Project Proposal

Computer Science and Engineering

COURSE CODE : CSE-437

Team Member	Submitted To
1.Udipta Mondal ID : 221-209-038 **3.Sumaiya Akter Mitu** ID : 221-121-038 **4.Faria Afrin Dola** ID : 221-023-038	**Zakaria Shams Siam** **Lecturer to** **Presidency University**

DATE :06.11..24

Project Proposal : Healthcare Diagnosis and Monitoring

Dengue fever is a disease caused by the dengue virus; it is spread by the Aedes aegypti mosquito which are rampant in tropical islands such as Singapore. Weather also plays an important correlation on the rising number of dengue incidents. This is likely due to Singapore having high average temperatures and precipitation rate across the entire island, stagnant puddles in warm temperatures becomes breeding grounds for these mosquitos.Given a dataset about dengue occurrences, alongside records of Singapore's average temperature and rainfall, for this project I attempt to train a regression model that could predict dengue occurrences in thefuture. This model will include yearly dengue cases and climate factors(mean temperature, relative humidityand rainfall)for the past decade since 2012.The domain-specific area of the dataset would specifically focus on the relationship between dengueoccurrences, temperature as well as rainfall in that year. The regression model could be used to betterunderstand how weather patterns influence the transmission of dengue and to improve dengue forecastingefforts in Singapore. It is important to predict the number of dengue incidents so that the government andthe people will be ready to prevent a dengue outbreak when the number of dengue incidents is predicted to be high.

Objectives of the project:

The objective of this project is to visualize the data provided and utilize machine learning algorithms to train a model that could accurately predict future dengue outbreaks. The impact and contribution of such a model would be significant, as it could help public health officials and policymakers to take proactive measures to prevent the spread of dengue fever before any huge outbreaks. For example, if the algorithm is able to predict that the incidence of dengue fever is likely to increase based on recent rainfall statistics, officials could take steps to reduce mosquito breeding sites, distribute insect repellent, or implement other public health measures to mitigate the risk of dengue transmission. Additionally, if the algorithm predicts that certain areas are at higher risk of experiencing outbreaks due to frequent rainfall and high humidity, resources could be directed to those areas to better address the threat of dengue. Finally, the results of a dengue forecast algorithm may contribute to the overall understanding of the factors that influence the transmission of dengue fever, which could inform the development of more effective prevention and control strategies. By identifying patterns and relationships in the data that are associated with increased risk of dengue transmission, researchers and public health officials can gain valuable insights into the disease and how it spreads, which can inform the development of targeted interventions.

[Problem statement] :

Dengue is a viral infection transmitted to humans through the bite of infected mosquitoes and is found in tropical and sub-tropical climates worldwide, mostly in urban areas. we face challenges in accurately diagnosing diseases such as dengue fever. Our System should help in providing accurate and timely diagnoses by leveraging machine learning algorithms trained on relevant medical data.

Dataset:

The dataset is named 'Weekly Number of Dengue Dataset,' and it was compiled by Singapore's Ministry of Health. The dataset may be obtained at Data.gov.sg, the official government website that makes available to the general public datasets on a range of statistics. There are 5 independent variables that would be collected from the two dataset that span over the term of a decade. These variables consists of 'year', 'Week No.', 'Daily Rainfall', 'Mean temperature' and 'dengue count'. Weekly information of dengue clusters with infection records comes from National Environment Agency of Singapore NEA. The mean temperature and rainfall values are recorded in a separate CSV file presumably acquired by web scrapping past climate trends in Singapore from WeatherSpark.com, a metrology observation website.

Objectives of the project:

The objective of this project is to visualize the data provided and utilize machine learning algorithms to train a model that could accurately predict future dengue outbreaks. The impact and contribution of such a model would be significant, as it could help public health officials and policymakers to take proactive measures to prevent the spread of dengue fever before any huge outbreaks. For example, if the algorithm is able to predict that the incidence of dengue fever is likely to increase based on recent rainfall statistics, officials could take steps to reduce mosquito breeding sites, distribute insect repellent, or implement other public health measures to mitigate the risk of dengue transmission. Additionally, if the algorithm predicts that certain areas are at higher risk of experiencing outbreaks due to frequent rainfall and high humidity, resources could be directed to those areas to better address the threat of dengue.Finally, the results of a dengue forecast algorithm may contribute to the overall understanding of the factors that influence the transmission of dengue fever, which could inform the development of more effective prevention and control strategies. By identifying patterns and relationships in the data that are associated with increased risk of dengue transmission, researchers and public health officials can gain valuable insights into the disease and how it spreads, which can inform the development of targeted interventions.

