Dengue Shock Syndrome prediction using ML/AI

Dr. Abhijit Ray
Mr. Padmanava Rudra

Introduction : Dengue Fever is a mosquito-borne viral fever caused by the dengue virus. Though mostly prevalent in the tropical countries, Dengue Fever has a global annual incidence of approximately 390 million cases per year, of which around 500,000 develop severe illness leading to almost 25,000 deaths. Deaths from Dengue Fever occur mostly as a result of hemorrhagic shock due to thrombocytopenia-induced bleeding. Though some risk factors such as a recurrent infections, low base-level platelet count etc have been linked to severe diseases, most cases of Dengue Hemorrhagic Shock have no apparent pre-existing cause.

Dengue Fever : Dengue is the most common arthropod-borne viral disease in man. Dengue fever can be asymptomatic in some cases. However, the most common symptoms include high-grade fever, headache, stomach ache, rash, myalgia, and arthralgia. Due to the severity of the arthralgia, Dengue fever is often called 'Break-Bone Fever'. While the majority of cases are self-limiting, some patients develop Dengue Hemorrhagic Shock, which are often fatal.
Despite intensive research, the underlying mechanisms causing severe dengue is still not well understood. Even though it is clear that both viral and host factors play important roles in the course of infection, a fundamental knowledge gap still remains to be filled regarding host cell tropism, crucial host immune response mechanisms, and viral markers for virulence.

Dengue Virus : The dengue virus is a single positive-stranded RNA virus of the family Flaviviridae; genus Flavivirus. It is spread by mosquito of Aedes sp. The Dengue Virus has 4 serotypes (DENV 1 to 4). All four DENV serotypes have emerged from sylvatic strains in the forests of South-East Asia. Difference in clinical pictures have been reported across the different serotypes. Cases infected with DENV-1 were more likely to present with red eyes whereas presence of joint pain and lower platelet count was associated with DENV-2 cases. However, DENV-1 was associated with a greater risk of Dengue Hemorrhagic Fever than DENV-2.
The lack of specific antivirals and vaccines have made control of Aedes mosquito the only measure of prevention against Dengue Fever. Different of mosquito control has been tried across the globe with varying rates of success. Some of these methods include:

- Chemical interventions : Though a wide range of chemicals are used as insecticides, organophosphates and pyrethroids (like Permethrin) are mainly used against *Aedes* spp. mosquitoes. They act on both larval and adult stages of the mosquito life cycle. Bacillus thuringiensis israelensis, a gram-positive, spore-forming bacterium is pathogenic to mosquitoes and thus used as a chemical larviciding agent.

- Habitat management : A combination of methods, such as grassroots community involvement, coordination of public health systems, health education, and punitive measures in vector control have been used by many countries. However, by nature, *Aedes* mosquitoes prefer both urban and semi-urban areas. In urban areas, breeding places are difficult to eliminate due to overcrowding, pollution, suboptimal waste water drainage, and inappropriate garbage and waste disposal. In semi-urban areas, there are plenty of natural water collection sites that are also difficult to eliminate.

- Non-chemical larviciding : Non-chemical methods, such as oil coating and vector trapping, interrupt the life cycle of the vector, thus limiting propagation. Non-chemical methods, such as oil coating and vector trapping, interrupt the life cycle of the vector, thus limiting propagation. *Wolbachia* spp. are used as a population control method for the *Aedes* mosquito. *Wolbachia* is an intracellular, maternally-inherited, endosymbiotic bacteria found in insects. *Wolbachia* causes reproductive modifications, such as cytoplasmic incompatibility, which results in the generation of unviable offspring when an uninfected female mates with a *Wolbachia*-infected male.

- Genetic techniques : Genetic manipulation comprises several techniques for vector control. The sterile insect technique involves creation and release of sterile males into the environment, reducing the reproductive potential of the target wild population thus eventually reducing the population size. Sterility is induced traditionally by radiation and lately by genetic manipulation. These methods cause death of mosquito offspring following mating.

Dengue Shock Syndrome : There are two schools of thought regarding the pathophysiology of Dengue Hemorrhagic Fever (DHF) and Dengue Shock Syndrome (DSS). One theory suggests more virulent strains of the dengue virus are the main cause behind DHF/DSS. According to the other theory, DHF/DSS results from abnormal and exaggerated host immune responses; in particular, the production of dengue virus cross-reactive antibodies – which augments the infection. In primary infection with the dengue virus, cross-reactive antibodies that lack neutralising activity are produced. During secondary infection by a different serotype, the dengue virus and non-neutralising antibodies form virus–antibody complexes. The Fc portion of these antibodies bind to FcγRI- and FcγRII-bearing cells, resulting in an increased number of cells being infected by the dengue virus. This phenomenon is known as antibody-dependent enhancement and is believed to play an important part in the pathogenesis of shock.

