PROJECT REPORT

ON

FALSE DATA INJECTION ATTACKS

Submitted in partial fulfillment of the requirements

Submitted By

S. NIDHI SHREE   245320748043

Under the guidance

Of

Dr. Rajasekaran

Ph.D., PMP

Department of CSE(AIML)

NEIL GOGTE INSTITUTE OF TECHNOLOGY
A Unit of Keshav Memorial Technical Education (KMTES)
Approved by AICTE, New Delhi & Affiliated to Osmania University, Hyderabad

PROJECT SCHOOL CERTIFICATE

Title: False Data Injection
Mentor: Dr. Rajasekaran
Session Duration :12Weeks [13thOct2022 - 12th Jan2023]

Class: CSM A
Student Name: S. NIDHI SHREE
Roll No: 245320748043

Signature of Faculty

Signature of student

PROJECT OBJECTIVE

Predictive maintenance techniques are designed to help determine the condition of in-service equipment to estimate when maintenance should be performed. This approach promises cost savings over routine or time-based preventive maintenance because tasks are performed only when warranted. Thus, it is regarded as condition-based maintenance carried out as suggested by estimations of the degradation state of an item.

This project deals with the problem of false data injection. In this application, I am currently using “PDM Dataset”, Using this Dataset I am trying to predict the life cycle of ratio using the PdM method (predictive maintenance). The main objective comes into place when an external unauthorized user makes changes in the dataset without the knowledge of the person handling the data, the predictions also get affected.

The project can be divided into the following parts: -

- Prediction of the life cycle
- Making a Graph of truly identified life cycles using CMAPPS Dataset.
- Injecting False Data into CMAPPS Dataset.
- Making another graph using the same method but using the false CMAPPS dataset.
- Making a website, displaying the graphs of different algorithms.

The above parts may further be classified into different parts.

TECHNOLOGY STACK

The technologies used in this application are:

- Python 3.10.0
- MATLAB
- Django

The modules used in this application are:

- TensorFlow
- Sci-kit-learn
- matplotlib
- pandas
- NumPy
- Keras
- Seaborn

The algorithms used in this application are:

- kNN (k-Nearest Neighbors)
- Logistic Regression
- Random Forests
- Autoencoders

DEVELOPED SOLUTION

PdM DATASET:

The dataset consists of multiple multivariate time series. Each data set is further divided into training and test subsets. Each time series is from a different engine i.e., the data can be considered to be from a fleet of engines of the same type. Each engine starts with different degrees of initial wear and manufacturing variation which is unknown to the user. This wear and variation are considered normal, i.e., it is not considered a fault condition. There are three operational settings that have a substantial effect on engine performance. These settings are also included in the data. The data is contaminated with sensor noise.

The engine is operating normally at the start of each time series and develops a fault at some point during the series. In the training set, the fault grows in magnitude until system failure. In the test set, the time series ends some time prior to system failure. The objective of the competition is to predict the number of remaining operational cycles before failure in the test set, i.e., the number of operational cycles after the last cycle that the engine will continue to operate. Also provided is a vector of true Remaining Useful Life (RUL) values for the test data.

The data are provided as a zip-compressed text file with 26 columns of numbers, separated by spaces. Each row is a snapshot of data taken during a single operational cycle, each column is a different variable. The columns correspond to:

1) unit number
2) time, in cycles
3) operational setting 1
4) operational setting 2
5) operational setting 3
6) sensor measurement 1
7) sensor measurement 2

…

26) sensor measurement 26

Using these sensors, we are going to predict its lifetime.

PREDICTIVE MAINTENANCE:

Predictive maintenance software uses data science and predictive analytics to estimate when a piece of equipment might fail so that corrective maintenance can be scheduled before the point of failure. The goal is to schedule maintenance at the most convenient and most cost-efficient moment, allowing equipment’s lifespan to be optimized to its fullest, but before the equipment has been compromised.

This predictive maintenance can be analyzed using the previous data of the equipment. CMAPPS dataset is a perfect example to make predictive maintenance of the cycle of the plane, using all the sensors.

MAKING GRAPHS OF PdM USING PDM DATASET:

I have used "Machine Learning" algorithms to predict the cycle's condition. To achieve this there are different steps:

- **DATA PROCESSING:**

  This is one of the most important steps in this application. Using data pre-processing techniques or worse state.

  My 1st step was to remove all the unrequired features which would confuse the ml algorithm. "NumPy", and "Pandas", are widely used in this process of pre-processing the data.

  My next step was to create EOL Labels (End of life cycle Labels) for making this I used a sample formula, The formula for EOL is: "id – cycle – 1".

  Using these EOL values the next step was to create LR (Life Ratio), the formula I have used to make these Life Ratio Labels is " current cycle/ current EOL of the cycle ".