Target Population :

The target population for this project are the Doctor's ,Nurses, and Medical technicians.This system should help them to accurate and timely diagnoses. with this real-time monitoring system, they can mitigate the spread of diseases like dengue fever. Also this system will help the local people to predict dengue .Sometimes local people are unable to connect the doctor but if they

have the proper test result then our system will help them to predict their problem .At the same time this will help the Medical technicians and the local people .

Proposed Approach : First we will collect the data then we will Clean the collected data.

Programming Languages : Python(Pandas, Matplotlib, Scikit, Numpy).

Implementation :

[Data Pre-processing**]-** The datasets were relatively clean, except for the dengue incidence that were recorded in a seperate dataset whereas climate data are stored in another. Dengue haemorrhagic fever cases should also removed from the dengue dataset, as such cases were far and few and not the main focus of this project. Therefore, the first process below will be used to clean the dengue dataset and convert the database into First Normal Form(1NF).

The table must be single valued and should not consist of missing values:

In [4]:
```
#Check if there is any missing variables
#Or any value that is not positive integer

missing_value = ['N/a','na','nan',np.nan]
df2 = pd.read_csv('Weekly Dengue Cases.csv', na_values = missing_value)
df1 = pd.read_csv('Singapore.csv', na_values = missing_value)
df2.isnull().sum()
df1.isnull().sum()
#Drop any NaN values
df2 = df2.dropna()
df1 = df1.dropna()

display(df2)
```

	year	eweek	type_dengue	number
0	2014	1	Dengue	436.0
1	2014	1	DHF	1.0
2	2014	2	Dengue	479.0
3	2014	2	DHF	0.0
4	2014	3	Dengue	401.0
...	...	...	...	...
523	2018	50	DHF	1.0
524	2018	51	Dengue	127.0
525	2018	51	DHF	1.0
526	2018	52	Dengue	160.0

Remove Dengue haemorrhagic fever(DHF) cases since it is not relevant in this project:

```
# Dropping any DHF

df2 = df2[df2["type_dengue"].str.contains("DHF") == False]
df2 = df2.rename(columns={'number': 'Dengue_number'})

# new CSV file

df2.to_csv('New_Dengue_Cases.csv')

#Dataset
df2
```

Out[5]:

	year	eweek	type_dengue	Dengue_number
0	2014	1	Dengue	436.0
2	2014	2	Dengue	479.0
4	2014	3	Dengue	401.0
6	2014	4	Dengue	336.0
8	2014	5	Dengue	234.0
...	...	...	...	...
518	2018	48	Dengue	109.0
520	2018	49	Dengue	113.0

Remove climate file years to only contain year 2014-2018:

```
#year between (14-18)

d1 = df1.drop(df1[df1['Year'] < 2014].index, inplace = True)


d1 = df1.drop(df1[df1['Year'] > 2018].index, inplace = True)\

# new CSV file
df1.to_csv('New_Singapore.csv')

df1
```

Out[6]:

	Year	Week No.	Daily Rainfall Total (mm)	Mean Temperature (C)
104	2014	1	3.456210	26.592556
105	2014	2	8.061746	26.538159
106	2014	3	0.025827	26.285000
107	2014	4	0.000000	25.810238
108	2014	5	0.000357	26.223190
...	...	...	...	...
360	2018	48	8.713978	27.479643
361	2018	49	8.521088	27.514524
362	2018	50	11.923381	27.171958

Merge the two files to create a new comprehensive dataset:

The datasets were relatively clean, since the missing variables were removed. The table does not contain composite or multi-valued attributes since they are fitted equally in each column,each row of table is unique and does not contain any repeating values. By fufilling these requirements the

dataset is transformed into First Normal Form (

```
Mean Temperature (C)          float64
Dengue_number                 float64
dtype: object
```

	Unnamed: 0	Year	Week No.	Daily Rainfall Total (mm)	Mean Temperature (C)	Dengue_number
0	104	2014	1	3.456210	26.592556	436.0
1	105	2014	2	8.061746	26.538159	479.0
2	106	2014	3	0.025827	26.285000	401.0
3	107	2014	4	0.000000	25.810238	336.0
4	108	2014	5	0.000357	26.223190	234.0
...	...	...	...	...	...	...
256	360	2018	48	8.713978	27.479643	109.0
257	361	2018	49	8.521088	27.514524	113.0
258	362	2018	50	11.923381	27.171958	107.0
259	363	2018	51	4.726233	28.054615	127.0
260	364	2018	52	0.753174	28.279193	160.0

261 rows × 6 columns

```
In [8]: #Measures Mean(average) of per year
```

Statistical Analysis:

This section will be focused on identifying key series of the dataset and provide statistical summary of the data. The central tendency of the data can be used to decribe the data by identifying summary statistics such as mean,median and mode.

