MACHINE LEARNING

ASSIGNMENT 1

“TRANSFUSION DATA ANALYSIS”

1. Using Logistic Regression
    - Gradient Descent
    - Newton’s Method

2. Using Naive Bayes Classification

1. Task Introduction

- We have been provided with the Transfusion dataset that contains data samples of various donors.

- Our task is to classify that wether the donor donated the blood in march 2017 provided other data columns.

- In all, Dataset has 4 features and our assumptions are that these 4 features followes normal distribution, samples are Independently and identically distributed (IID) and that the features are linearly related to the target.

- Following all these assumptions, we are ready to use Logistic Regression to begin our analysis of this linear relationship that persists in our dataset throughout all samples.

2. Description the data and Preprocessing

- Pipeline of our preprocessing will be as follows.

    1. Look for duplicate rows and remove them.

    2. Look for potential null/nan values and use appropriate strategy to treat them.

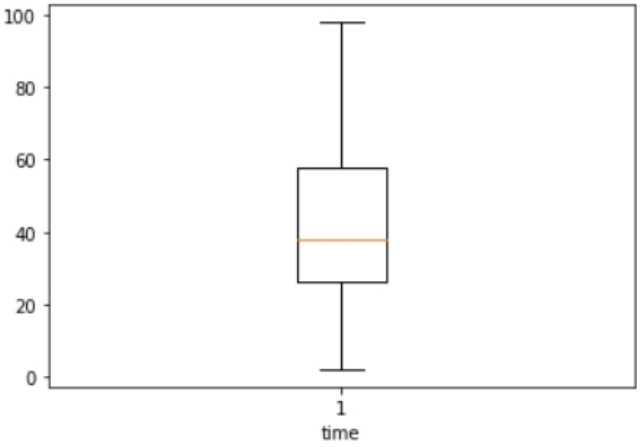
    3. Look for non adherent outliears and remove them or replace them with boundary values.

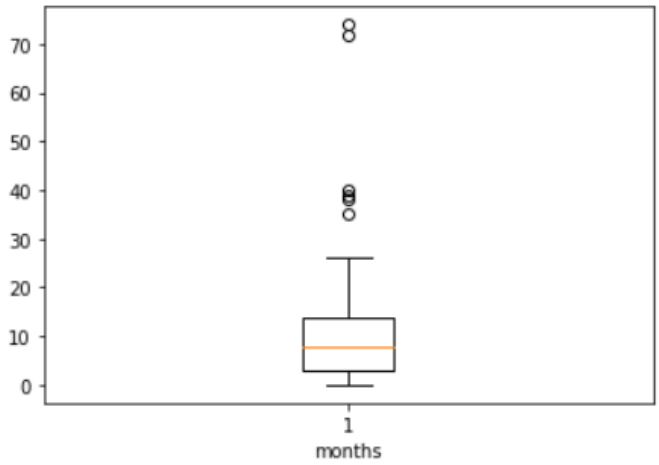
    4. Correlation of analysis of data columns with the target column and other features.

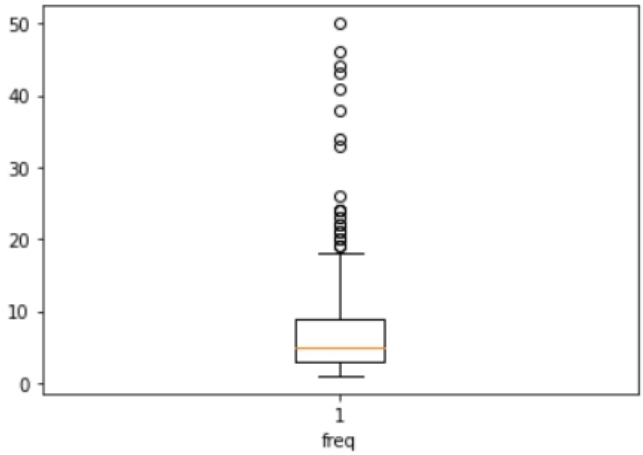
    5. Standardisation of data values to map them to standard scale.