  These EOL labels are thereafter used to label the cycle, the following labels are:

  1. If LR<=0.6: Good
  2. If LR lies in the range of 0.6 and 0.8: Moderate
  3. If LR>0.8: Warning

  These labels are predicted using machine learning models.

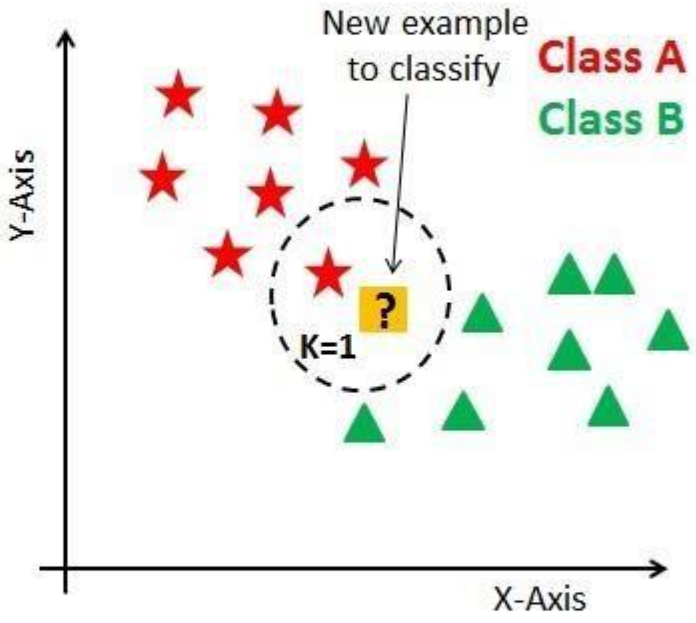
- **PREDICTING LABELS**

  For this step we will be using different models:

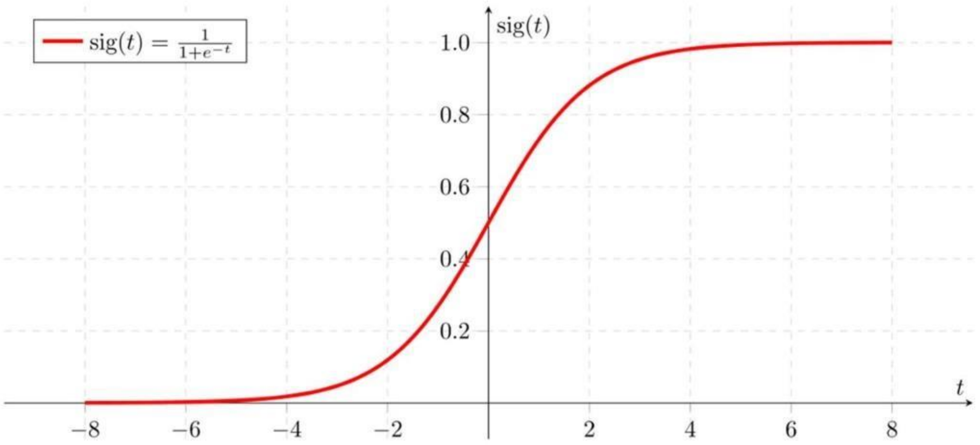
  1. **KNN (k- NEAREST NEIGHBORS):**

     In KNN, K is the number of nearest neighbors. The number of neighbors is the core deciding factor. K is generally an odd number if the number of classes 2. When K=1, then the algorithm is known as the nearest neighbor algorithm. This is the simplest case. Suppose P1 is the point, for which the label needs to predict.

First, you find the one closest point to P1 and then the label of the nearest point assigned to P1. Therefore, in the below example the “?” will be a star.