Deaths from Severe Dengue occur mostly as a result of progressively worsening shock and multi-organ failure. Though the exact mechanism is not fully understood, increased vascular permeability is thought to occur largely due to malfunction of vascular endothelial cells induced by cytokines or chemical mediators.

Cytokine	Response	Duration
TNF-α, IL-2, IL-6, and IFN-γ	Th1	First 3 days of infection
IL-10, IL-5, and IL-4	Th2	After the first 3 days

Table 1 : Cytokines in DHF

In presence of enhancing antibodies, monocytes infected with dengue virus produce TNF-α, which induces plasma leakage7. Dengue virus has also been reported to infect endothelial cells and cause direct damage through apoptosis8. Dengue infections are associated with reduced numbers of CD4+ T-cells, CD8+ T-cells, and natural killer cells. Levels of these cells are lowest at the point when the fever settles and the onset of shock takes place, and increase subsequently.

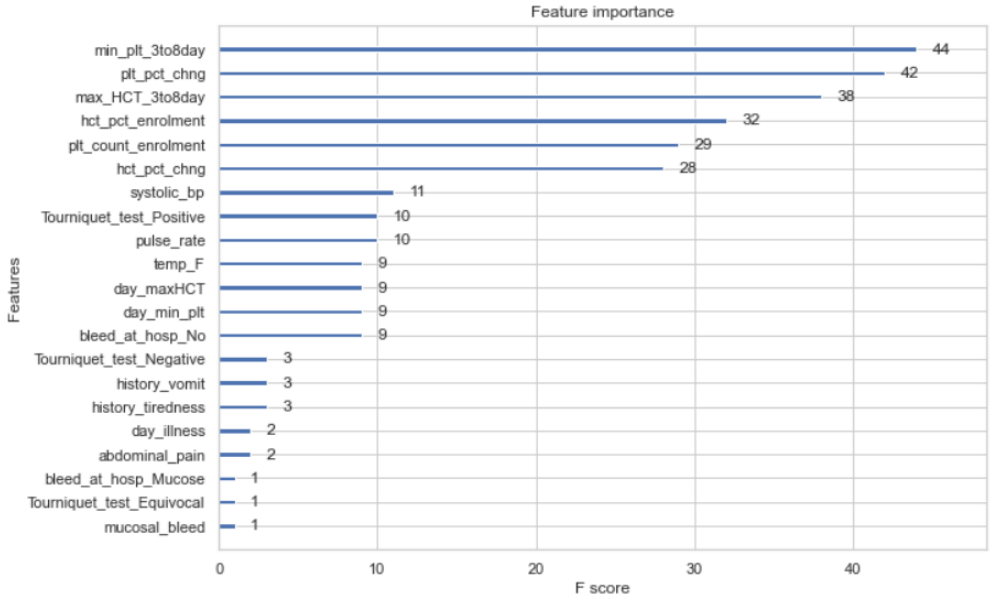
Thrombocytopenia : The cause of thrombocytopenia in Dengue Fever is not very well understood. It is believed that DENV directly or indirectly affects bone marrow progenitor cells by inhibiting their function to reduce the proliferative capacity of hematopoietic cells. DENV can induce bone marrow hypoplasia during the acute phase of the disease. Significant deregulation of the plasma kinin system due to the immunopathogenesis of dengue (probably cytokine production) has been reported. In addition, DENV infection induces platelet consumption due to disseminated intravascular coagulation (DIC), platelet destruction due to increased apoptosis, lysis by the complement system and by the involvement of anti-platelet antibodies.

As is evident from the above text, no clear cause-effect relationship could be established between Dengue Fever and thrombocytopenia leading to DHF/DSS. Thus, it has previously been almost impossible for clinicians across the world to interpret which patients were most at risk of DSS. Due to this lack of interpretation, the diagnosis of DSS in Dengue Fever is often too late. By the time clinicians receive reports from the lab showing the extremely low platelet count, the patient may be in a very critical state.

Aim : A study was conducted to observe the relationship between daily platelet counts and Hematocrit (HCT) percentage during treatment days to predict DSS by a supervised machine learning model.
After initial cleansing, total patient as recorded in the dataset is 2296 and DSS was attacked to 141 patients.
The dataset contains observational data of following information
Variables:
- st_no : Patient study number
- age : Age at enrolment (year)
- sex : Gender (Female, Male)
- wt : Weight at enrolment (kg)
- day_ill : Day of illness at enrolment
- his_tired : History of tiredness at enrolment (Yes, No)
- his_vomit : History of vomiting at enrolment (Yes, No)
- ttest : Tourniquet test result at enrolment (Positive, Equivocal, Negative)
- temp : Temperature at enrolment (0C)
- pulse : Pulse rate at enrolment (count per minute)
- sys_bp : Systolic blood pressure at enrolment (mmHg)
- mucosal_bleed : Mucosal bleeding at enrolment (Yes, No)
- abdominal_pain : Abdominal pain at enrolment (Yes, No)
- liver : Liver size at enrolment (cm)
- hct_bsl : Haematocrit level at enrolment (%)
- plt_bsl : Platelet count at enrolment (cells/mm3)
- serotype2 : Serotype determined by PCR (DENV-1, DENV-2, DENV-3, DENV-4, Mixed, Negative)
- serology : Immune status determined by ELISA (Secondary dengue, Primary dengue, Possible primary, Unclassifiable)
- to_PICU : Referred to PICU (Yes, No)
- shock : Dengue shock syndrome (Yes, No)
- doi_shock : Day of illness at shock (day)
- bleed_hos : Bleeding during hospitalization (No, Skin, Mucose, Other)
- minPLT_3to8 : Platelet nadir (cells/mm3)
- dminPLT_3to8 : Day of illness of platelet nadir (day)
- maxHCT_3to8 : Maximum haematocrit (%)
- dmaxHCT_3to8 : Day of illness of maximum haematocrit (day)
- maxhemo_3to8 : Overall haemoconcentration (%)

