROC curve (area = 0.89)
macro-average ROC curve (area = 0.87)
Confusion Matrix - Test Data
True label
Not Dead
Dead
TN=1532
FP=5
FN=342
TP=674
Not Dead
Dead
Predicted label

SVM Classifier Model

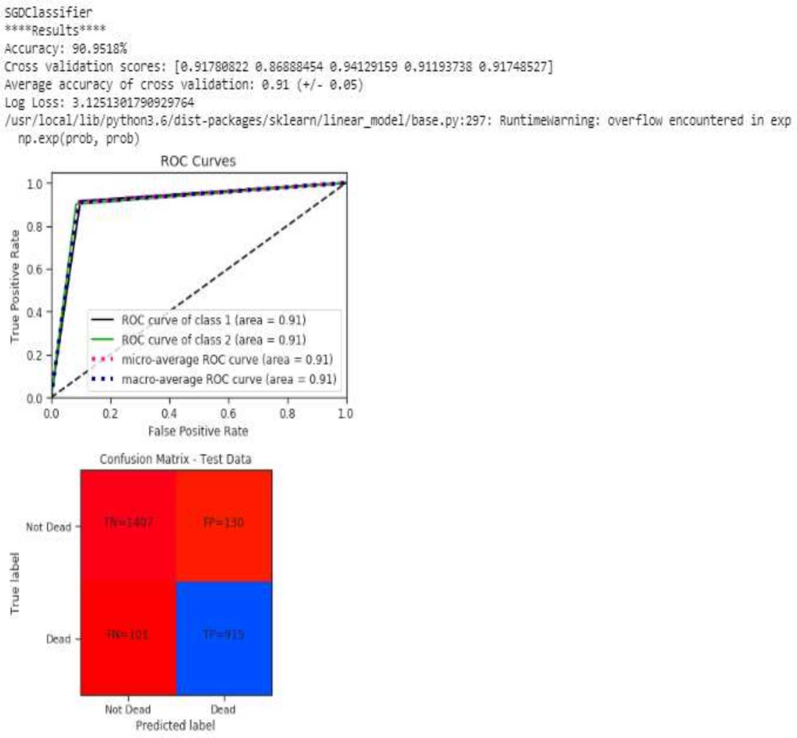


```
SVC
****Results****
Accuracy: 60.2037%
Cross validation scores: [0.63013699 0.63013699 0.63013699 0.63013699 0.63064833]
Average accuracy of cross validation: 0.63 (+/- 0.00)
Log Loss: 0.6734782397358767
```

ROC Curves
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve of class 1 (area = 0.50)
ROC curve of class 2 (area = 0.50)
micro-average ROC curve (area = 0.60)
macro-average ROC curve (area = 0.50)
Confusion Matrix - Test Data
True label
Not Dead
Dead
TN=1537
FP=0
TP=0
Not Dead
Dead
Predicted label

Supervised Modelling Continued..

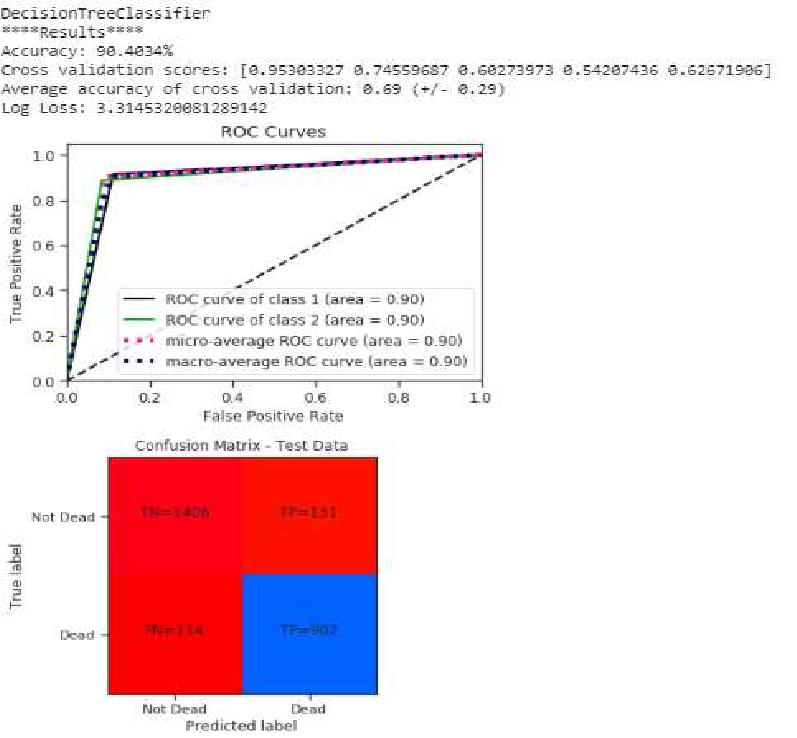
SGD Classifier Model



```
SGDClassifier
****Results****
Accuracy: 90.9518%
Cross validation scores: [0.91780822 0.86888454 0.94129159 0.91193738 0.91748527]
Average accuracy of cross validation: 0.91 (+/- 0.05)
Log Loss: 3.1251301790929764
/usr/local/lib/python3.6/dist-packages/sklearn/linear_model/base.py:297: RuntimeWarning: overflow encountered in exp
  np.exp(prob, prob)
```