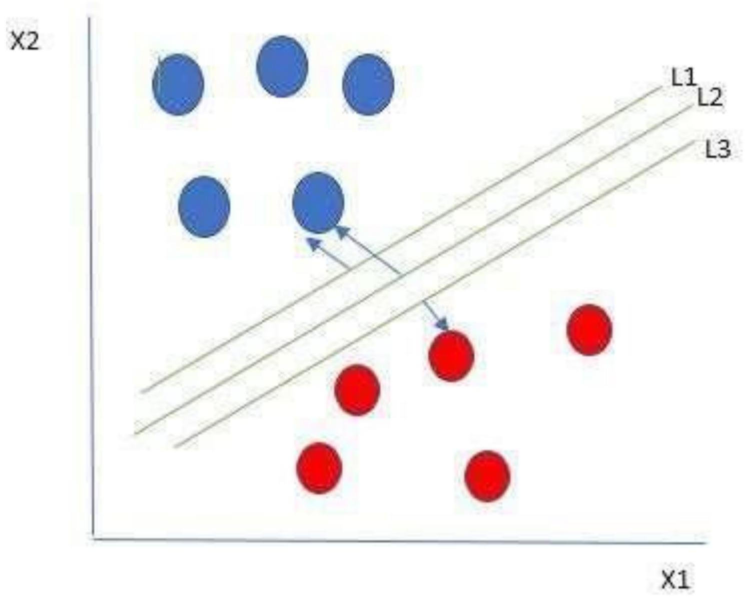


2. LOGISTIC REGRESSION:



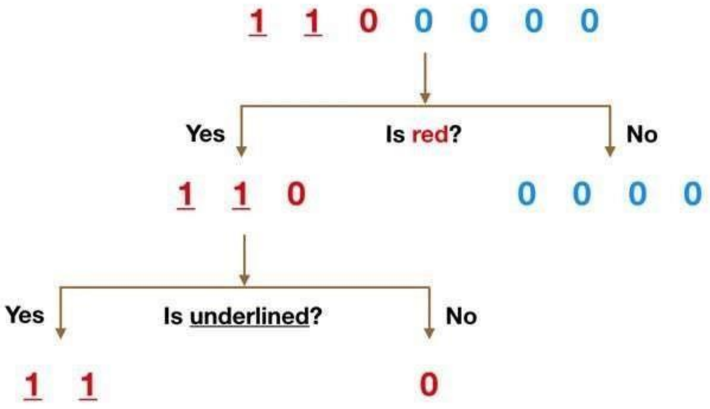
If we ignore the fact that logistic regression has regression in its name, logistic regression is a classification technique that is one of the most widely used methods for classification. This takes the Regression model to classification by using the sigmoid function. This sigmoid function is used to make the predicted value as a number, but this number lies between 0-1, this process makes the model classify theta as one part of the model has values between 0-0.5, and another as it lies between 0.5-1.

3. SVM (SUPPORT VECTOR MACHINE):

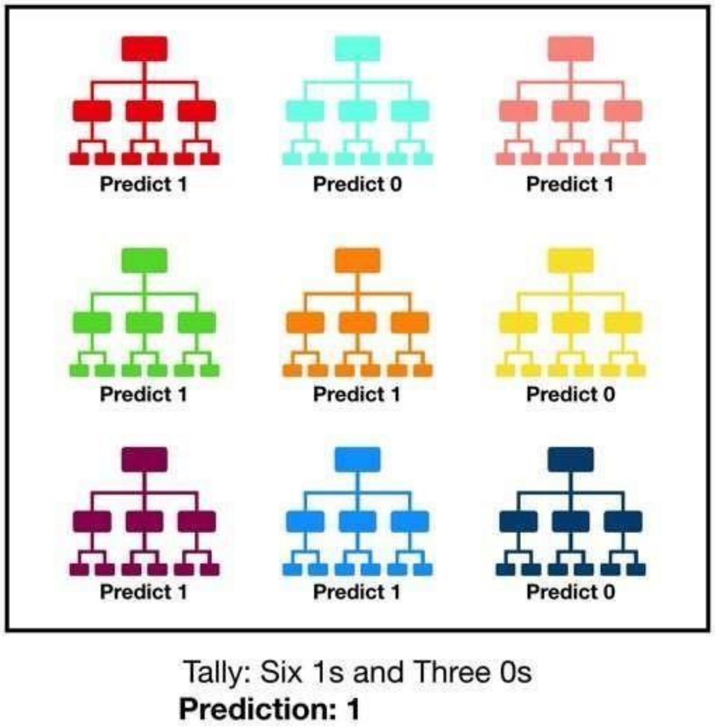


The SVM kernel is a function that takes low-dimensional input space and transforms it into higher-dimensional space, i.e., it converts a non-separable problem to a separable problem. It is mostly useful in non-linear separation problems. Simply put the kernel, does some extremely complex data transformations and finds out the process to separate the data based on the labels or outputs defined.

4. RANDOM FOREST:



Random forest, as name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction (see figure below).



5. AUTO ENCODERS:

Logistic Regression is a Supervised Machine Learning Algorithm that is used for classification problems.

In the predictive analysis of the CMAPSS dataset, we implement this classification technique by using the aid of Life Ratio (LR) which is a ratio between the End of Life and the Cycle.

We classify the data provided using this LR into three different conditions

- Good
- Moderate
- Warning

Then after labeling the data, we use Logistic Regression to classify the data and train the model according to it.

Then we plot the actual data and the predicted data

The plot with minimum deflections and maximum accuracy indicates the true data

The plot with maximum deflections indicates False data.

• MAKING FALSE DATA:

I have used MATLAB for making false data. False data can be of different types, which may be randomized insertion, deletion of the column, insertion of the different column, continuous insertion of values, or manipulation of the values. I have focused to discuss the randomized manipulation of the data.

The 1st step is to make a random variable every time in a column, the technique is to create a random variable of 1-5 so that if the value is 1 or 2 we will manipulate the data, in other cases, we will be not disturbing the data.

