Fertility dataset

PROJECT (CLASSIFICATION)

TEAM-4

A.AKHILA
CH.RAHUL
R.MITALI
K.BHANU PRASAD REDDY
V.NANDINI

ABOUT DATA

Fertility is most likely if the semen discharged in a single ejaculation (ejaculate) contains at least 15 million sperm per milliliter. Sperm concentration are related to socio-demographic data, environmental factors, health status, and life habits in UC Irvine machine learning repository.

OBJECTIVE

So the main objective is to come up with better analysis and solutions.The task here is to build a Multivariate logistic regression model to predict "**TARGET Output**" Diagnosis.

THE PATH

To fit a best model by using a Standard Machine Learning Algorithm.

DATA AND DATA QUALITY CHECK

DATA INTRODUCTION :

Data consist of 100 volunteers provide a semen sample analyzed according to the WHO 2010 criteria. Sperm concentration are related to socio-demographic data, environmental factors, health status, and life habits

VARIABLES :

The attributes which are identified , which consist of the Season(winter,spring,Summer, fall.), Age(18-36), Childish diseases(yes or no), Accident or serious trauma(yes or no), Surgical intervention(yes or no), High fevers in the last year(less than three months ago,more than three months ago,no.), Frequency of alcohol consumption(several times a day,every day,several times a week,once a week,hardly ever or never),Smoking habit(never,occasional,daily). Number of hours spent sitting per day(yes or no), Output:Diagnosis(normal(N),altred(O)).

ATTRIBUTE INFORMATION

- **Season** in which the analysis was performed. 1) winter, 2) spring, 3) Summer, 4) fall. (-1, -0.33, 0.33, 1)
- **Age** at the time of analysis. 18-36 (0, 1)
- **Childish diseases** (ie , chicken pox, measles, mumps, polio) 1) yes, 2) no. (0, 1)
- **Accident or serious trauma** 1) yes, 2) no. (0, 1)
- **Surgical intervention** 1) yes, 2) no. (0, 1)
- **High fevers** in the last year 1) less than three months ago, 2) more than three months ago, 3) no. (-1, 0, 1)
- **Frequency of alcohol consumption** 1) several times a day, 2) every day, 3) several times a week, 4) once a week, 5) hardly ever or never (0, 1)

- Smoking habit 1) never, 2) occasional 3) daily. (-1, 0, 1)
- Number of hours spent sitting per day (0, 1)

Output (Dependent variable)

Diagnosis normal (N), altered (O)

MISSING VALUES

We didnt find any Missing or NULL values in our dataset. and our Target variable was replaced to 0's and 1's, and outliers were identified in the columns.

