PROJECT REPORT

USER AUTHENTICATION FROM MOUSE LOG DATA USING SVM CLASSIFIER

GROUP 1:

MEMBERS:RAVI JATINBHAI KAKAIYA(21EE64R06)

SOURABH BHATTACHARYA (21EE64R18)

RISHABH TRIPATHI (21EE64R11)

SHUBHANKAR SARKAR (21EE64R01)

ANKIT KUMAR (20MI91R03)

SANTANU KUNDU (18EC10052)

PROBLEM NUMBER: 06

Acknowledgment

We would like to express our deepest appreciation to all those who provided us the possibility to complete this report. We cordially acknowledge the contributions of our project guide **Bijaylaxmi Das Ma'am** under **Prof. Sudipta Mukhopadhyay**whose suggestions were very insightful regarding the Successful completion of MIES end semester project.

Q) PROBLEM AND STATEMENT:

Authenticate the user by training and testing your assigned classifier by the data acquired for continuous authentication data. Use five-fold validation to report results.

DETAILS SPECIFIC TO OUR TEAM:

We were assigned the group 1 hence according the question distribution protocol we were allotted the dataset of the next group i.e., the group 2. Regarding the question, We were supposed to use Support Vector Machine (SVM) Classifier for user identification using 'MouseLog' data.

FORMAL DEFINITION OF ALL THE MODULES REFERRED TO:

1) Math module:

A part of the standard python library, the Math module assists for vivid mathematical operations as well as some very important values in the literature as well.

2) String module:

The String module plays a very crucial role when trying to implement the array of bytes representing the unicode characters.

3) OS module:

The OS module comes under the Python's standard utility modules which provides a convenient way of using the functions of the operating system. Hence this module provided an interface for dealing with the operarting systems. To interact with the file systems functions such as *os* and *os.path* comes in handy.

4) Datetime module:

A date in Python is not a data type of its own, but we can import a module named date time to work with dates as date objects.

5) Numpy module:

NumPy is a module for Python. The name is an acronym for "Numeric Python" . It is an extension module for Python, mostly written in C. This makes sure that the precompiled mathematical and numerical functions and functionalities of NumPy guarantee great execution speed.

6)Itertools module:

This module implements a number of iterator building blocks inspired by constructs from APL, Haskell, and SML. Each has been recast in a form suitable for Python.The module standardizes a core set of fast, memory efficient tools that are useful by themselves or in combination. Together, they form an "iterator algebra" making it possible to construct specialized tools succinctly and efficiently in pure Python.

7) Matplotlib module:

Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits.