In this way, I have got the data that is randomly manipulated, as so that the data is 40 percent false.

• GRAPHS:

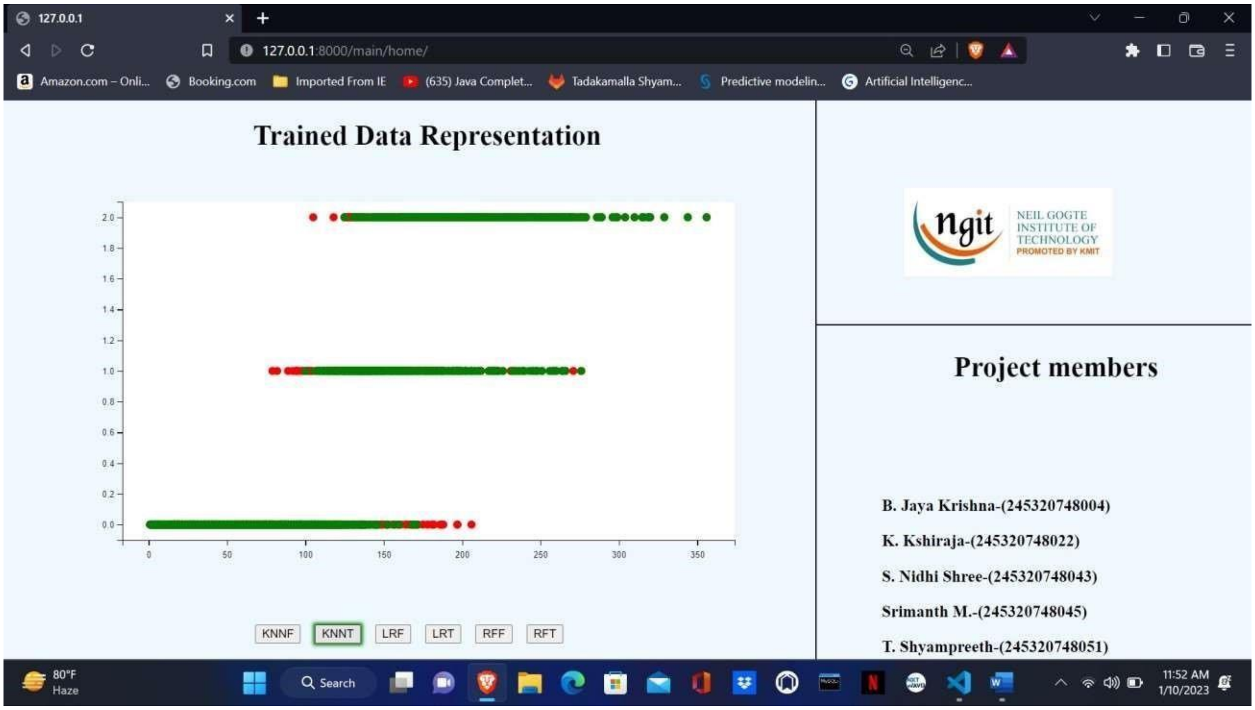
For making the visual representation of how the changes will be in the model, using the false data and the true data, I have used matplotlib a library in python that is used to plot different graphs for the model with false data and without false data.

I am plotting the data in such a way that X-Axis has the cycle number and Y-Axis has the label, we will be plotting the same graph using different colors,

1) Red color dots represent the ACTUAL LABELS.
2) Green color dots represent the PREDICTED LABELS.

The following are screenshots of the project where we can visualize the actual and predicted values.