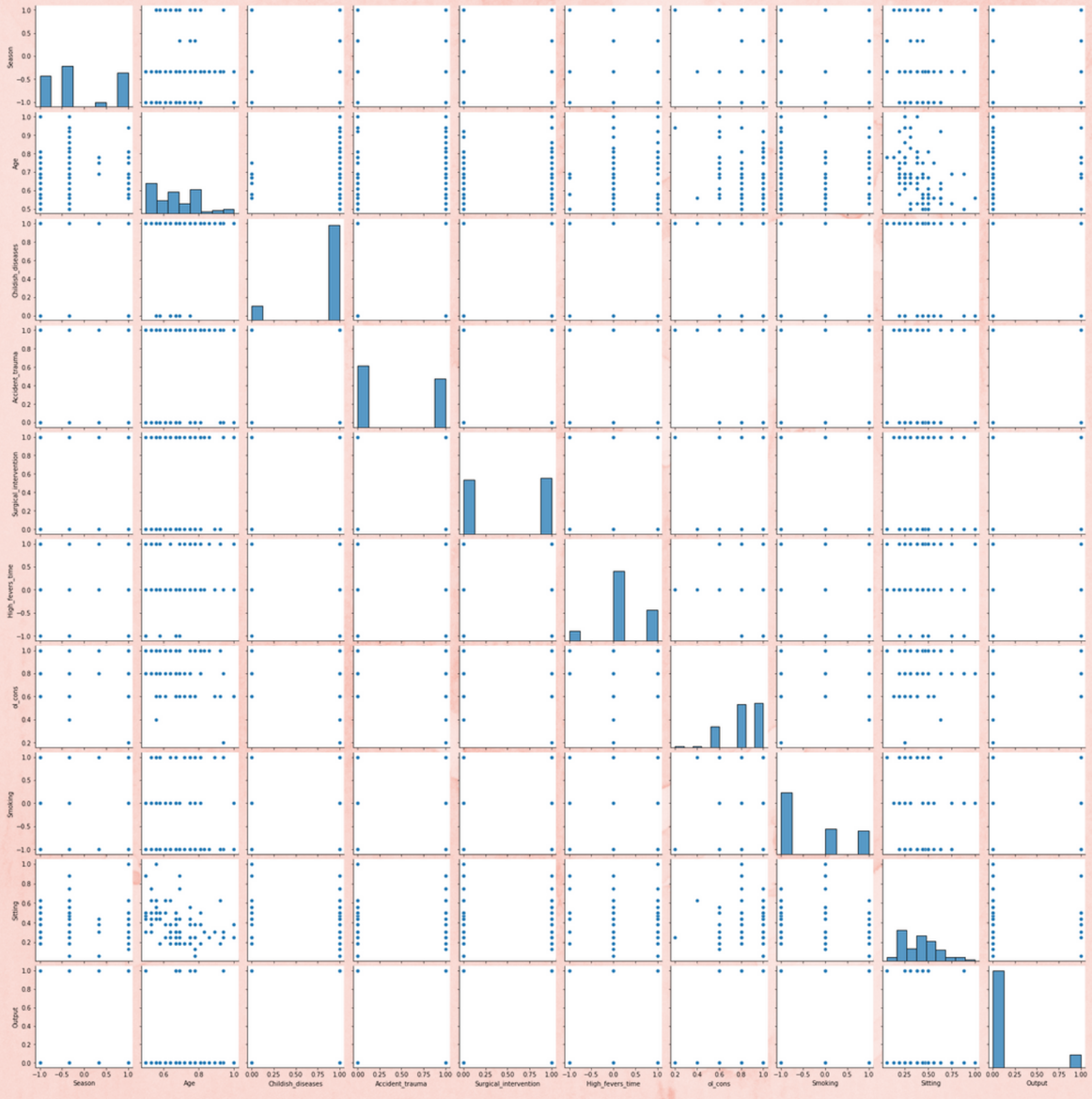
	Season	Age	Childish_diseases	Accident_trauma	Surgical_intervention	High_fevers_time	ol_cons	Smoking	Sitting	Output
0	-0.33	0.69	0	1	1	0	0.8	0	0.88	0
1	-0.33	0.94	1	0	1	0	0.8	1	0.31	1
2	-0.33	0.50	1	0	0	0	1.0	-1	0.50	0
3	-0.33	0.75	0	1	1	0	1.0	-1	0.38	0
4	-0.33	0.67	1	1	0	0	0.8	-1	0.50	1
...	...	...	...	...	...	...	...	...	...	...
95	-1.00	0.67	1	0	0	0	1.0	-1	0.50	0
96	-1.00	0.61	1	0	0	0	0.8	0	0.50	0
97	-1.00	0.67	1	1	1	0	1.0	-1	0.31	0
98	-1.00	0.64	1	0	1	0	1.0	0	0.19	0
99	-1.00	0.69	0	1	1	0	0.6	-1	0.19	0

We have our dependent variable as categorical so we replaced Normal and Altered into binary values i.e., 0's and 1's .

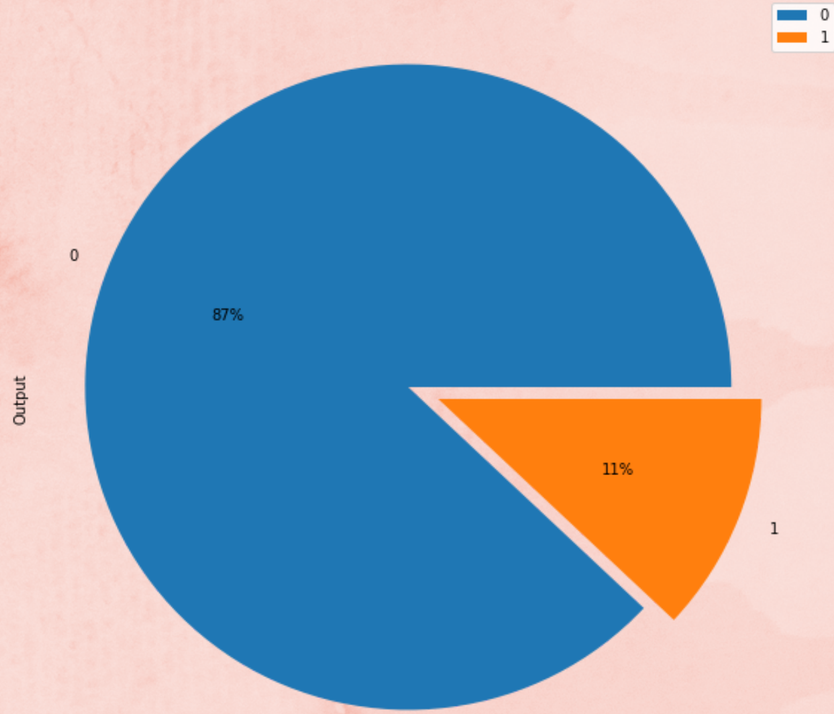
EDA(Exploratory Data Analysis)