#Measures Mean(average) of per year:

```
            'Average Temperature (C)': Temperature_mean,
            'Avergae Dengue Cases': Dengue_mean}
df = pd.DataFrame(data)
display(df)
```

	Year	Average Rainfall (mm)	Average Temperature (C)	Avergae Dengue Cases
0	2014	5.653466	27.857213	345.396226

	Year	Average Rainfall (mm)	Average Temperature (C)	Avergae Dengue Cases
0	2015	4.907074	28.112822	216.961538

	Year	Average Rainfall (mm)	Average Temperature (C)	Avergae Dengue Cases
0	2016	5.960347	28.362796	251.173077

	Year	Average Rainfall (mm)	Average Temperature (C)	Avergae Dengue Cases
0	2017	7.080943	27.85924	52.884615

	Year	Average Rainfall (mm)	Average Temperature (C)	Avergae Dengue Cases
0	2018	6.707895	27.823571	62.634615

```
In [9]: # Calculate the median of the dataset
```

#Measures *Median stat* of per year:

```
    data = {'Year': [i] ,
            'Median Rainfall (mm)': Rainfall_median,
            'Median Temperature (C)': Temperature_median,
             'Median Dengue Cases': Dengue_median}
    df = pd.DataFrame(data)
    display(df)
```

	Year	Median Rainfall (mm)	Median Temperature (C)	Median Dengue Cases
0	2014	5.452107	27.964286	291.0

	Year	Median Rainfall (mm)	Median Temperature (C)	Median Dengue Cases
0	2015	3.587293	28.121978	225.5

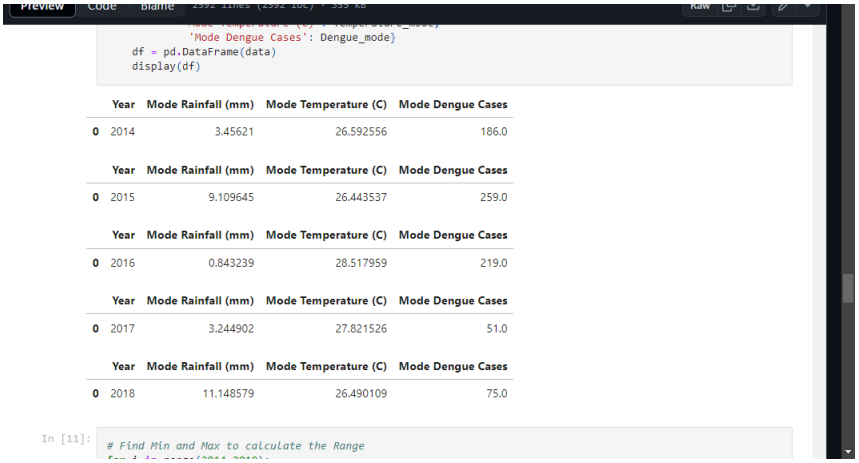
	Year	Median Rainfall (mm)	Median Temperature (C)	Median Dengue Cases
0	2016	5.05426	28.322957	217.0

	Year	Median Rainfall (mm)	Median Temperature (C)	Median Dengue Cases
0	2017	6.313049	27.90179	51.0

	Year	Median Rainfall (mm)	Median Temperature (C)	Median Dengue Cases
0	2018	5.762674	27.832007	56.0

```
In [10]: #Measures Mode stat per year ############################
         ## Mode is the highest occurring value throughout the year
```

#Measures Mode stat per year:



```
                'Mode Dengue Cases': Dengue_mode}
    df = pd.DataFrame(data)
    display(df)
```

	Year	Mode Rainfall (mm)	Mode Temperature (C)	Mode Dengue Cases
0	2014	3.45621	26.592556	186.0

	Year	Mode Rainfall (mm)	Mode Temperature (C)	Mode Dengue Cases
0	2015	9.109645	26.443537	259.0