ROC Curves
1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
ROC curve of class 1 (area = 0.91)
ROC curve of class 2 (area = 0.91)
micro-average ROC curve (area = 0.91)
macro-average ROC curve (area = 0.91)
Confusion Matrix - Test Data
Not Dead
Dead
True label
TN=1407
FP=130
FN=101
TP=915
Not Dead
Dead
Predicted label

Decision Tree Classifier Model

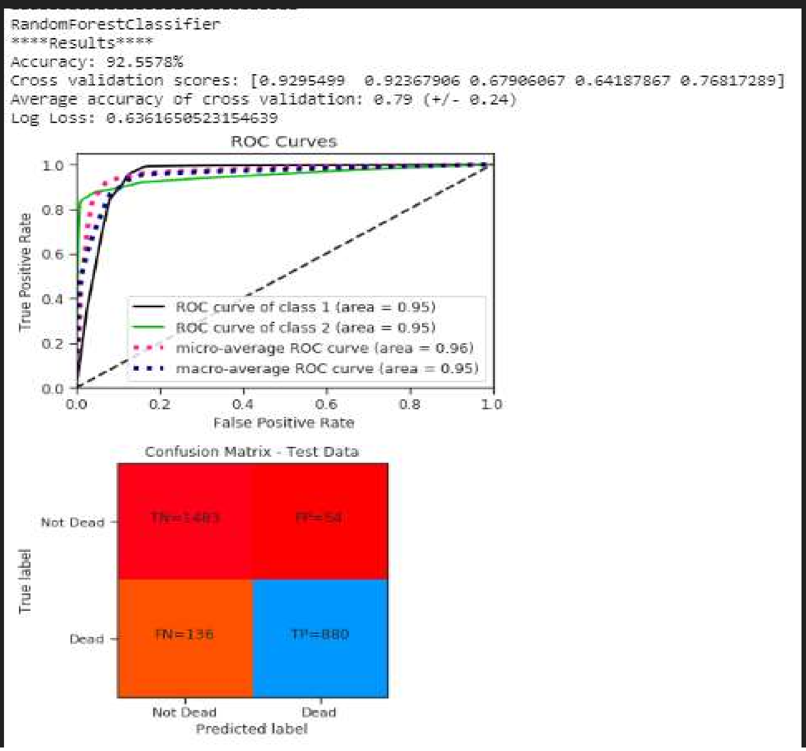


```
DecisionTreeClassifier
****Results****
Accuracy: 90.4034%
Cross validation scores: [0.95303327 0.74559687 0.60273973 0.54207436 0.62671906]
Average accuracy of cross validation: 0.69 (+/- 0.29)
Log Loss: 3.3145320081289142
```

ROC Curves
1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
ROC curve of class 1 (area = 0.90)
ROC curve of class 2 (area = 0.90)
micro-average ROC curve (area = 0.90)
macro-average ROC curve (area = 0.90)
Confusion Matrix - Test Data
Not Dead
Dead
True label
TN=1406
FP=131
FN=134
TP=902
Not Dead
Dead
Predicted label

Supervised Modelling Continued..