PAIR PLOT

relations between all features and distributions of all features

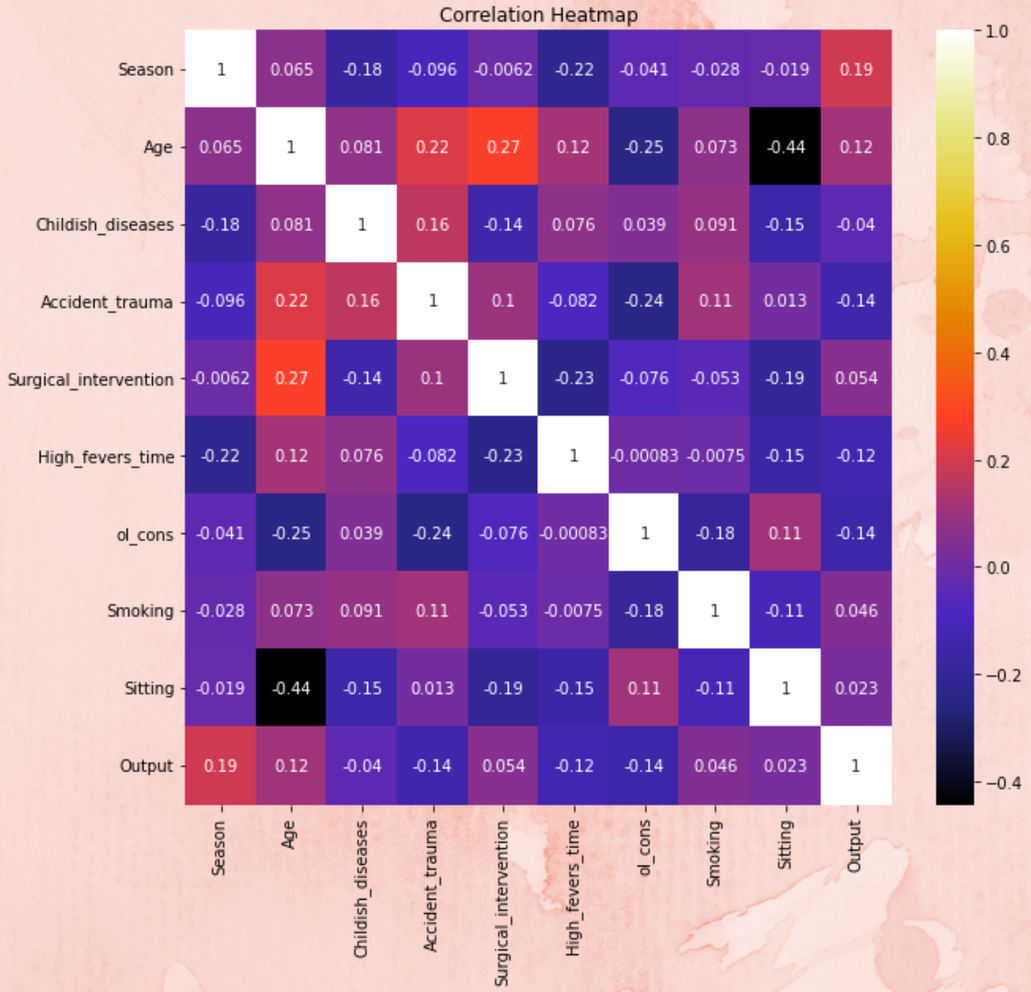


PIE PLOT



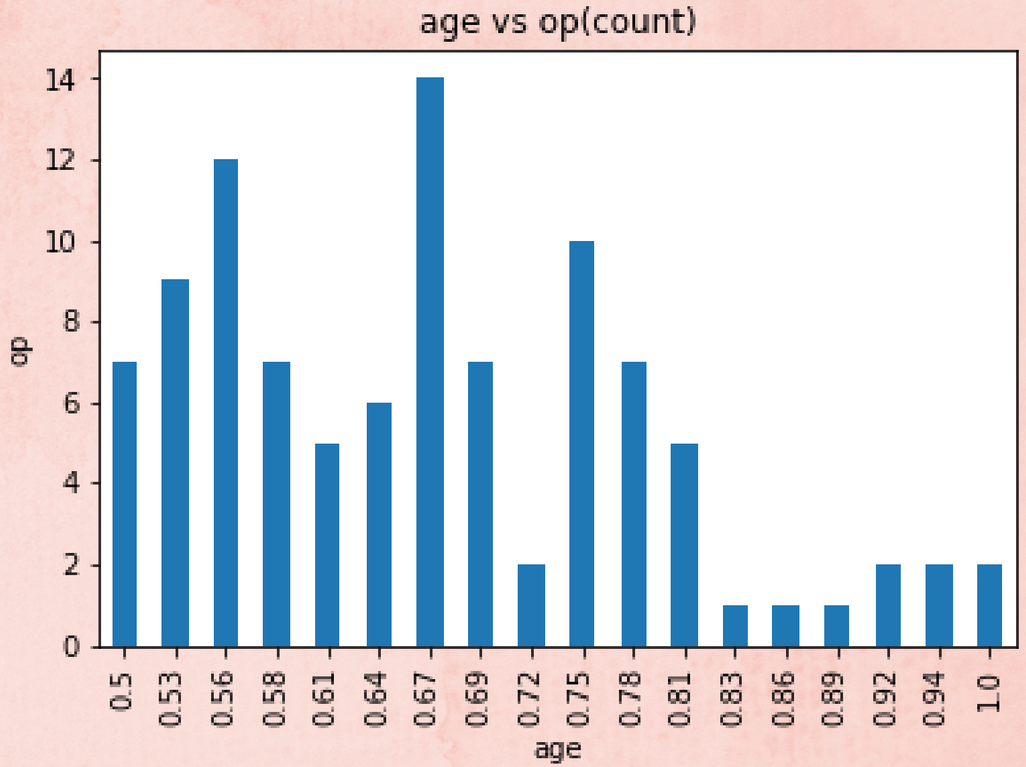
The pie plot shows the percent of normal and alter diagnosis