Method : Logistic regression was used as the main statistical model for all univariate and multivariable analyses. The clinical data of daily Hematocrit (HCT) levels and platelet counts were assessed during treatment days and multiple classifier models were fitted with data and best accuracy was achieved by the classifier LightGBM.

Statistical Analysis : Initial statistical analysis is done and after feature importance following result is found.



After analysis of feature importance on predicted variable (DSS) following features are considered in the final dataset.

```
In [8]: data.describe().T
```

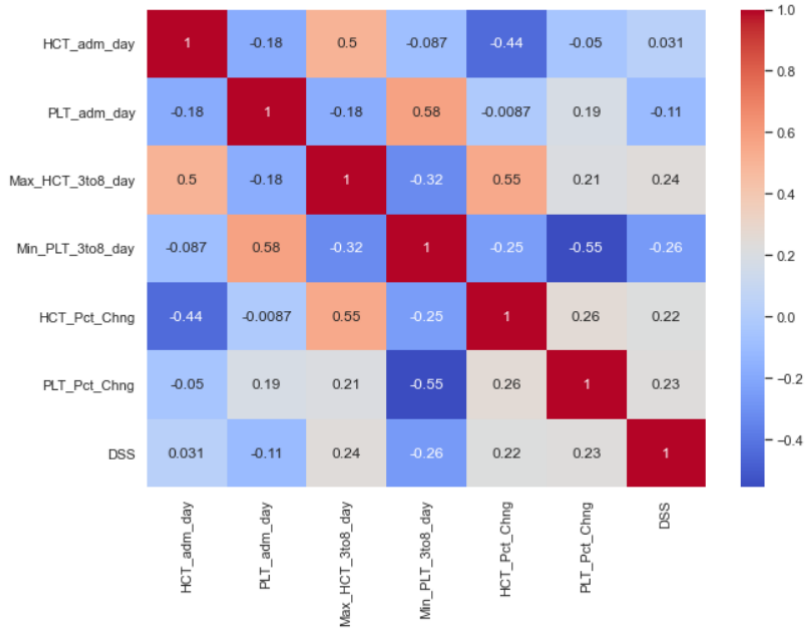
Out[8]:

	count	mean	std	min	25%	50%	75%	max
HCT_adm_day	2296.00	39.76	3.96	23.10	37.27	39.50	42.00	64.00
PLT_adm_day	2296.00	141945.42	61208.77	7800.00	98000.00	136000.00	177000.00	429000.00
Max_HCT_3to8_day	2296.00	44.32	4.65	30.30	41.00	44.00	47.00	61.00
Min_PLT_3to8_day	2296.00	75283.04	47150.78	6050.00	40775.00	66000.00	99000.00	379000.00
HCT_Pct_Chng	2296.00	11.98	11.68	-25.93	2.83	9.77	18.16	84.85
PLT_Pct_Chng	2296.00	44.56	30.03	-670.56	25.45	47.43	65.41	95.51
DSS	2296.00	0.06	0.24	0.00	0.00	0.00	0.00	1.00