SCREENSHOTS



Trained Data Representation

Project members

B. Jaya Krishna-(245320748004)

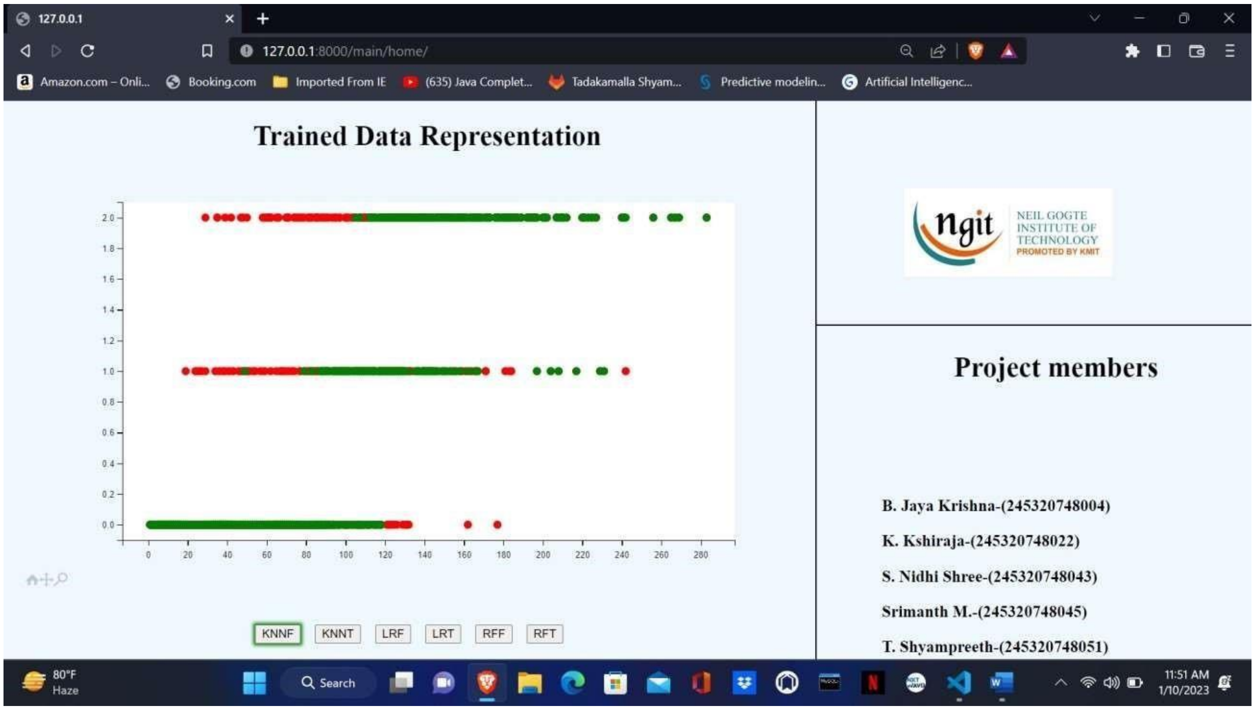
K. Kshiraja-(245320748022)

S. Nidhi Shree-(245320748043)

Srimanth M.-(245320748045)

T. Shyampreeth-(245320748051)

Figure 1: kNN without False Data Injection



Trained Data Representation

Project members

B. Jaya Krishna-(245320748004)

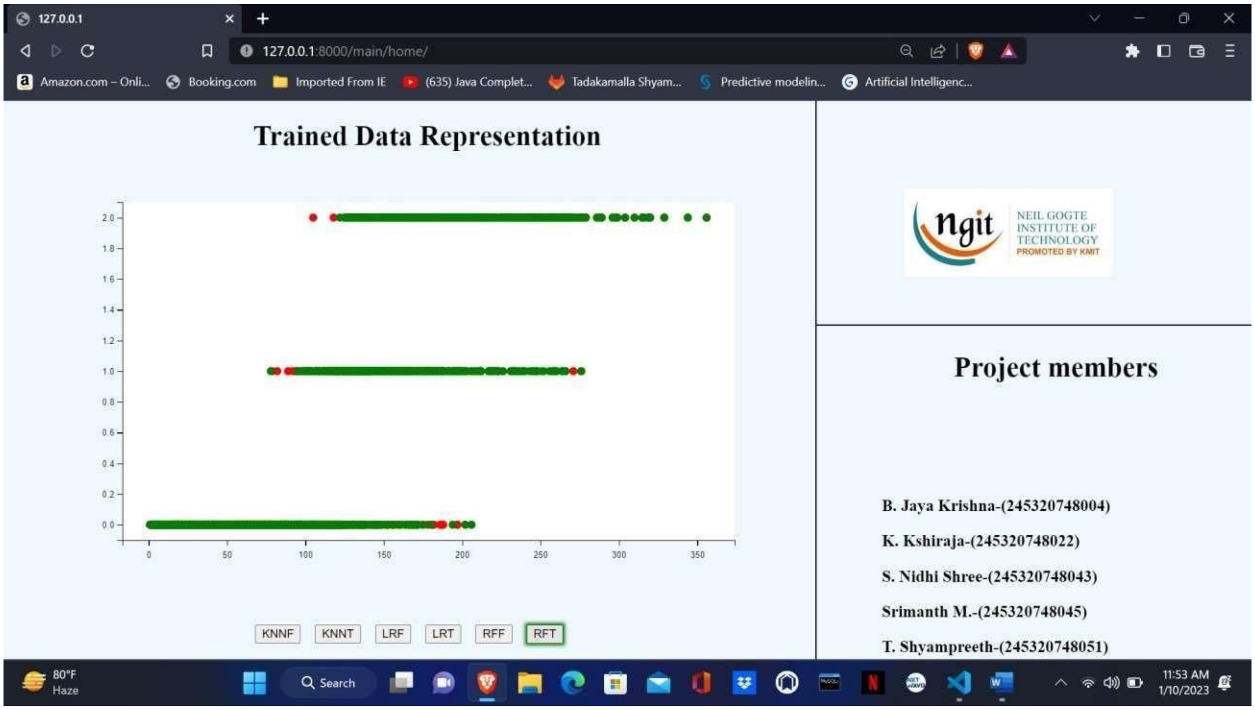
K. Kshiraja-(245320748022)