Random Forest Classifier Model

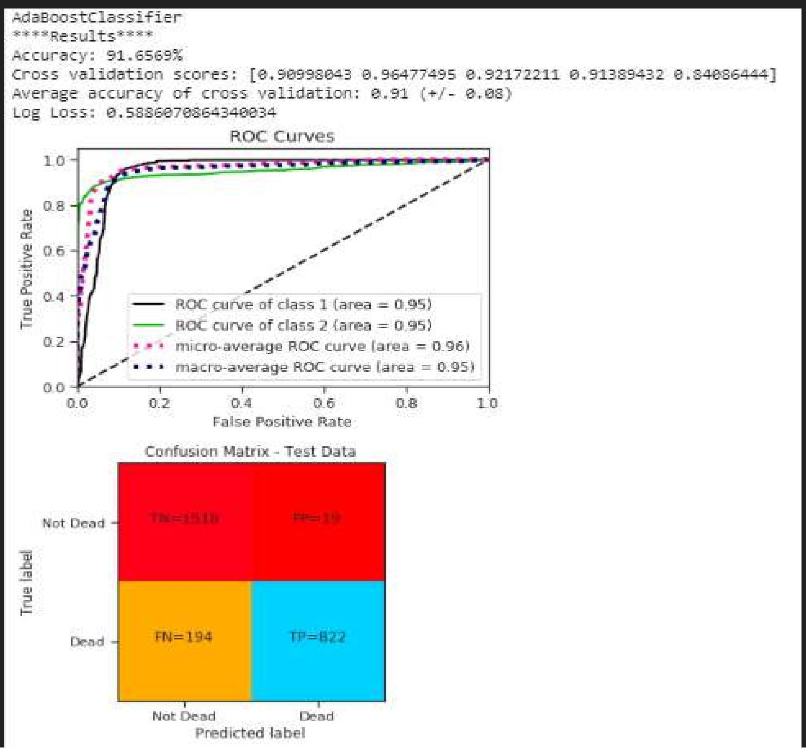


```
RandomForestClassifier
****Results****
Accuracy: 92.5578%
Cross validation scores: [0.9295499  0.92367906 0.67906067 0.64187867 0.76817289]
Average accuracy of cross validation: 0.79 (+/- 0.24)
Log Loss: 0.6361650523154639
```

ROC Curves
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve of class 1 (area = 0.95)
ROC curve of class 2 (area = 0.95)
micro-average ROC curve (area = 0.96)
macro-average ROC curve (area = 0.95)
Confusion Matrix - Test Data
True label
Predicted label

	Not Dead	Dead
Not Dead	TN=1483	FP=54
Dead	FN=136	TP=880

ADA Boost Classifier Model



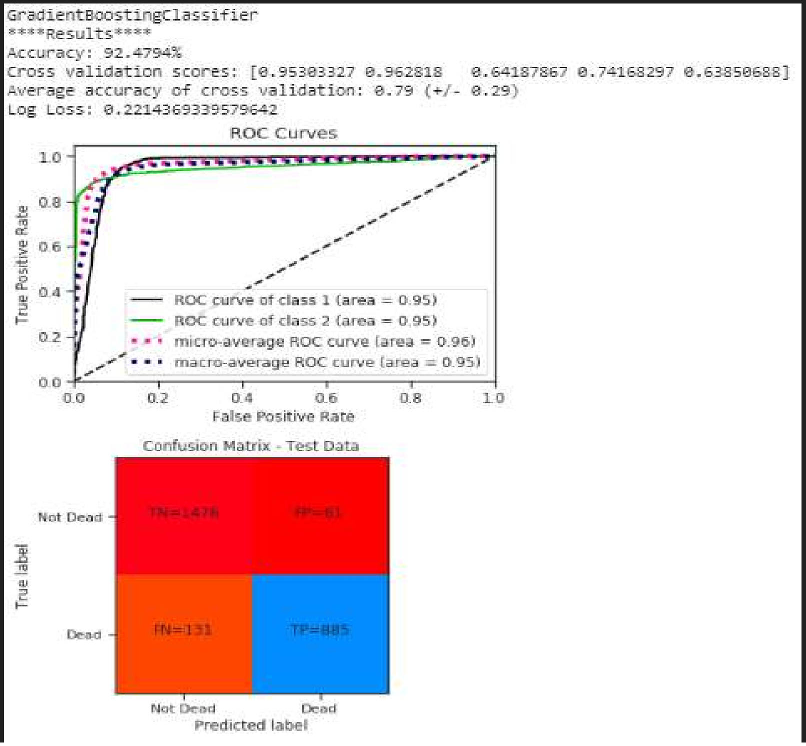
```
AdaBoostClassifier
****Results****
Accuracy: 91.6569%
Cross validation scores: [0.90998043 0.96477495 0.92172211 0.91389432 0.84086444]
Average accuracy of cross validation: 0.91 (+/- 0.08)
Log Loss: 0.5886070864340034
```

ROC Curves
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve of class 1 (area = 0.95)
ROC curve of class 2 (area = 0.95)
micro-average ROC curve (area = 0.96)
macro-average ROC curve (area = 0.95)
Confusion Matrix - Test Data
True label
Predicted label

	Not Dead	Dead
Not Dead	TN=1518	FP=19
Dead	FN=194	TP=822

Supervised Modelling Continued..