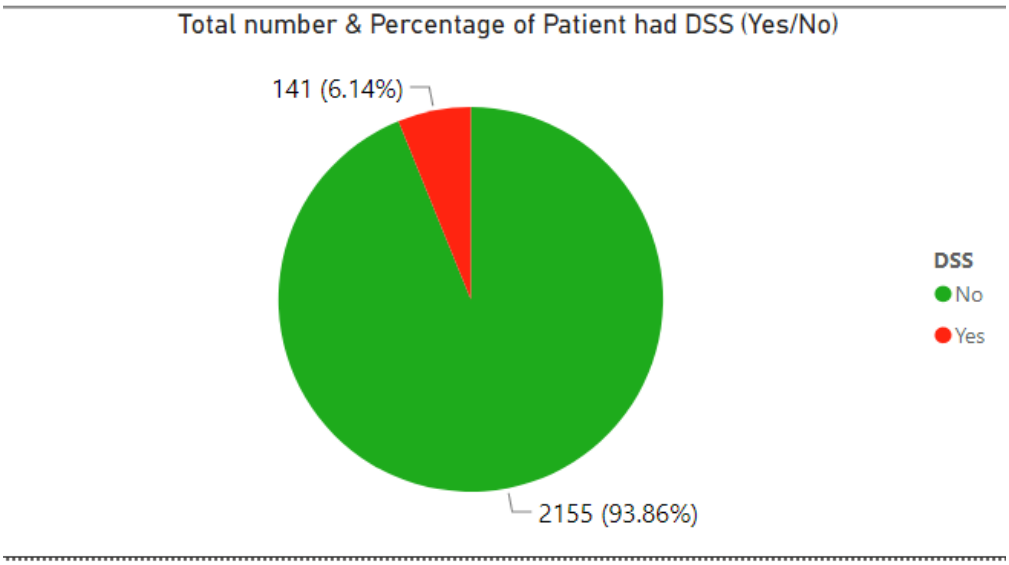
hct_bsl -> renamed as HCT_adm_day
plt_bsl -> renamed as PLT_adm_day
maxHCT_3to8 -> renamed as Max_HCT_3to8_day
minPLT_3to8 -> renamed as Min_PLT_3to8_day
HCT_Pct_Chng -> The percentage of change in HCT from admission day to maximum HCT observed between 3 to 8 day of treatment.
PLT_Pct_Chng -> The percentage of change in PLT count from admission day to minimum PLT count observed between 3 to 8 day of treatment.
Shock -> renamed as DSS

Feature co-relation matrix is drawn as :

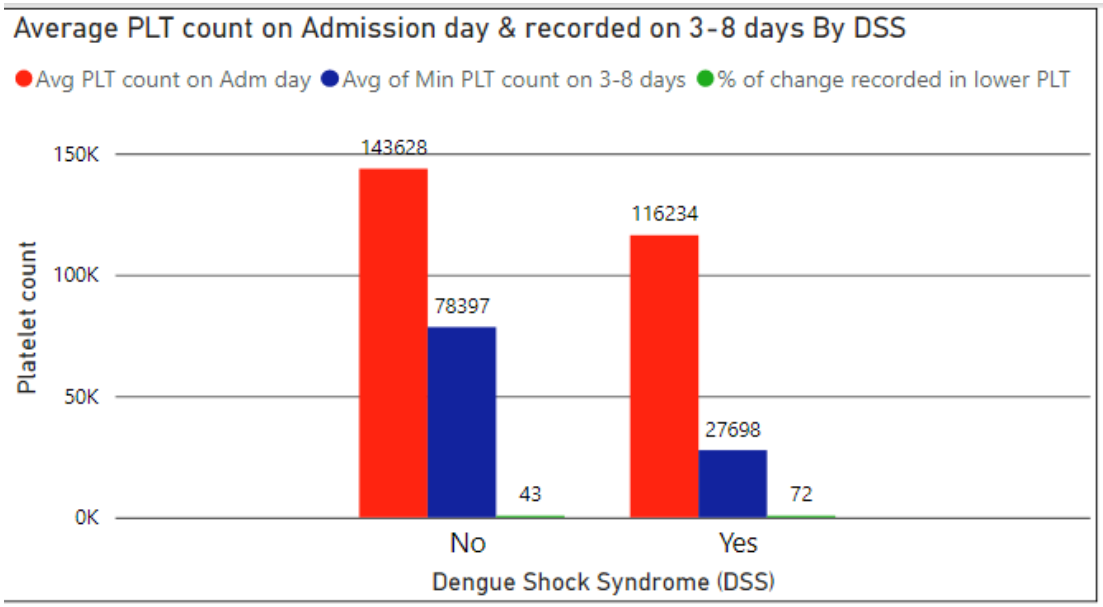
Out[34]: <AxesSubplot:>



Ratio of Patients had DSS (Yes/No)