S. Nidhi Shree-(245320748043)

Srimanth M.-(245320748045)

T. Shyampreeth-(245320748051)

Figure 2: kNN after injection of False Data



Trained Data Representation

Project members

B. Jaya Krishna-(245320748004)

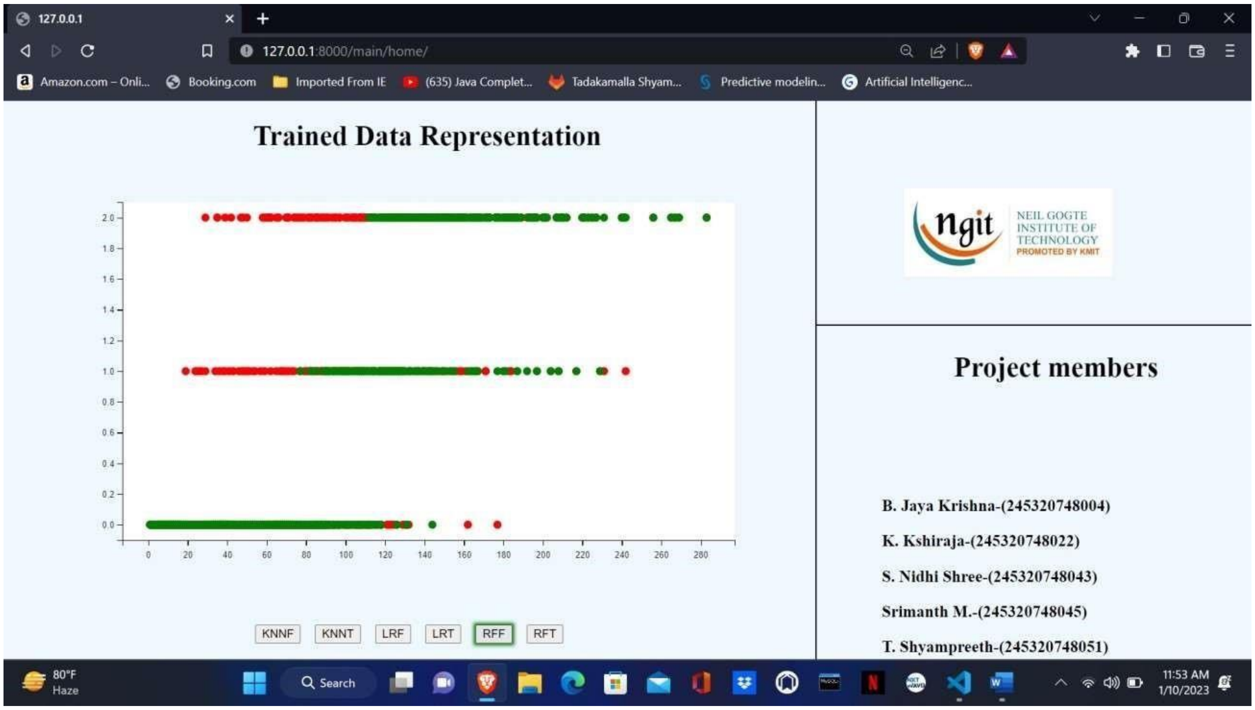
K. Kshiraja-(245320748022)

S. Nidhi Shree-(245320748043)

Srimanth M.-(245320748045)