HEAT MAP



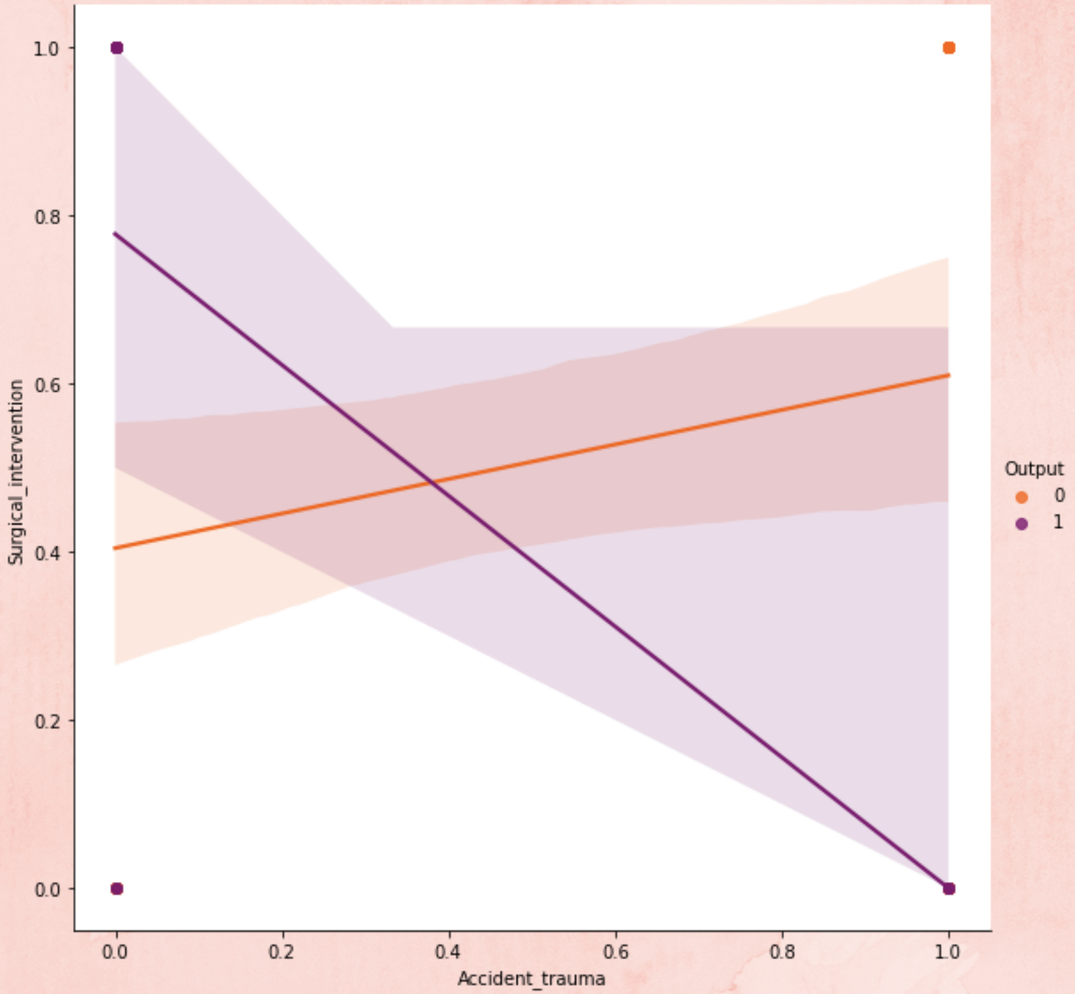
Season, age are highly positively correlated with target variable. Childish_diseases,accident trauma, high_fever_time, ol_cons are highly negatively correlated with target variable.