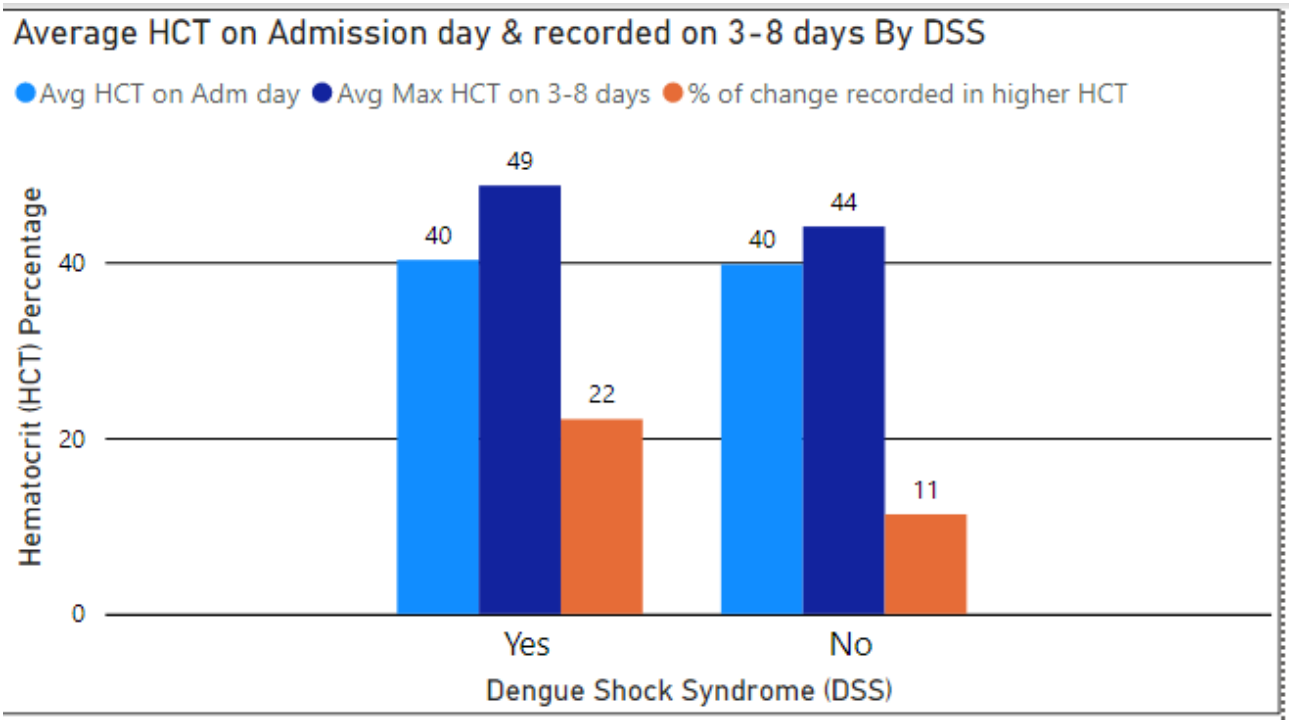
Platelet (PLT) count recorded on admission and 3-8 days of treatment



Observation :

- It is observed that platelet (PLT) count for patient who were suffered in DSS had gone below to an average of 72% and average minimum platelet count was observed as 27K
- Dengue patient who were not suffered from DSS, the platelet count gone below to an average of 43% and average minimum platelet count was observed as 78K
- Falling of platelet count during 3-8 days of treatment, below 30K may lead to DSS

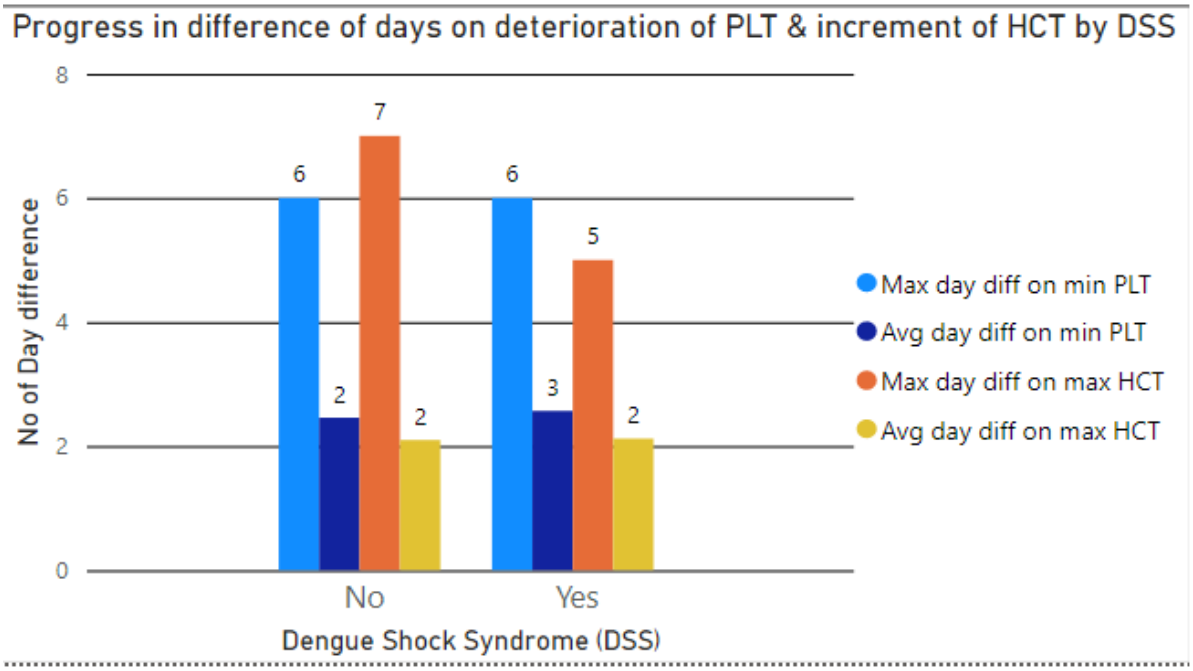
Hematocrit(HCT) recorded on admission and 3-8 days of treatment