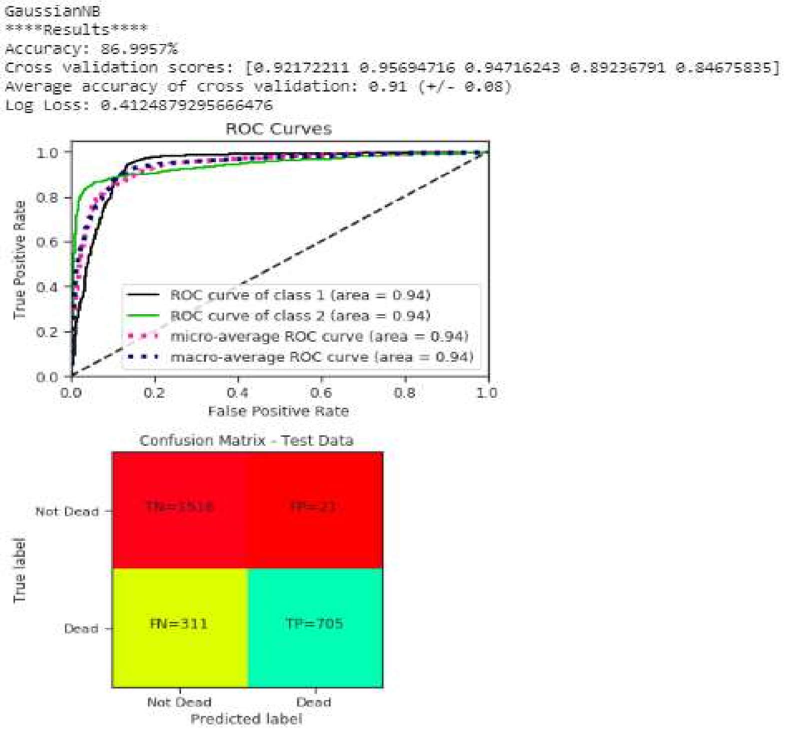
Gradient Boost Classifier Model



```
GradientBoostingClassifier
****Results****
Accuracy: 92.4794%
Cross validation scores: [0.95303327 0.962818   0.64187867 0.74168297 0.63850688]
Average accuracy of cross validation: 0.79 (+/- 0.29)
Log Loss: 0.2214369339579642
```

ROC Curves
True Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
ROC curve of class 1 (area = 0.95)
ROC curve of class 2 (area = 0.95)
micro-average ROC curve (area = 0.96)
macro-average ROC curve (area = 0.95)
Confusion Matrix - Test Data
True label
Not Dead
Dead
TN=1478
FP=61
FN=131
TP=885
Not Dead
Dead
Predicted label

Gaussian Naïve Bayes Classifier Model

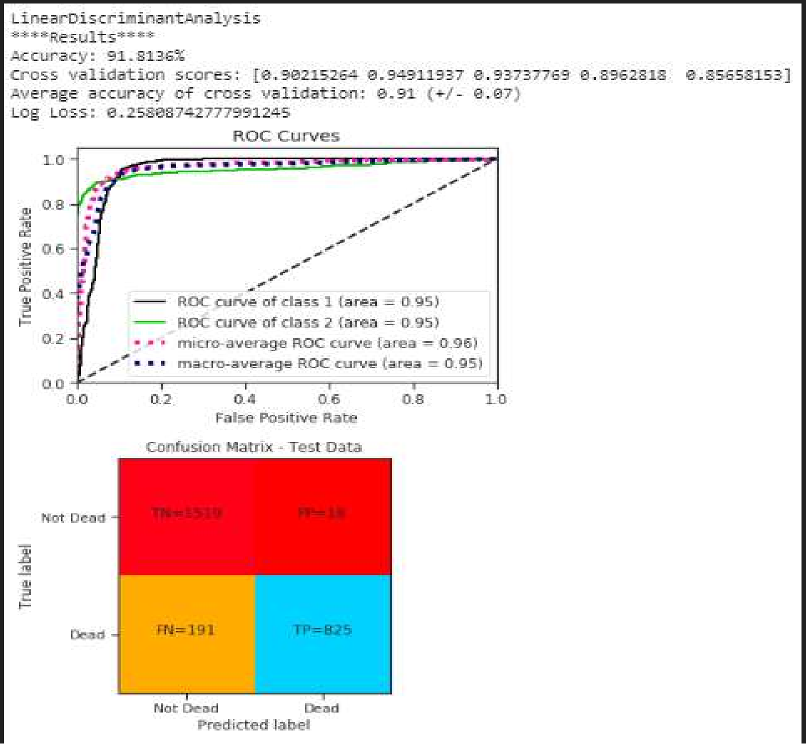


```
GaussianNB
****Results****
Accuracy: 86.9957%
Cross validation scores: [0.92172211 0.95694716 0.94716243 0.89236791 0.84675835]
Average accuracy of cross validation: 0.91 (+/- 0.08)
Log Loss: 0.4124879295666476
```