BAR GRAPH



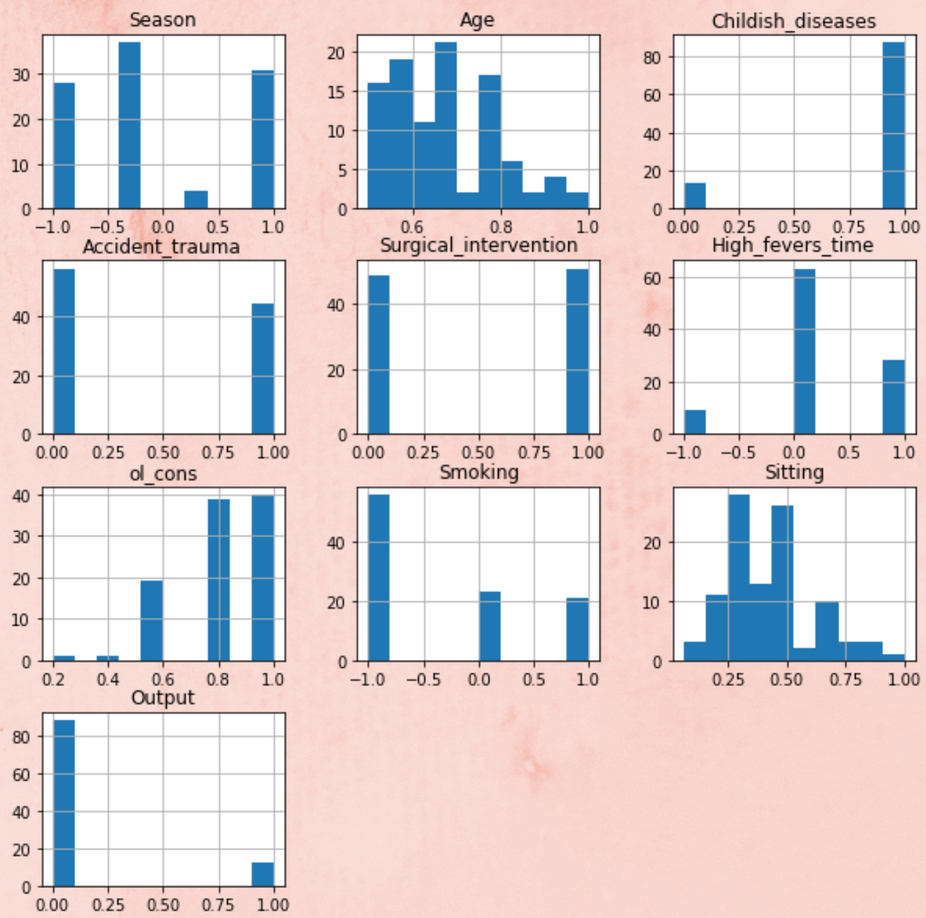
The bar plot shows the relation of age vs target variable.