Observation:

- It is observed that percentage in increment of HCT for patient who were suffered in DSS had gone up to an average of 22% and average maximum percentage of HCT was observed as 49%
- Dengue patient who were not suffered from DSS, the change in increment of HCT percentage were found up to average 11% and average maximum HCT percentage was observed as 44%
- Change in increment of certain HCT percentage during 3-8 days of treatment, may lead to DSS

Progression in deterioration of Platelet count and increment of HCT percentage during 3-8 days of treatment



Observation:

- Maximum difference in days found to trace minimum platelet count was 6 days and average difference was 3 days and for HCT increment it was 5 and 2 days respectively for DSS suffered patient.
- Maximum difference in days found to trace minimum platelet count was 6 days and average difference was 2 days and for HCT increment it was 7 and 2 days respectively for patient who were not suffered DSS
- Change in platelet count and HCT percentage in progression of treatment is an important observation to assess possibility of DSS. It is found that day 4 is a critical day to observe the patient's platelet count and HCT percentage which may lead to DSS

Preparation of Model:

After initial cleaning and preprocessing, data set is split into 70:30 for training and test set by using sci-kit learn package and then fitted into classifier pipeline using Lazypredict package with 27 classifier models. Best results are found in following classifier models

```
In [109]: print(models)
```

```
                                    Accuracy  Balanced Accuracy  ROC AUC  F1 Score  \
Model
ExtraTreesClassifier                    0.97               0.97     0.97      0.97
XGBClassifier                           0.97               0.97     0.97      0.97
RandomForestClassifier                  0.97               0.97     0.97      0.97
LGBMClassifier                          0.97               0.97     0.97      0.97
AdaBoostClassifier                      0.97               0.97     0.97      0.97
QuadraticDiscriminantAnalysis           0.97               0.97     0.97      0.97
SVC                                     0.97               0.97     0.97      0.97
```

Training data set and result with LGBMClassifier is as below

```
In [47]: X_train, X_test, y_train, y_test = train_test_split(x_smote,y_smote, test_size=0.30,random_state=42)
```

```
In [44]: import lightgbm as lgb
         clf = lgb.LGBMClassifier()
         clf.fit(X_train, y_train)
```

```
Out[44]: LGBMClassifier()
```

```
In [46]: clf.score(X_test,y_test)
```

```
Out[46]: 0.9158200290275762
```

```
In [47]: print(classification_report(y_test, lr.predict(X_test)))
```

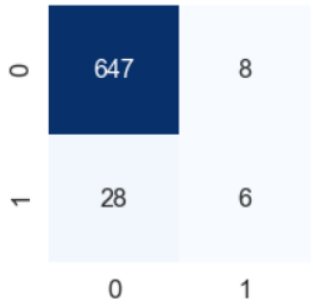
```
              precision    recall  f1-score   support

           0       0.96      0.99      0.97       655
           1       0.43      0.18      0.25        34

    accuracy                           0.95       689
   macro avg       0.69      0.58      0.61       689
weighted avg       0.93      0.95      0.94       689
```

Accuracy of the model is shown as 91.5 % which is a good score but the accuracy is biased on Non DSS class as clearly evidenced from the confusion matrix as below. The reason behind is that predicted variable DSS is by nature a binary class (Yes/No) with imbalance number of observations (Non DSS = 2155 observations as the majority class and DSS = 141 observations as the minority class).
For any classification model observation of predicted class distribution should be balance so that in training model should not be biased on majority class (Non DSS) over minority class (DSS).

```
In [43]: #Calling function
         conf_matrix(y_test, pred_test)
```



To resolve this problem there is a technique available called SMOTE (Synthetic Minority Oversampling Technique). In this technique, minority class is synthetically oversampled to maintain the balance of predicted class for proper training in the model.
SMOTE works by selecting examples that are close in the feature space, drawing a line between the examples in the feature space and drawing a new sample at a point along that line.

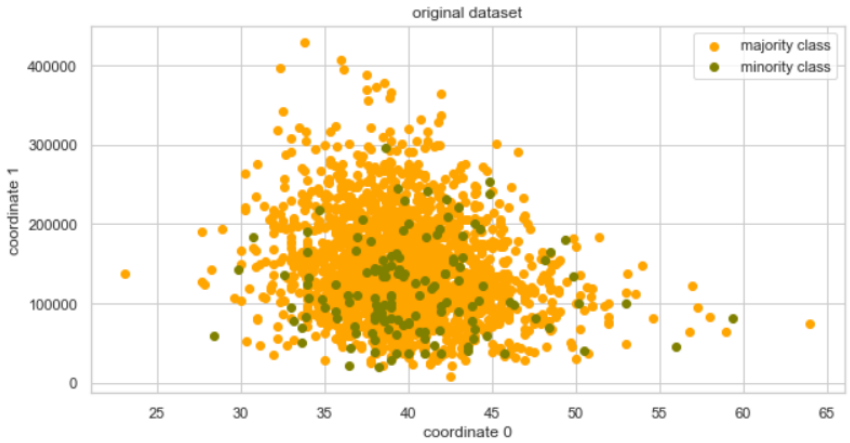
SMOTE first selects a minority class instance A at random and finds its k nearest minority class neighbors. The synthetic instance is then created by choosing one of the k nearest neighbors B at