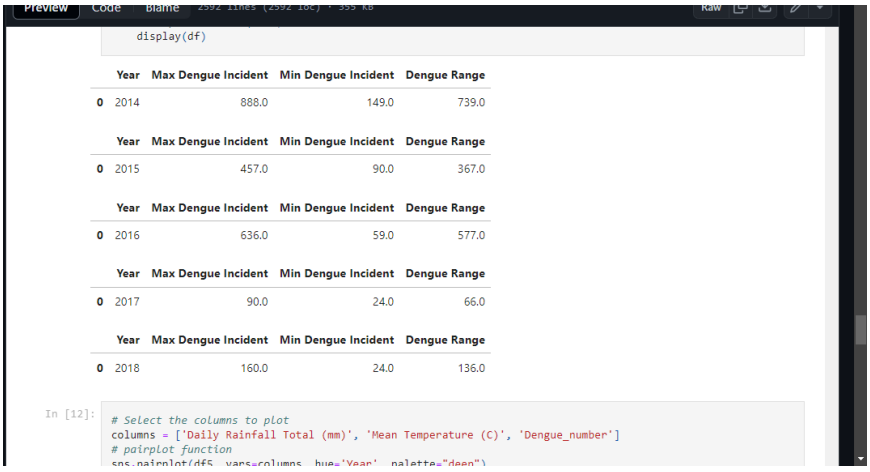
	Year	Mode Rainfall (mm)	Mode Temperature (C)	Mode Dengue Cases
0	2016	0.843239	28.517959	219.0

	Year	Mode Rainfall (mm)	Mode Temperature (C)	Mode Dengue Cases
0	2017	3.244902	27.821526	51.0

	Year	Mode Rainfall (mm)	Mode Temperature (C)	Mode Dengue Cases
0	2018	11.148579	26.490109	75.0

```
In [11]: # Find Min and Max to calculate the Range
```

#Measures *Min and Max*:



```
    display(df)
```

	Year	Max Dengue Incident	Min Dengue Incident	Dengue Range
0	2014	888.0	149.0	739.0

	Year	Max Dengue Incident	Min Dengue Incident	Dengue Range
0	2015	457.0	90.0	367.0

	Year	Max Dengue Incident	Min Dengue Incident	Dengue Range
0	2016	636.0	59.0	577.0

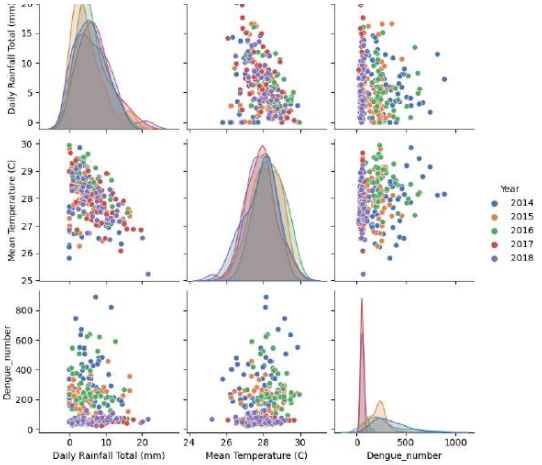
	Year	Max Dengue Incident	Min Dengue Incident	Dengue Range
0	2017	90.0	24.0	66.0

	Year	Max Dengue Incident	Min Dengue Incident	Dengue Range
0	2018	160.0	24.0	136.0

```
In [12]: # Select the columns to plot
         columns = ['Daily Rainfall Total (mm)', 'Mean Temperature (C)', 'Dengue_number']
         # pairplot function
```

Data Visualization:

In this segment, I will be utilizing data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in the data. This is to allow us to understand and explore the variables in great depth before moving on to building a model. Histogram & Scatter Plot Each Year Based on the plot below we can observe the recording of each independent variable druing the years from 2014-2018. Furthermore, the scatter plot also shows us how each variable might have a correlation effect on the other.



Apply Logistic Regression:

```
## Logistic Regression
from sklearn.linear_model import LogisticRegression

logreg_model = LogisticRegression()
logreg_model.fit(X_train, y_train)
y_pred_logreg = logreg_model.predict(X_test)

# Evaluation
print("Accuracy Score : ",accuracy_score(y_test, y_pred_logreg))
print("Precision Score : ",precision_score(y_test, y_pred_logreg,average='macro'))
print("Recall Score : ",recall_score(y_test, y_pred_logreg,average='macro'))
```

```
Accuracy Score :  0.9622641509433962
Precision Score :  0.5
Recall Score :  0.4811320754716981
/opt/conda/lib/python3.10/site-packages/sklearn/metrics/_classification.py:1344: UndefinedMetricWarn
is ill-defined and being set to 0.0 in labels with no true samples. Use `zero_division` parameter to
is behavior.
  _warn_prf(average, modifier, msg_start, len(result))
```

```
## Decision Tree
```

Logistic Regression: A supervised learning algorithm used for binary classification tasks. It predicts the probability that an instance belongs to a particular class.