LM PLOT



to draw a scatter plot onto a FacetGrid

HISTOGRAM

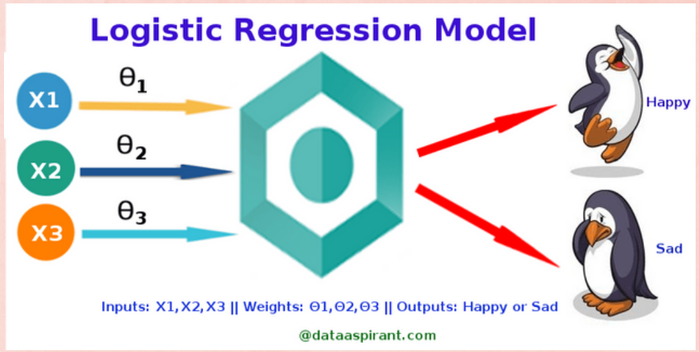


ML ALGORITHM : LOGISTIC REGRESSION

TRAIN TEST RATIO	ACCURACY
70-30	90%
80-20	90%
75-25	88%
60-40	90%

<---- Best ratio: 70-30

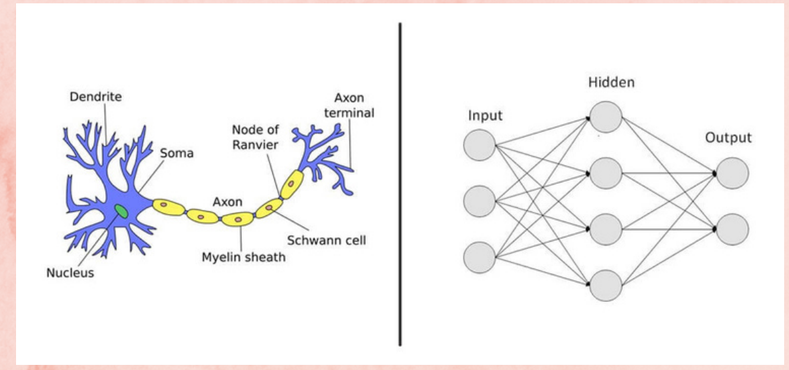
We applied variuos train test proportions and obtained the accuracy to be between 90-100%



ML ALGORITHM : NEURAL NETWORK

Sample of Experiments worked on

TRAIN TEST RATIO	ARCHITECTURE	EPOCHS	ACCURACY
75-25	9-12-6-3-1	50	0.88
80-20	9-8-4-7-1	100	0.95
80-20	9-5-6-8-1	125	0.9
70-30	9-8-8-5-1	50	0.8
80-20	9-7-6-5-1	75	0.9
60-40	9-6-7-2-1	75	0.825



<--- On working with different train test splits and no.of Architectures highest Accuracy obtained was for the ratio-

TRAIN TEST RATIO	ARCHITECTURE	EPOCHS	ACCURACY
80-20	9-5-5-5-1	70	0.9
80-20	9-2-3-4-1	80	0.9
80-20	9-3-4-6-1	90	0.9
70-30	9-5-6-7-1	75	0.933
70-30	9-7-8-9-1	105	0.833
70-30	9-10-1-2-1	205	0.9