random and connecting A and B to form a line segment in the feature space. The synthetic instances are generated as a convex combination of the two chosen instances A and B. There are several SMOTE variants available. To find out the best suited SMOTE variant, **smote_variants** package is used and finally **ASMOBD** variant is chosen to get the optimum result.

Class distribution before using SMOTE technique is as below

In [17]:
```
plt.figure(figsize=(10, 5))
plt.scatter(X[y == 0][:,0], X[y == 0][:,1], label='majority class', c='orange')
plt.scatter(X[y == 1][:,0], X[y == 1][:,1], label='minority class', c='olive')
plt.title('original dataset')
plt.xlabel('coordinate 0')
plt.ylabel('coordinate 1')
plt.legend()
```

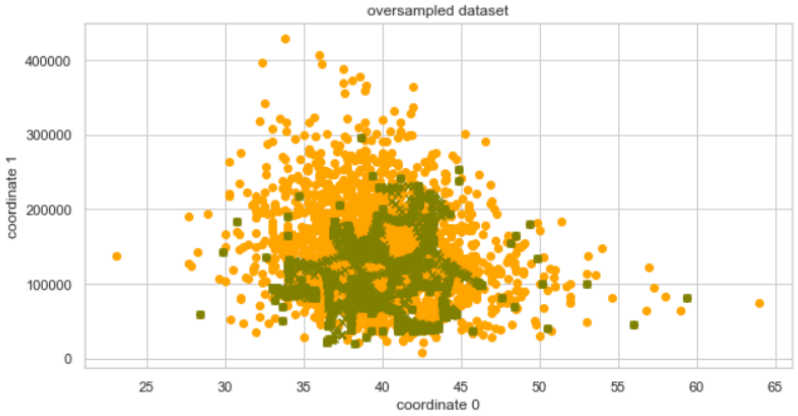
Out[17]: <matplotlib.legend.Legend at 0x1b79f84c6a0>



After using the SMOTE variant ASMOBD class distribution is found as below

In [19]:
```
plt.figure(figsize=(10, 5))

plt.scatter(X[y == 0][:,0], X[y == 0][:,1], label='majority class', c='orange', marker='o')
plt.scatter(X[y == 1][:,0], X[y == 1][:,1], label='minority class', c='olive', marker='o')
plt.scatter(X_samp[y_samp == 1][:,0], X_samp[y_samp == 1][:,1], label='new minority samples', c='olive', marker='x')
plt.title('oversampled dataset')
plt.xlabel('coordinate 0')
plt.ylabel('coordinate 1')
plt.show()
```



Model Training

After SMOTE distribution of data set, model is trained with **LGBMClassifier** as below steps

In [22]:
```
# build the lightgbm model
import lightgbm as lgb
clf = lgb.LGBMClassifier()
clf.fit(X_train, y_train)
```