ROC Curves
True Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
ROC curve of class 1 (area = 0.94)
ROC curve of class 2 (area = 0.94)
micro-average ROC curve (area = 0.94)
macro-average ROC curve (area = 0.94)
Confusion Matrix - Test Data
True label
Not Dead
Dead
TN=1518
FP=21
FN=311
TP=705
Not Dead
Dead
Predicted label

Supervised Modelling Continued..

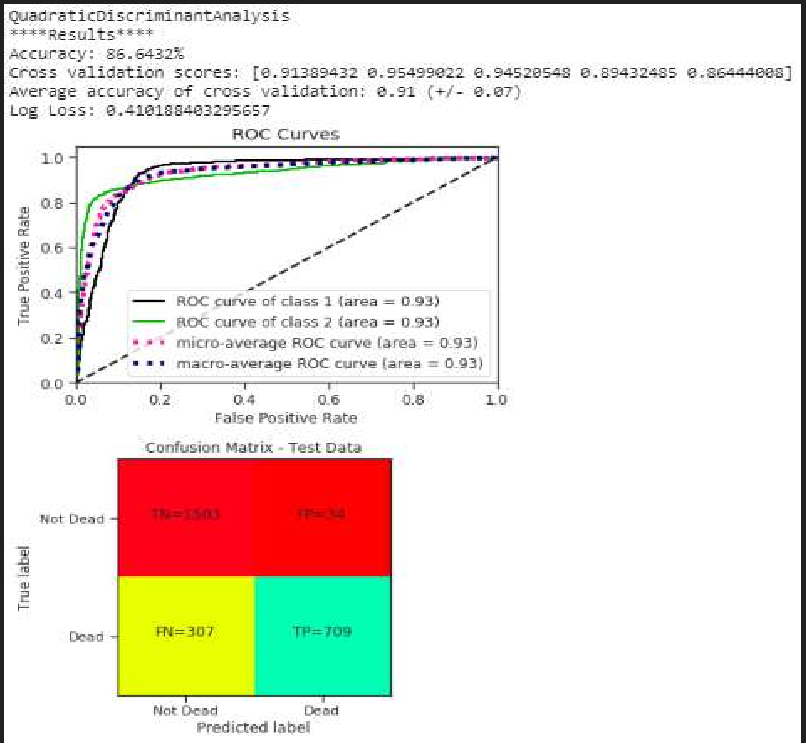
Linear Discriminant Analysis Model



```
LinearDiscriminantAnalysis
****Results****
Accuracy: 91.8136%
Cross validation scores: [0.90215264 0.94911937 0.93737769 0.8962818  0.85658153]
Average accuracy of cross validation: 0.91 (+/- 0.07)
Log Loss: 0.25808742777991245
```

ROC Curves
1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
ROC curve of class 1 (area = 0.95)
ROC curve of class 2 (area = 0.95)
micro-average ROC curve (area = 0.96)
macro-average ROC curve (area = 0.95)
Confusion Matrix - Test Data
Not Dead
Dead
True label
TN=1519
FP=18
FN=191
TP=825
Not Dead
Dead
Predicted label