T. Shyampreeth-(245320748051)

Figure 3: Random Forest without False Data Injection



Trained Data Representation

Project members

B. Jaya Krishna-(245320748004)

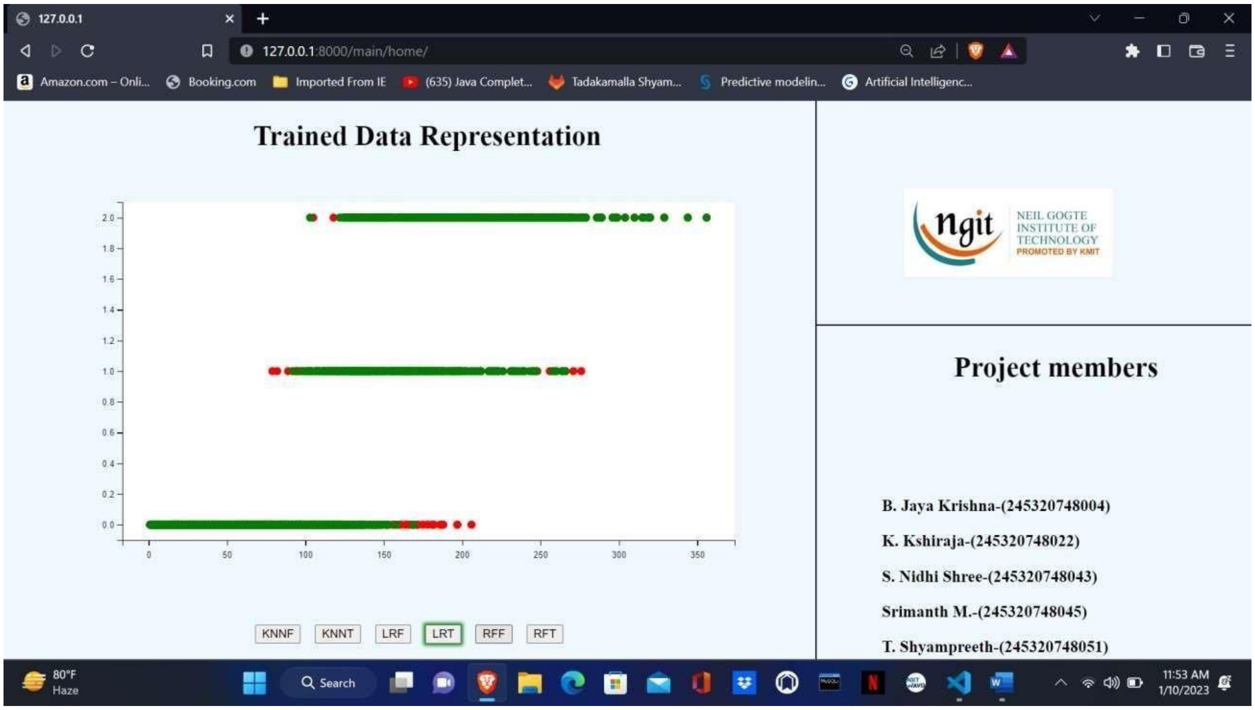
K. Kshiraja-(245320748022)

S. Nidhi Shree-(245320748043)

Srimanth M.-(245320748045)

T. Shyampreeth-(245320748051)

Figure 4: Random Forest after injection of False Data



Trained Data Representation

Project members

B. Jaya Krishna-(245320748004)

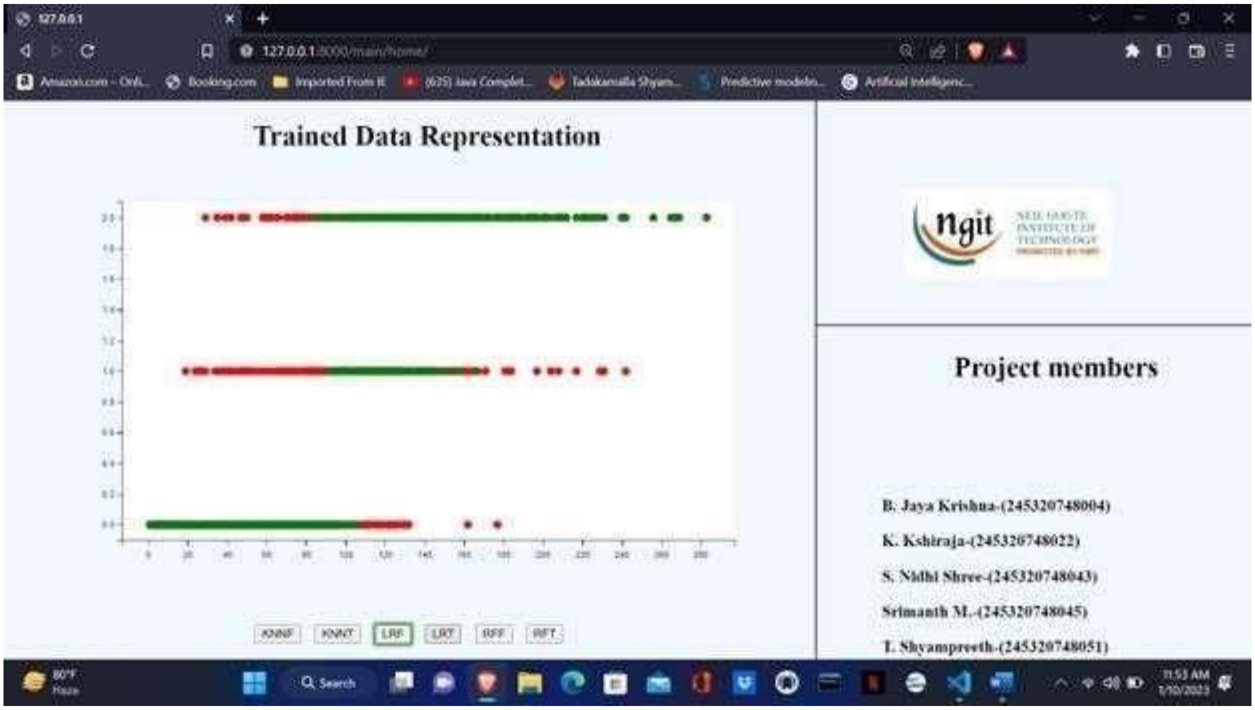
K. Kshiraja-(245320748022)