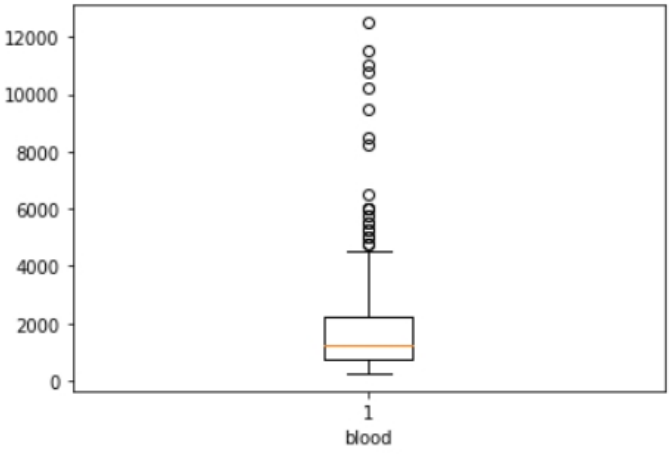
- Data can be described in various ways, We primarily be focussing on Box-Plots and Correlation matrix as our data is simple and numerically/contigously distributed.

➢ Box plots of features having potential outliers









- Statistics for Dataset.

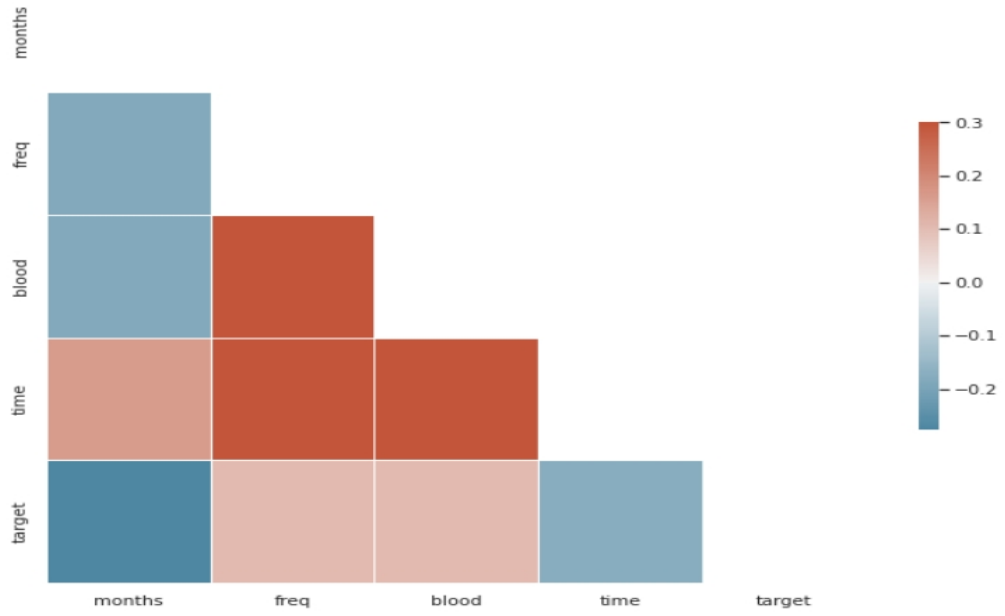
	months	freq	blood	time	target
count	748.000000	748.000000	748.000000	748.000000	748.000000
mean	9.506684	5.514706	1378.676471	34.282086	0.237968
std	8.095396	5.839307	1459.826781	24.376714	0.426124
min	0.000000	1.000000	250.000000	2.000000	0.000000
25%	2.750000	2.000000	500.000000	16.000000	0.000000
50%	7.000000	4.000000	1000.000000	28.000000	0.000000
75%	14.000000	7.000000	1750.000000	50.000000	0.000000
max	74.000000	50.000000	12500.000000	98.000000	1.000000

Here we see the some data columns has values that are more than 3 times the standard deviation from the mean.