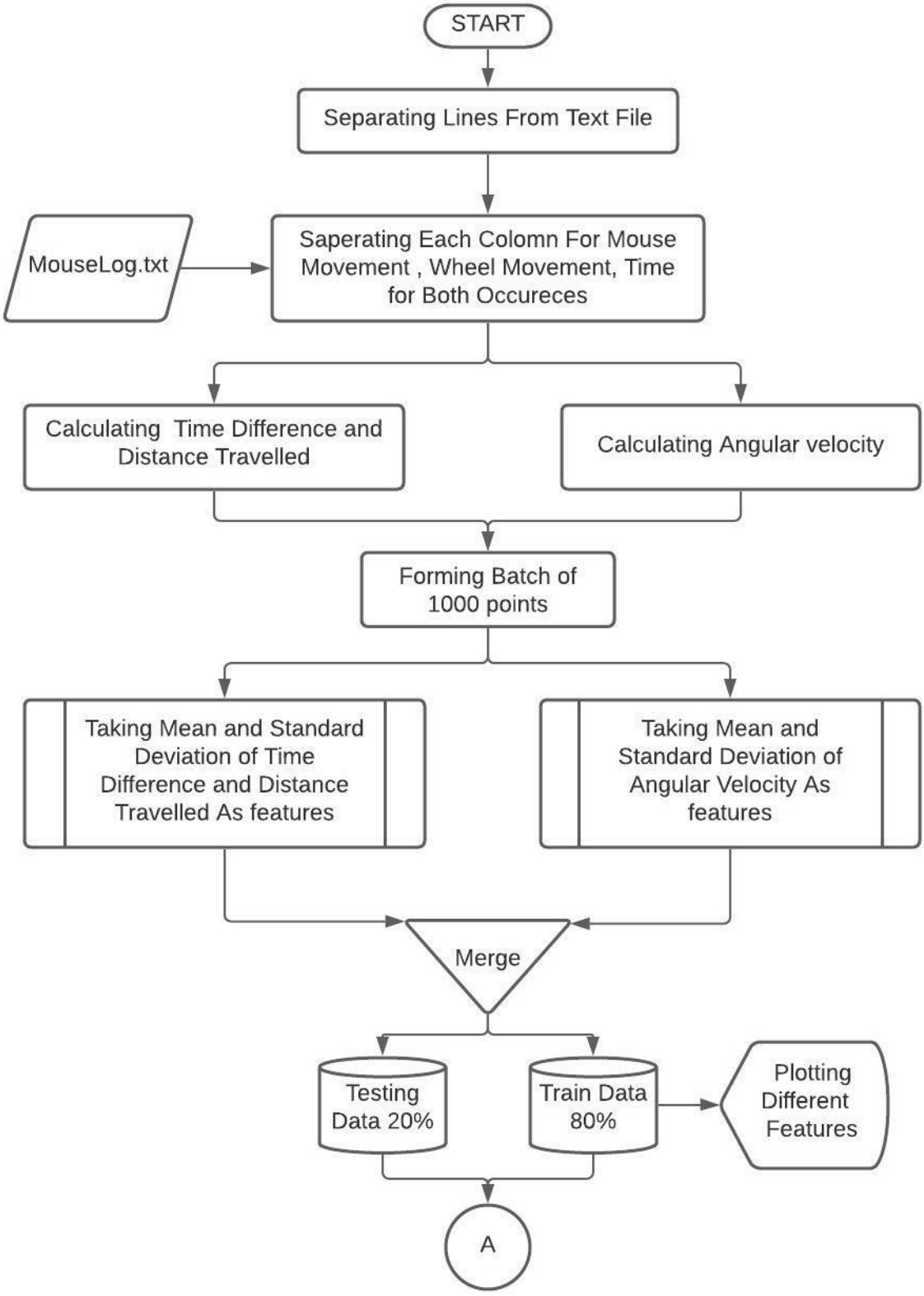
8) Sklearn module:

a)SVC: From Sklearn we are using the SVM module which houses the linear support vector classifier. To put in short, a formal definition can go like this support vectors are data points that are closer to the hyper plane and influence the position and orientation of the hyper plane. Using these support vectors, we maximize the margin of the classifier. Deleting the support vectors will change the position of the hyper plane. These are the points that help us build our SVM.