S. Nidhi Shree-(245320748043)

Srimanth M.-(245320748045)

T. Shyampreeth-(245320748051)

Figure 5: Logistic Regression without injection of False Data



Trained Data Representation

Project members

B. Jaya Krishna-(245320748004)

K. Kshiraja-(245320748022)

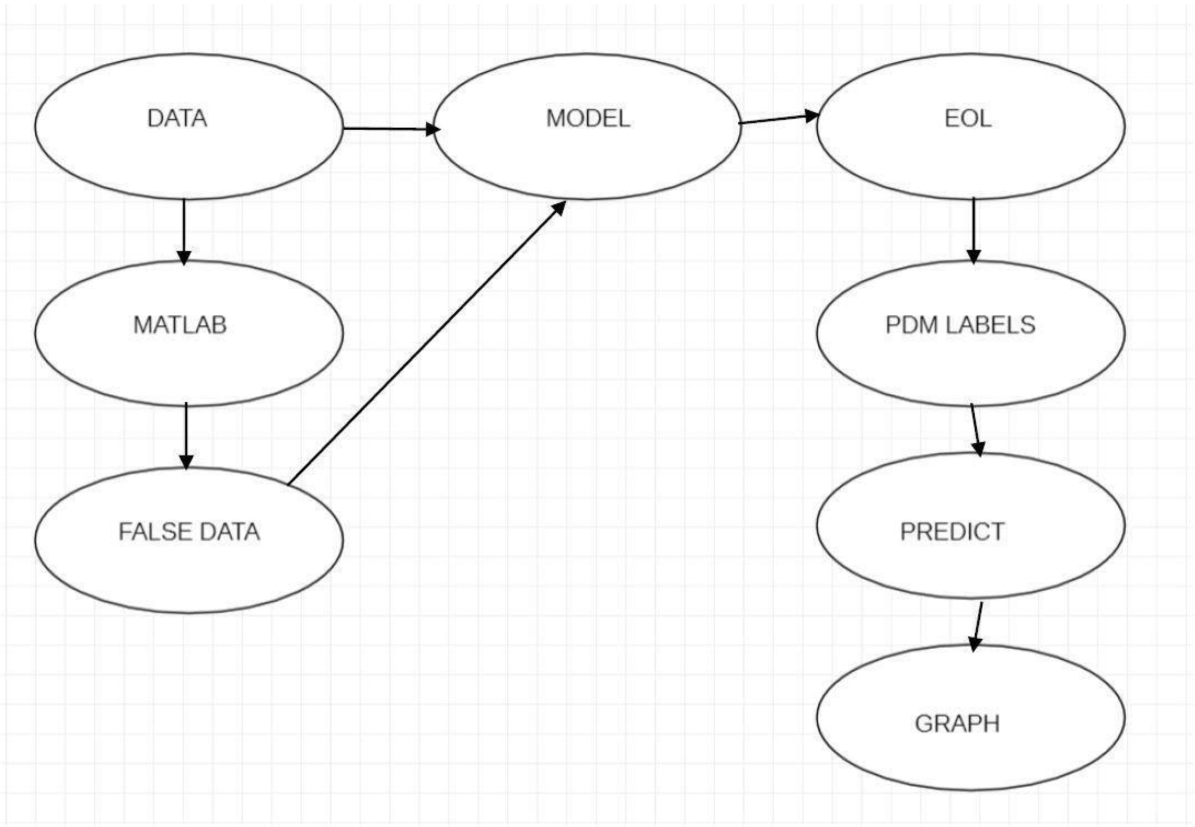
S. Nidhi Shree-(245320748043)

Srimanth M.-(245320748045)

T. Shyampreeth-(245320748051)

Figure 6: Logistic Regression after insertion of False Data

SYSTEM WORKFLOW



CONCLUSION AND FUTURE SCOPE

From the above graphs, we can see understand that the red dots which are not covered by the green dots are the falsely predicted labels. The Above graphs have a common pattern. In which the true Dataset-trained models are having less number of red dots displaying and the false dataset-trained models have more red dots. This can be understood as the false dataset trained models making a greater number of false predictions which intern make more red dots appear as the green dots (predicted dots) would not be covering them.

From this application, we have understood that the purity of the data is the most important thing to predict the label. This application is not yet completed, there is still room for improvement. For example, we can increase the numerical algorithms. Also, this application is not real-time so we can modify this application to intake data stream directly from the sensors themselves. Also, the number of features used in the current version is less so we can increase the number of features.

GITHUB LINK:

https://github.com/nidhishree29/FALSE-DATA-INJECTION-PdM.git