Out[22]:
```
LGBMClassifier()
```

In [23]:
```
# predict the results
y_pred=clf.predict(X_test)
```

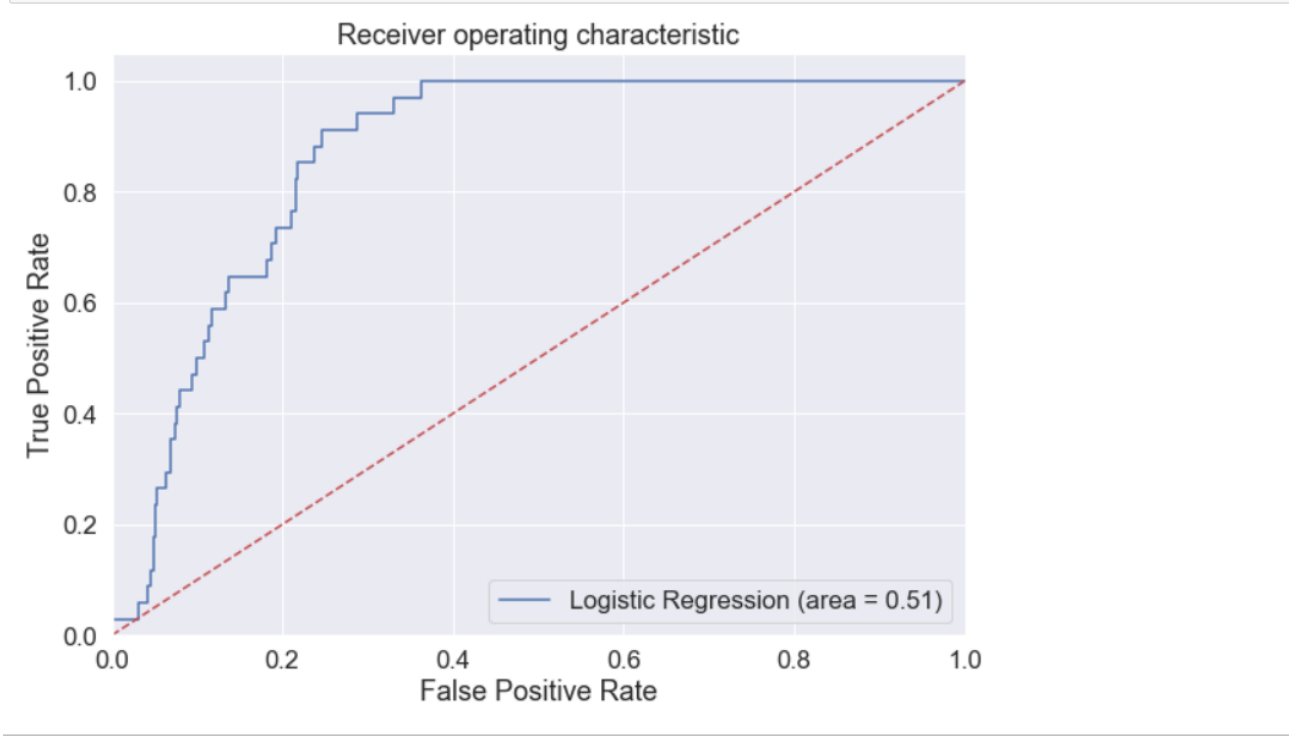
In [24]:
```
# view accuracy
from sklearn.metrics import accuracy_score
accuracy=accuracy_score(y_pred, y_test)
print('LightGBM Model accuracy score: {0:0.4f}'.format(accuracy_score(y_test, y_pred)))
print(classification_report(y_test, clf.predict(X_test)))
```

```
LightGBM Model accuracy score: 0.9420
              precision    recall  f1-score   support

           0       0.97      0.91      0.94       663
           1       0.91      0.97      0.94       630

    accuracy                           0.94      1293
   macro avg       0.94      0.94      0.94      1293
weighted avg       0.94      0.94      0.94      1293
```

AUC ROC curve is drawn as below

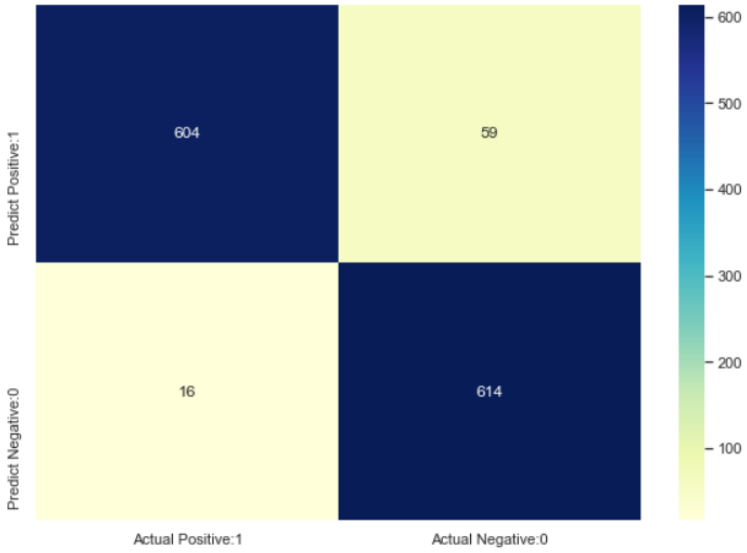


Accuracy is found as 94% with balance prediction of classes. Confusion matrix is evidently proved this result

In [27]:
```
# visualize confusion matrix with seaborn heatmap
from sklearn.metrics import confusion_matrix
cm = confusion_matrix(y_test, y_pred)
cm_matrix = pd.DataFrame(data=cm, columns=['Actual Positive:1', 'Actual Negative:0'],
                                  index=['Predict Positive:1', 'Predict Negative:0'])

sns.heatmap(cm_matrix, annot=True, fmt='d', cmap='YlGnBu')
```

Out[27]: <AxesSubplot:>



Out of **663** Non-DSS observations, model could predict correctly **604** observations - accuracy **91.1%**
Out of **630** DSS observations, model could predict correctly **614** observations - accuracy **97.4%**

Discussion :

As is evident from the above discussion, the model can predict the chance of Dengue Shock Syndrome with platelet and hematocrits counts of the 3rd day of Dengue Fever. The statistics shows Negative Predictive Value to be 91.1% while the Positive Predictive Value is as high as 97.4%. Accurate prediction of DSS early in the disease will aid clinicians in triaging the at-risk patients, continuous monitoring and timely interventions. This can reduce the mortality rate of the disease to a great extent.