Two most promising techniques are, To remove the entire row containing the non adherent outlying data and To replace the data column value with the boundry value i.e. the 3rd quantile data value of the respective column. We are using the former one, as the data samples are not sparse and data shows no sensitivity with minimal loss of samples.

Correlation Analysis:

- With reference to “Statistical Learning by Trevor – Stanford publication” There should be minimum correlation between features and quite a large correlation with the regression target.
- Reason for former is to explain as much variance in regression target, We need to have diversified and uncorrelated features as extreme correlation between features doesn’t account any additional info to explain the target, perhaps only make regression task harder by worsessing the over-fitting and increasing complexity of model.



- Fortunately, Our correlation matrix outrages all the flaws concluding from high positive/negative intra-sample correlation.

➢ Dealing with Null/Na/Nan values

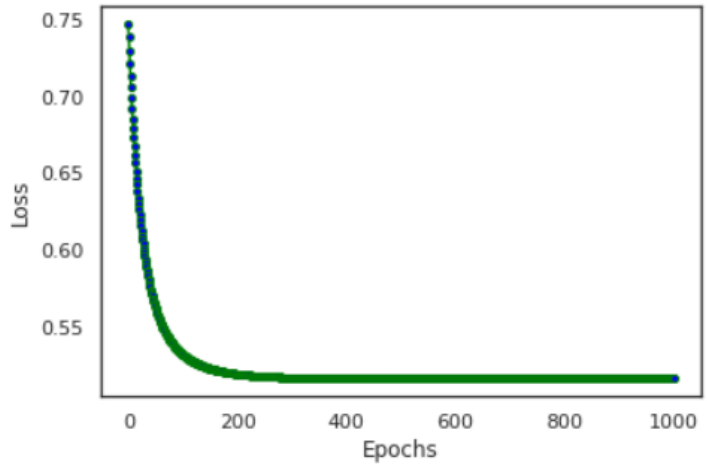
- Following figure depicts the amount of null values consitituted in our dataset.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 748 entries, 0 to 747
Data columns (total 5 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   months  748 non-null    int64
 1   freq    748 non-null    int64
 2   blood   748 non-null    int64
 3   time    748 non-null    int64
 4   target  748 non-null    int64
dtypes: int64(5)
memory usage: 29.3 KB
```

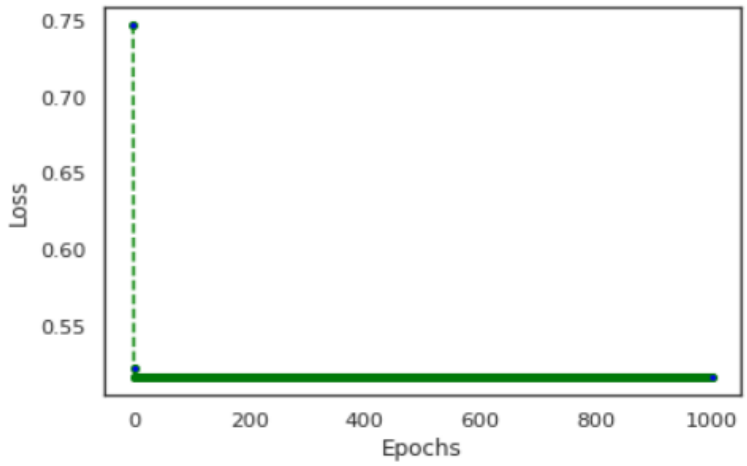
We see that no columns have null values.

➢ Training Model and Analysis of results