Accuracy: The proportion of correctly predicted instances.
Precision: The proportion of true positive predictions among all positive predictions.
Recall: The proportion of true positive predictions among all actual positives.

Apply Decision Tree:

```
is behavior.
  _warn_prf(average, modifier, msg_start, len(result))
```

15]:

```
## Decision Tree
from sklearn.tree import DecisionTreeClassifier

tree_model = DecisionTreeClassifier()
tree_model.fit(X_train, y_train)
y_pred_tree = tree_model.predict(X_test)

# Evaluation
print("Accuracy Score : ",accuracy_score(y_test, y_pred_tree))
print("Precision Score : ",precision_score(y_test, y_pred_tree,average='macro'))
print("Recall Score : ",recall_score(y_test, y_pred_tree,average='macro'))
```

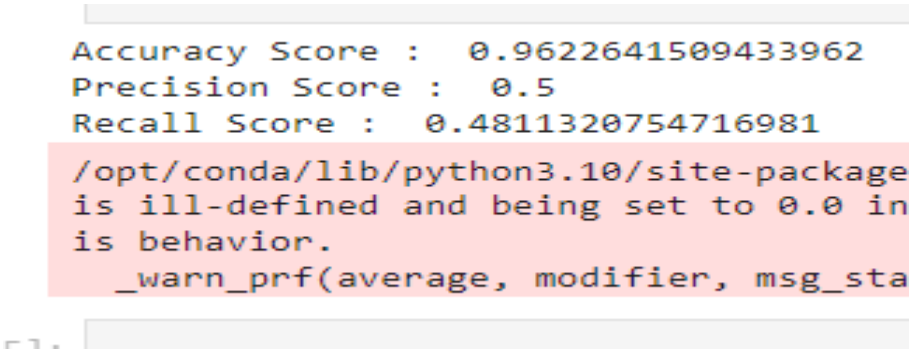
```
Accuracy Score :  0.6792452830188679
Precision Score :  0.5
Recall Score :  0.33962264150943394
/opt/conda/lib/python3.10/site-packages/sklearn/metrics/_classification.py:1344: UndefinedMetricWarning: Recall
is ill-defined and being set to 0.0 in labels with no true samples. Use `zero_division` parameter to control th
is behavior.
  _warn_prf(average, modifier, msg_start, len(result))
```

Decision Tree Classifier: A supervised learning algorithm used for both classification and regression tasks. It splits the data into branches based on feature values to make predictions. This approach ensures that the model is evaluated comprehensively, considering different aspects of its performance. Decision trees are intuitive and easy to interpret, making them a popular choice for many classification tasks. However, they can be prone to overfitting, especially if not properly tuned

Conclusion:

There are 2 different models developed to predict the dengue occurrences and each with varying degrees of accuracy Logistic Regression, Decision Tree. we can see In terms of Logistic Regression the

Accuracy Score is : 0.9622641509433962



Which is almost 1 . Logistic Regression good model for predicting the data.

Final Output : The application of the machine learning models for prediction of dengue outbreak can provide vital information to healthcare authorities so that they can better prepare for dengue fever outbreaks.

Motivation : The motivation behind this project is to improve public health and this system will control the spread of the disease .Nowadays our country people are not so much aware about Dengue and this is why so many people are affected by Dengue. Our System should help in providing accurate and timely diagnoses and this will help the people to be aware and this can mitigate the spread of dengue fever. There is no specific treatment for dengue and this is the main problem about this diseases. Most of the time we see lot of bad news about Dengue in Newspaper and Television and this is why we are decided to make a system on this problem .

Literature Review :

1. Publication :PLOS(https://doi.org/10.1371/journal.pntd.0005973)

The authors found that the support vector regression model performed the best. This model was able to forecast the dengue outbreaks in other provinces. Across five provinces confirmed SVR's superiority in tracking dengue.

2. Publication: Scientific Reports (https://www.nature.com/articles/s41598-020-79193-2#Abs1)

the study suggests that machine learning, particularly SVM's, can be a valuable tool for predicting dengue outbreaks, especially when combined with data balancing techniques. The findings highlight the potential for using these models to inform public health interventions and preventative actions.

3. Publication: SpringerLink (https://link.springer.com/article/10.1186/s12916-018-1108-5)

Overall, the paper underscores the importance of adopting a holistic approach to dengue prediction. The paper discusses the application of ML in dengue prediction, acknowledging its potential to revolutionize public health and provide new insights into infectious diseases