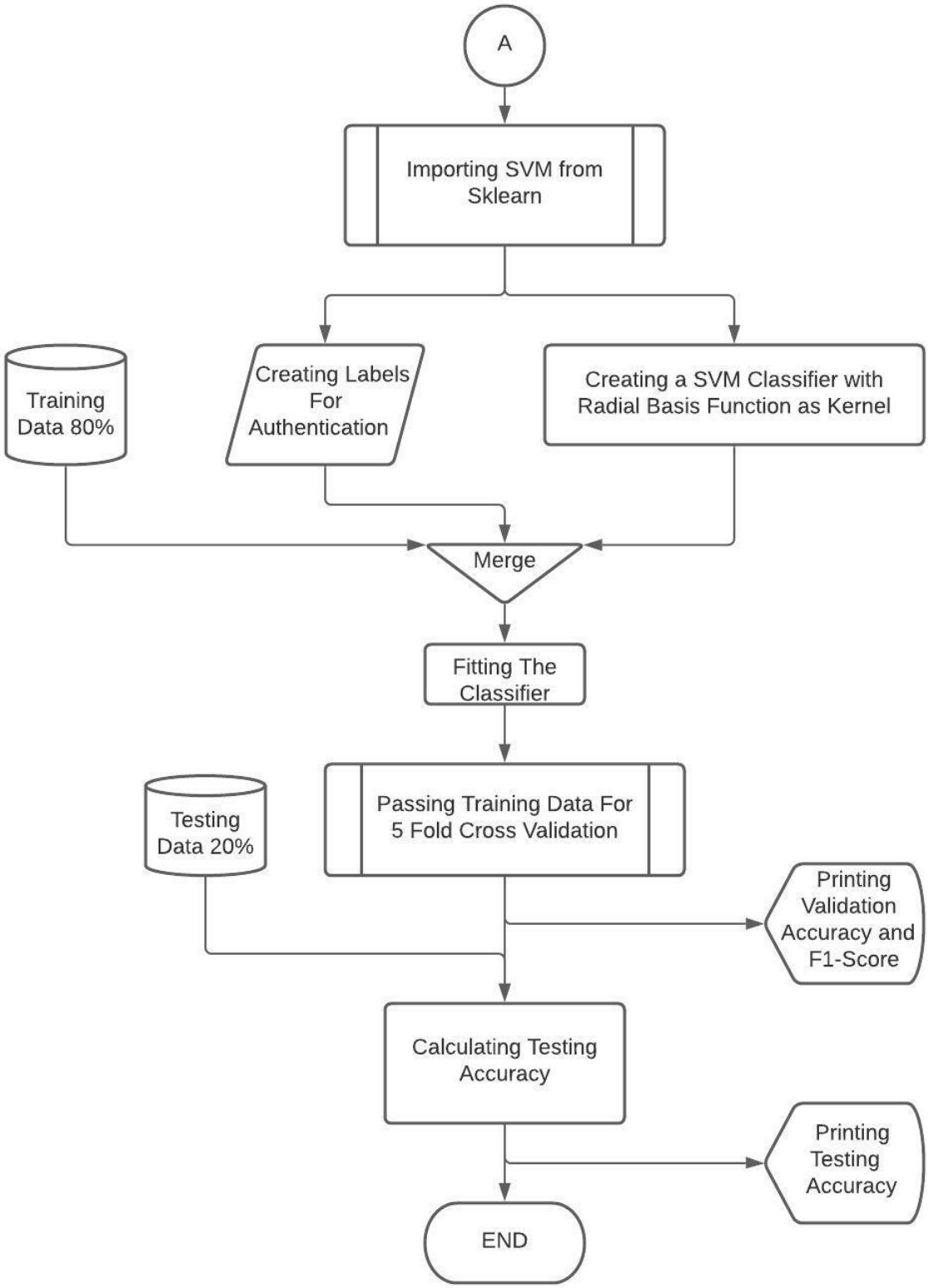
b) CROSS_VAL_SCORE: From SKLEARN library, MODEL_SELECTION module which houses the metric to quantify the cross validation operation. The cross_val_score() function will be used to perform the evaluation, taking the dataset and cross-validation configuration and returning a list of scores calculated for each fold.

c) METRICS: The sklearn.metrics module implements functions assessing prediction error for specific purposes. These metrics are termed distinctively as classification metrics, regression metrics and clustering metrics.

Flow-chart 1:


START
Separating Lines From Text File
MouseLog.txt
Saperating Each Colomn For Mouse Movement , Wheel Movement, Time for Both Occureces
Calculating Time Difference and Distance Travelled
Calculating Angular velocity
Forming Batch of 1000 points
Taking Mean and Standard Deviation of Time Difference and Distance Travelled As features
Taking Mean and Standard Deviation of Angular Velocity As features
Merge
Testing Data 20%
Train Data 80%
Plotting Different Features
A

Flow-Chart 2:


A
Importing SVM from Sklearn
Training Data 80%
Creating Labels For Authentication
Creating a SVM Classifier with Radial Basis Function as Kernel
Merge
Fitting The Classifier
Testing Data 20%
Passing Training Data For 5 Fold Cross Validation
Printing Validation Accuracy and F1-Score
Calculating Testing Accuracy
Printing Testing Accuracy
END

Support Vector Machine:

Support vector machine comes fromfamily of classifier. There can be many solutions for finding a hyper plane given by the Equation ($A^TX + B = 0$), but the confidence of separations/margin will vary in all the cases. Hence we have to look for an optimal solution in order to find such a separation regarding hyper planes having both high confidence level and high margin. This is where SVM comes in handy.