```
f1 score : [74.90196078]
```



➢ Training with Newton's optimisation method

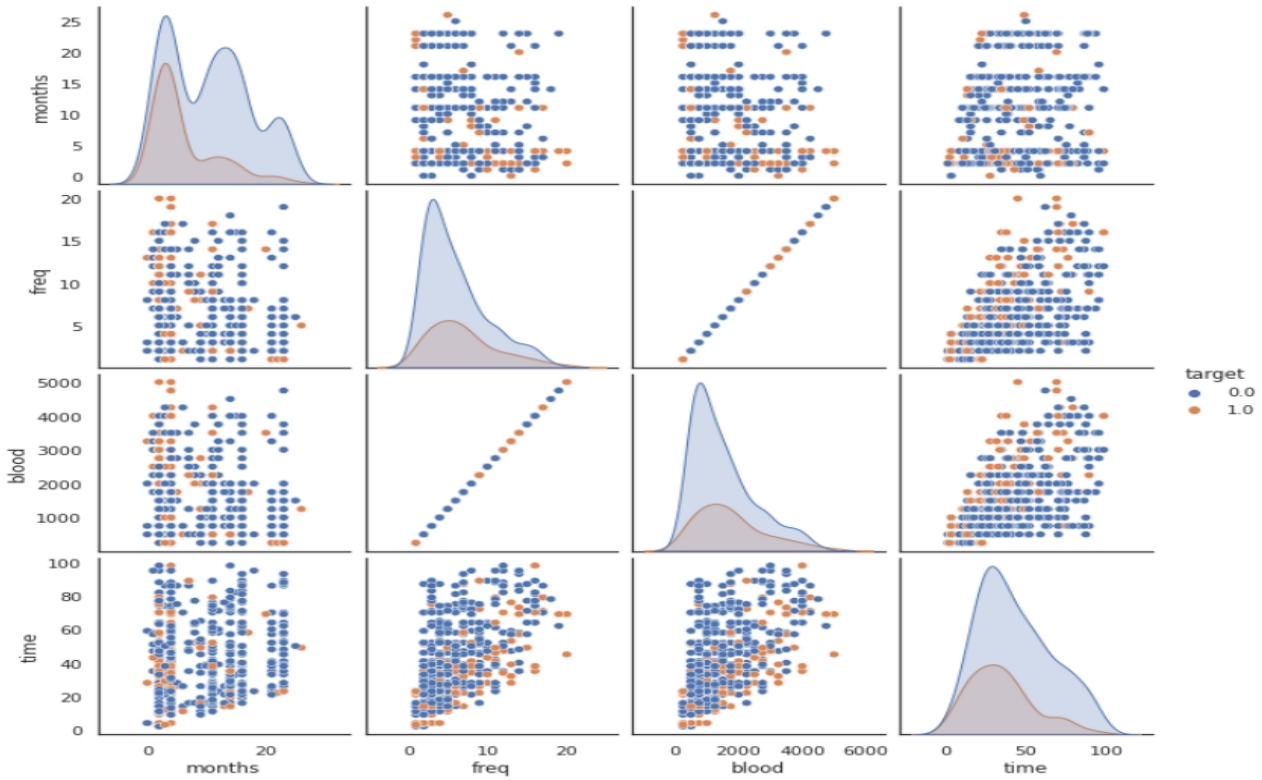


Naive Bayes Classification.

```
F1 score is 72.74509803921569
TRAINING ON ALL FEATURES


UNIVARIATE STARTS NOW!
TRAINING ON months
F1 score is 72.74509803921569
TRAINING ON freq
F1 score is 72.74509803921569
TRAINING ON blood
```

That’s the best we can achieve using all variables at once, but how to gain more?Well after looking at below figure:



We found that vairables that can be best modelled with MVG are freq_of_blood_donations, time_since_last_donation and total_blood_donated.

F1_score was really improved and we got the following results.

```
print("F1 score is " + str(nbc.f1_score(nbc.predict(X_train.to_numpy()), y_train)));
```

```
F1 score is 74.90196078431373
```

Here, we end our analysis of transfusion dataset.

For Reader,

After going through the python notebook and this report describing analysis process, if you really found this analysis intuitive and worth appreciating, I would say this task couldn't be accomplished without immense efforts made by our T.A. Shourabh Payal and Prof. Dinesh Babu.

Deepak Nandwani
MT2021037
Mtech. CSE