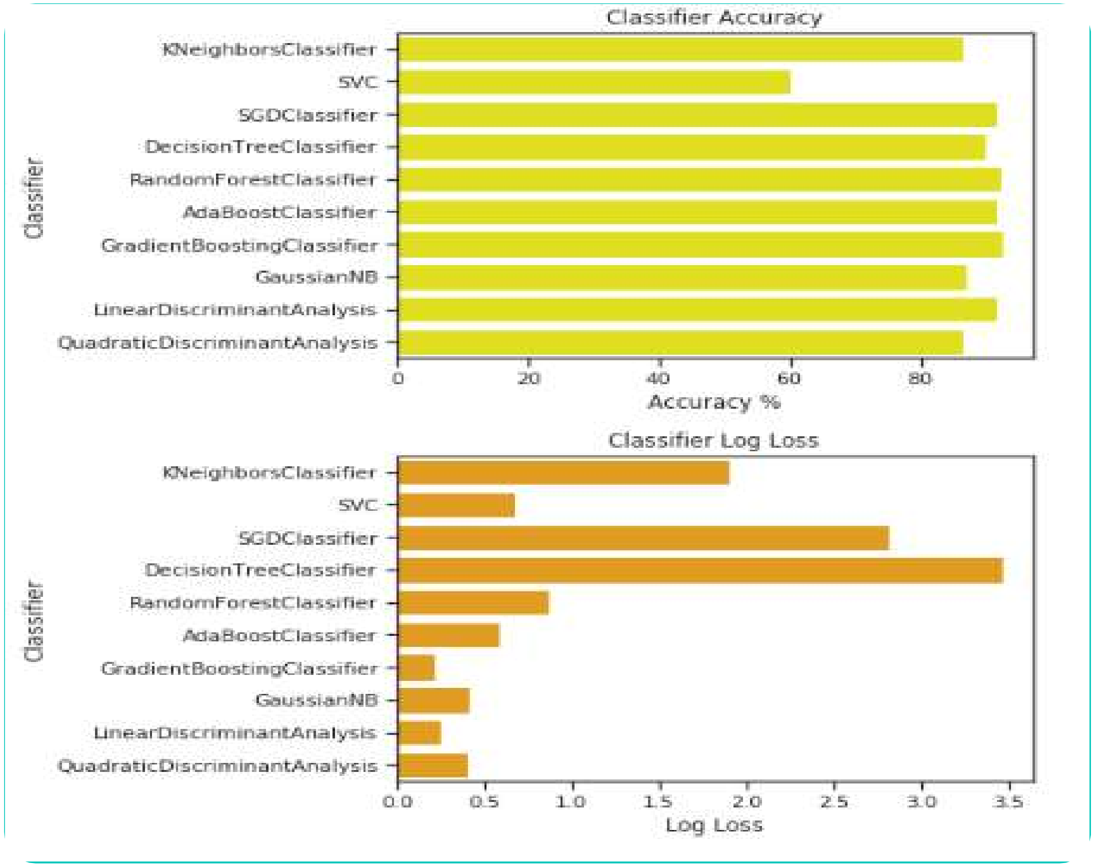
Quadratic Discriminant Analysis Model



```
QuadraticDiscriminantAnalysis
****Results****
Accuracy: 86.6432%
Cross validation scores: [0.91389432 0.95499022 0.94520548 0.89432485 0.86444008]
Average accuracy of cross validation: 0.91 (+/- 0.07)
Log Loss: 0.410188403295657
```

ROC Curves
1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
ROC curve of class 1 (area = 0.93)
ROC curve of class 2 (area = 0.93)
micro-average ROC curve (area = 0.93)
macro-average ROC curve (area = 0.93)
Confusion Matrix - Test Data
Not Dead
Dead
True label
TN=1503
FP=34
FN=307
TP=709
Not Dead
Dead
Predicted label

Best Model for patients 80 years and above



We are choosing Linear Discriminant Analysis as the best model in terms of prediction as well as log loss.

Conclusion

Linear Discriminant Analysis was able to provide an average prediction of 85% survival chances for patients 80 years and above with an accuracy of 91% and a Loss of 25% with consistent Cross Validation Scores. Even though it appears that elderly have good survival chances, we need to keep in mind that our predictions were based on the dataset surveyed between 1993 to 1996.

Our dataset also doesn’t capture some of the essential variables like height, weight, BMI, cholesterol, whether patients were smokers or alcoholics or diabetics and their measures, which, if provided, would help us make more accurate predictions.

Next Steps

Our predictions were for the 90s where medical facilities and surveying techniques were not as much advanced as it is in the current age. Our next steps can be to survey the current stroke patients with all the essential variables and risk factors and use iteration and evaluation techniques and re-use the models which we used for this dataset and compare our findings and see if the differences were statistically significant or not using T-tests and P-Values.

Prof. Anna Czlonkowska

Dr. hab Maciej Niewada

Prof. Peter Sandercock

Credits

Thank you!