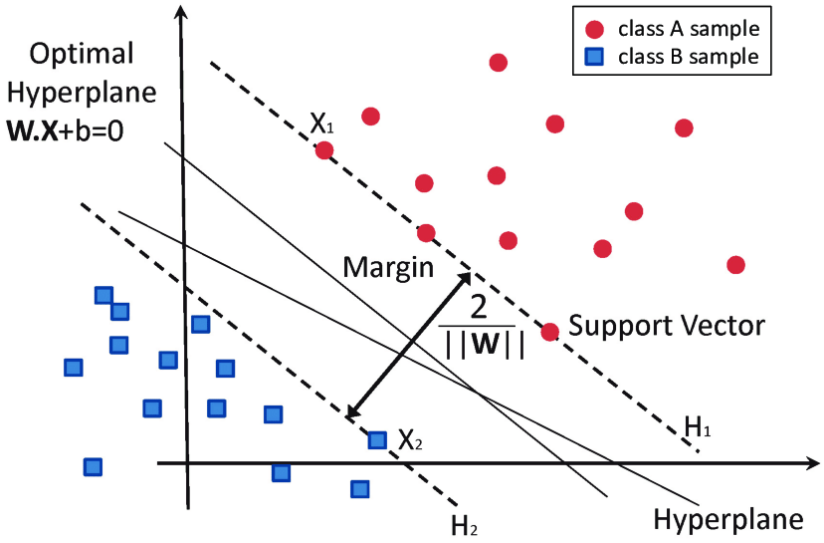


Fig 1

Kernels:

The datasets that we deal pretty often might not always be linear. Hence countering them we will be requiring nonlinear decision boundaries. To make our problem simpler the feature vectors are mapped to higher dimensional feature spaces. This mapping operations are carried out by what are called ‘Kernels’.

Radial Basis Function:

The radial basis function denoted by:

$$Y(x) = \sum_{i=1}^{N} Wi * \emptyset(||x - xi||)$$

It is a generalization of polynomial kernel function which maps a finite dimension input feature space into a infinite dimension feature space with assists in classification. The weights W_i are calculated by linear least square method since that is how they are related to the function. The X_i s are 'N' number of center points based on which given function is to be transformed.

Data Handling:

From the MouseLog.txt file we intend to extract the lines according to our requirement. Using OS library for python we are opening the file and separating out the each lines. And trying to access the each data values for every kind of mouse operations including (Mouse movement, Mouse Wheel movement, Mouse Pressed, Mouse released). After getting the values for the X and Y pixel coordinates and respective time values we intend to calculate those kind of features which separate out the Authenticate User from others.

Feature Extraction:

So now for that we can think of some features which performs well. Now for that first we are supposed to find the features like Distance Travelled, Time Difference and Angular Velocity. Over here the angular velocity represents the Proportion of Mouse scroll in particular time.

Train Split and Batch Formation:

Then for reducing the data complexity we formed the batches of 1000 data point. After forming batches and finding the Mean and Standard deviation for the batch wise data points. After finding this data values we put them in training set and done 80-20 Split.

Fitting The Classifier:

Then Creating the SVM Classifier using Sklearn library. Then created the label according to the authenticate user and unauthenticated user. And Fitting the classifier to the training set.

Five Fold Validation:

After that Fivefold validation is to be carried out. And gained results for all five folds and averaged out them to get final result. We used two method for result which are '**Accuracy**' and '**F1-Score**'. Doing that we shifted to Test our classifier. We than took 20% dataset and Tested the classifier using Accuracy as parameter.