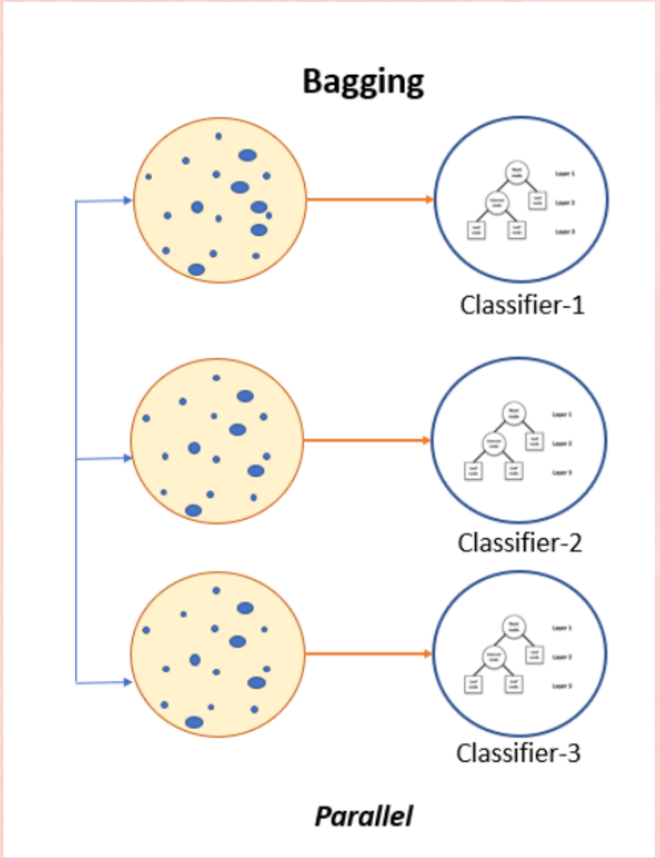
TRAIN TEST RATIO	ARCHITECTURE	EPOCHS	ACCURACY
75-25	9-8-7-2-1	100	0.88
75-25	9-9-8-7-1	200	0.88
75-25	9-6-5-3-1	300	0.88
75-25	9-7-5-2-1	50	0.92
75-25	9-2-3-6-1	150	0.88
75-25	9-9-9-9-1	70	0.84

TRAIN TEST RATIO	ARCHITECTURE	EPOCHS	ACCURACY
75-25	9-3-2-8-1	150	0.88
75-25	9-12-5-6-1	320	0.92
75-25	9-8-6-9-1	70	0.92
75-25	9-4-5-3-1	80	0.88
75-25	9-3-2-9-1	90	0.88
75-25	9-1-1-1-1	100	0.88

TRAIN TEST RATIO	ARCHITECTURE	EPOCHS	ACCURACY
75-25	9-2-3-11-1	105	0.88
80-20	9-7-5-2-1	50	0.9
80-20	9-2-3-6-1	150	0.9
80-20	9-9-9-9-1	70	0.9
80-20	9-3-2-8-1	150	0.9
80-20	9-12-5-6-1	320	0.9

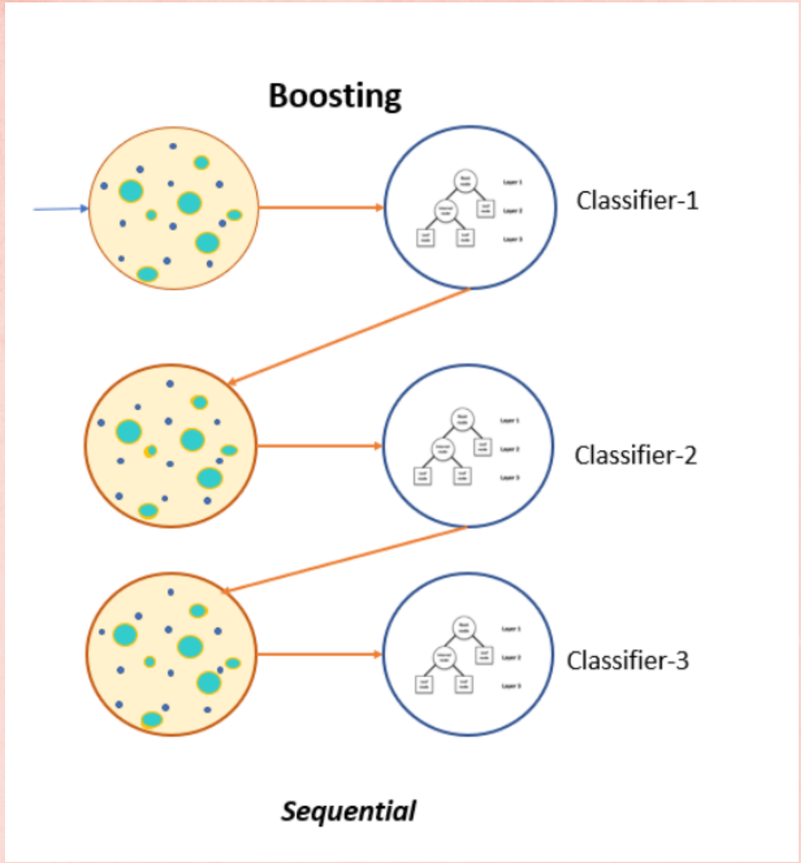
ML ALGORITHM : BAGGING

TRAIN TEST RATIO	ACCURACY	MODEL SCORE
80-20	0.867	0.88
75-25	0.834	0.84
70-30	0.867	0.84
60-40	0.853	0.88
65-35	0.854	0.88



ML ALGORITHM : BOOSTING

CLASSIFIERS	ACCURACY
ADA BOOST	90.0%
XG BOOST	90.0%
GRADIENT BOOST	97.0%



KNN CLASSIFIER

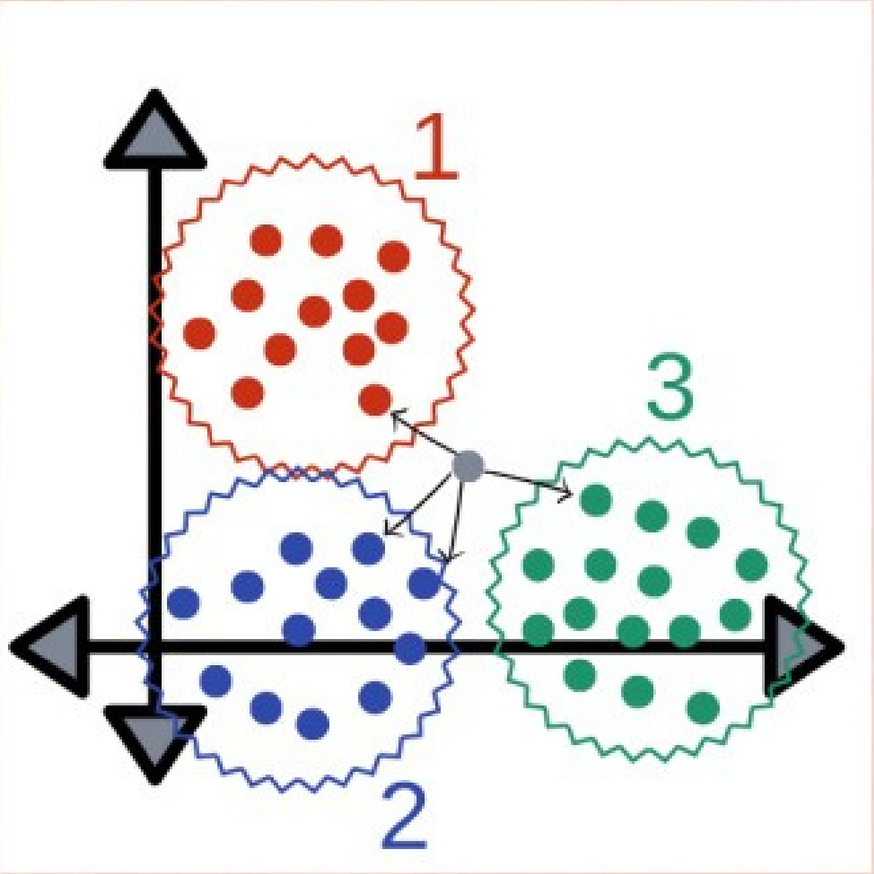
Confusion Matrix:

```
[[19  2]
 [ 0 19]]
```

Classification Report:

	precision	recall	f1-score	support
0	1.00	0.90	0.95	21
1	0.90	1.00	0.95	19
accuracy			0.95	40
macro avg	0.95	0.95	0.95	40
weighted avg	0.95	0.95	0.95	40

Accuracy: 0.95



SVM(Support Vector Machine)

Kernel as a Linear function

- SVM with respect to Linear as kernel
- Trained and Tested the model
- Confusion matrix
- Accuracy obtained = 0.566

MODELS	ACCURACY
GRADIENT BOOSTING	0.97
NEURAL NETWORKS	0.95
KNN	0.95
ADABOOST	0.9
XG BOOST	0.9
LOGISTIC REGRESSION	0.9
RANDOM FOREST	0.89
BAGGING	0.867
DECISION CLASSIFIER	0.72
SVM	0.56

CONCLUSION :

Performing all 4 Machine learning algorithms to the given dataset and analysing through exploratory data analysis we can conclude that Gradient boosting is the best model for given dataset with 97% of accuracy for 80-20 ratio.

THANKYOU

Team Members:

A.AKHILA
CH.RAHUL
R.MITALI
K.BHANU PRASAD REDDY
V.NANDINI

APPENDIX

Google Colaboratory
colab.research.google.com

(Logistic regression,Neural Networks)

Google Colaboratory
colab.research.google.com

(Decision trees,Random forests,Bagging,Boosting,KNN,SVM)