Feature Selection:

During the selection of the feature we tried two plot most feature in pair and observed the patterns between them. Then tried to train classifier using **Two, Four, and Six** features and combination of them at a time. These six feature included the **Mean and Standard deviation of time difference between two consecutive actions, Distance travelled by mouse between two time stamp, Angular velocity proportion.** And opted for the two features giving the best result. These features comes out to be '**Mean of Distance travelled of a batch**' and '**Mean of angular velocity of a batch**'.

Additional features Tested:

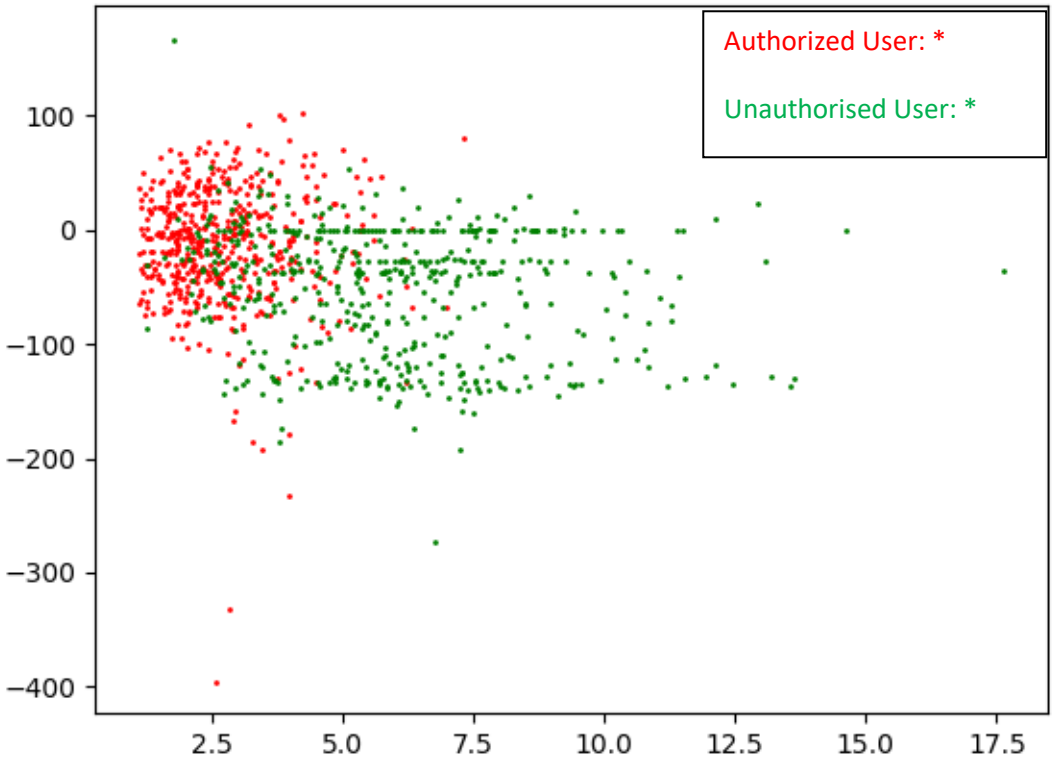
In order to get good results we opted for some additional feature which are acceleration and velocity of Mouse movement. But these features where unable to give a better separation between authorized and unauthorized user. So being unable to help for purpose we omitted those features. We tried for all those six features which were giving a better result for our classification task.

Intermediate Results:

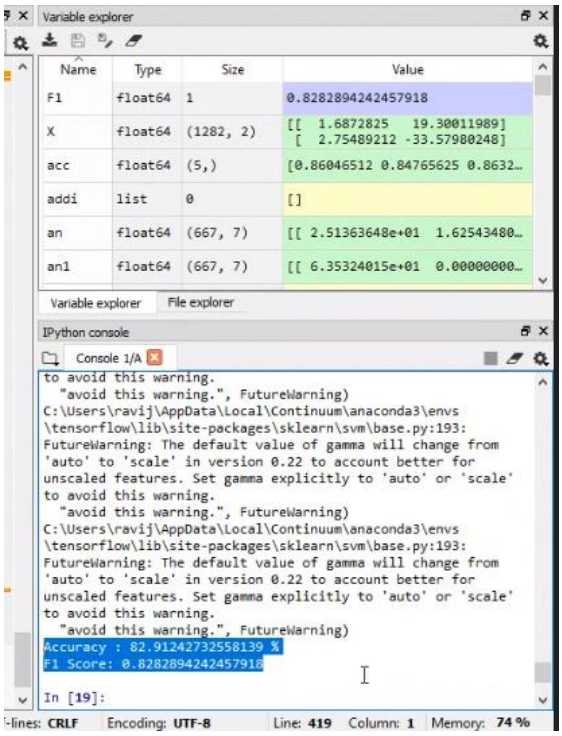
In order to test our model's performance we started with one week's data and observed the distribution of our data. Taken one of the 3 Users as authenticate user and created labels according to that.

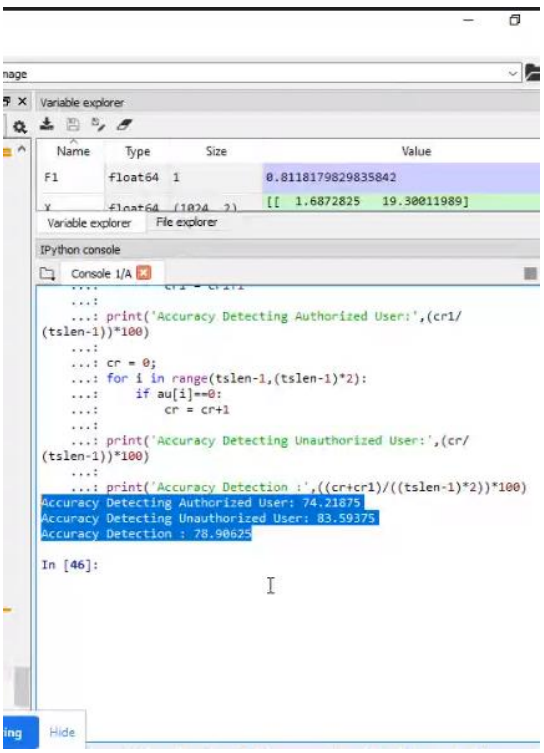
Intermediate Result after Data of Week 1+2:

Mean of Distance travelled vs. Mean of angular velocity



Validation and Testing accuracy



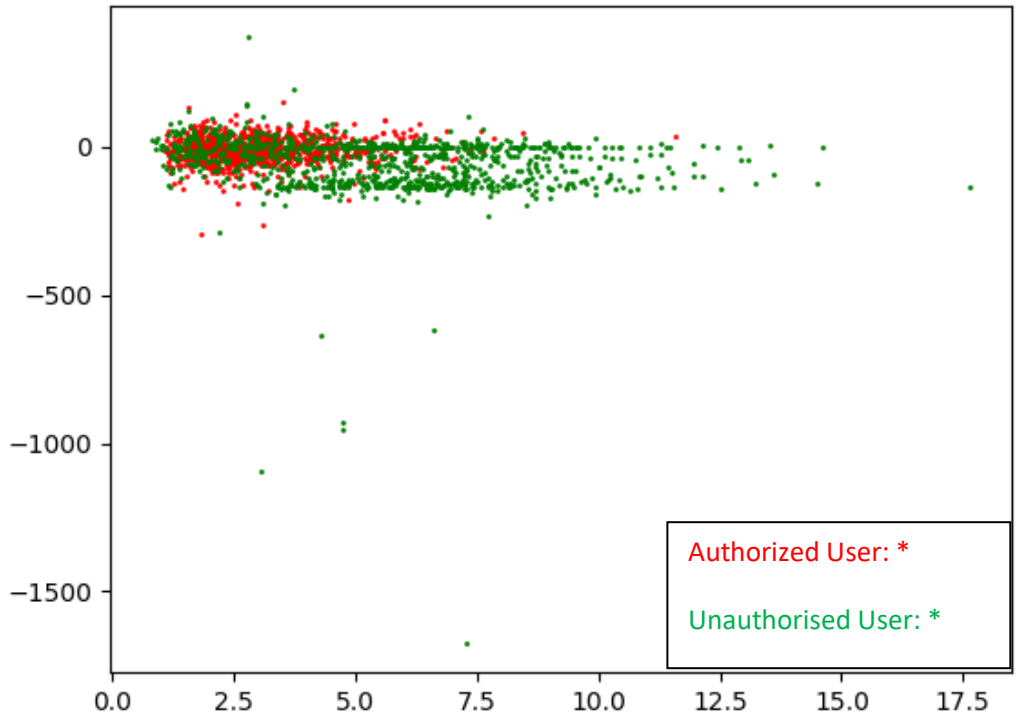


Accuracy:82.91% and F1 Score :0.828

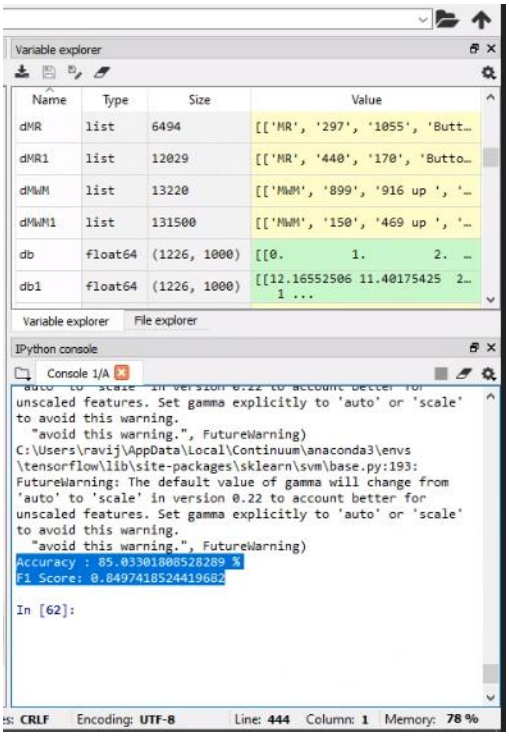
TestingAccuracy :78.90%

All week's Results:

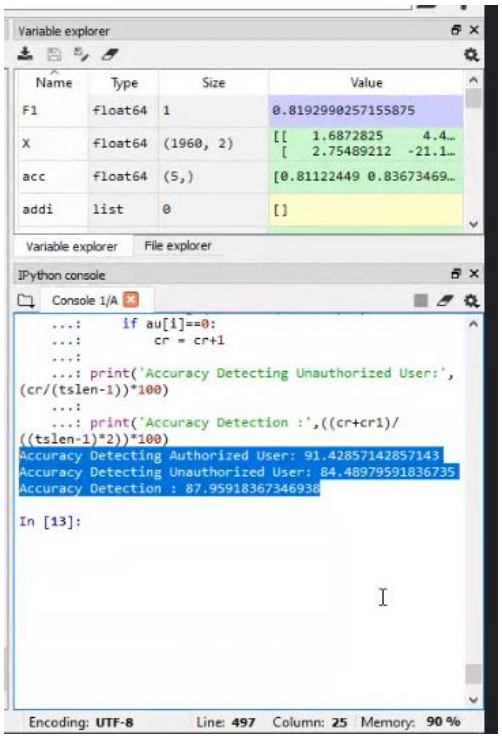
Mean of Distance travelled vs. Mean of angular velocity



Validation and Testing accuracy

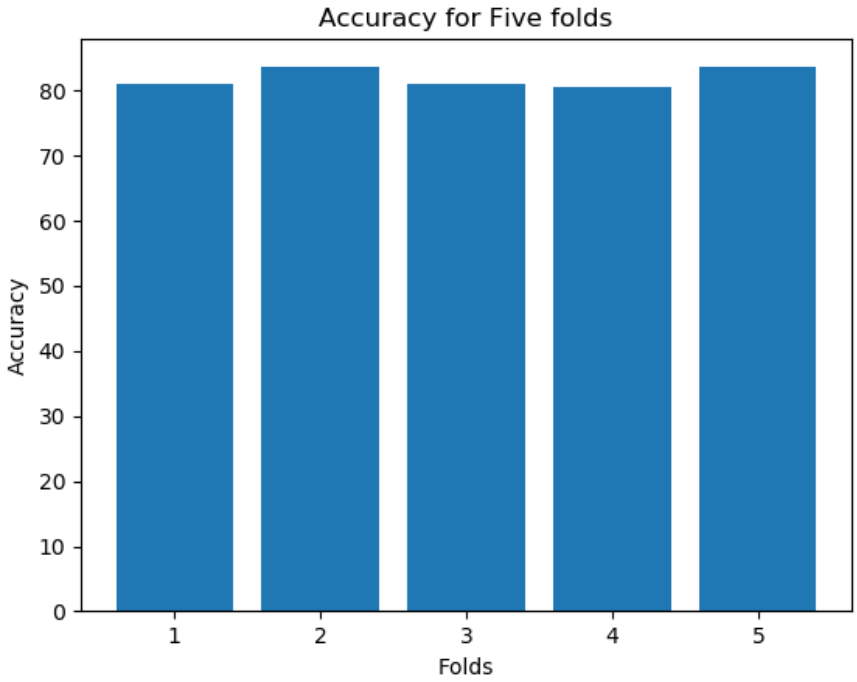


Accuracy:85% and F1-score :0.849

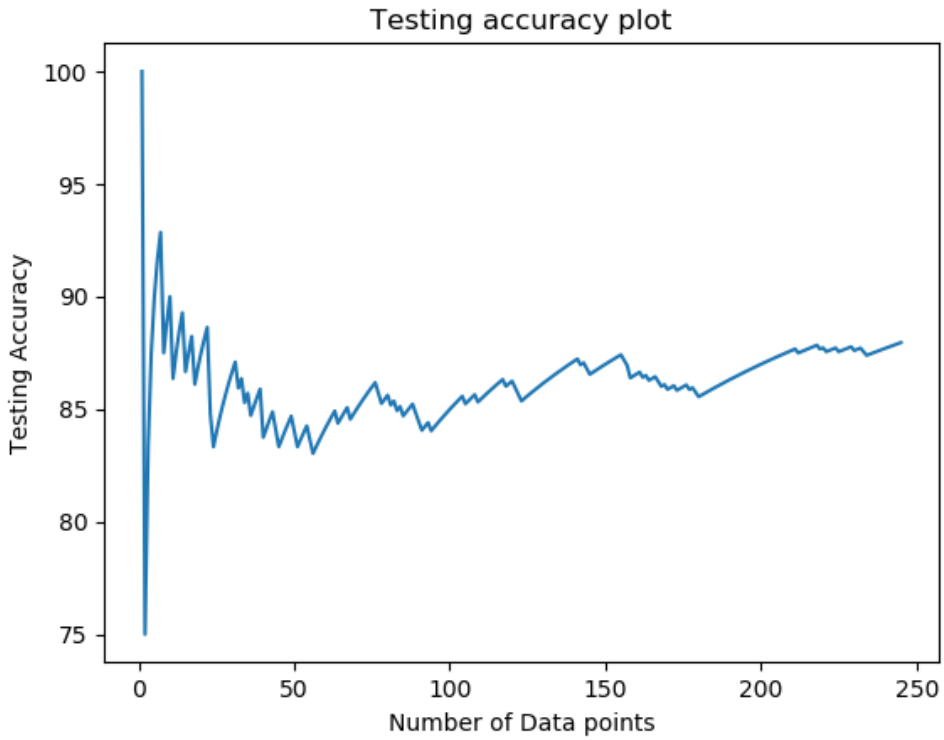


TestingAccuracy : 87.96%

Validation accuracy for 5 folds:

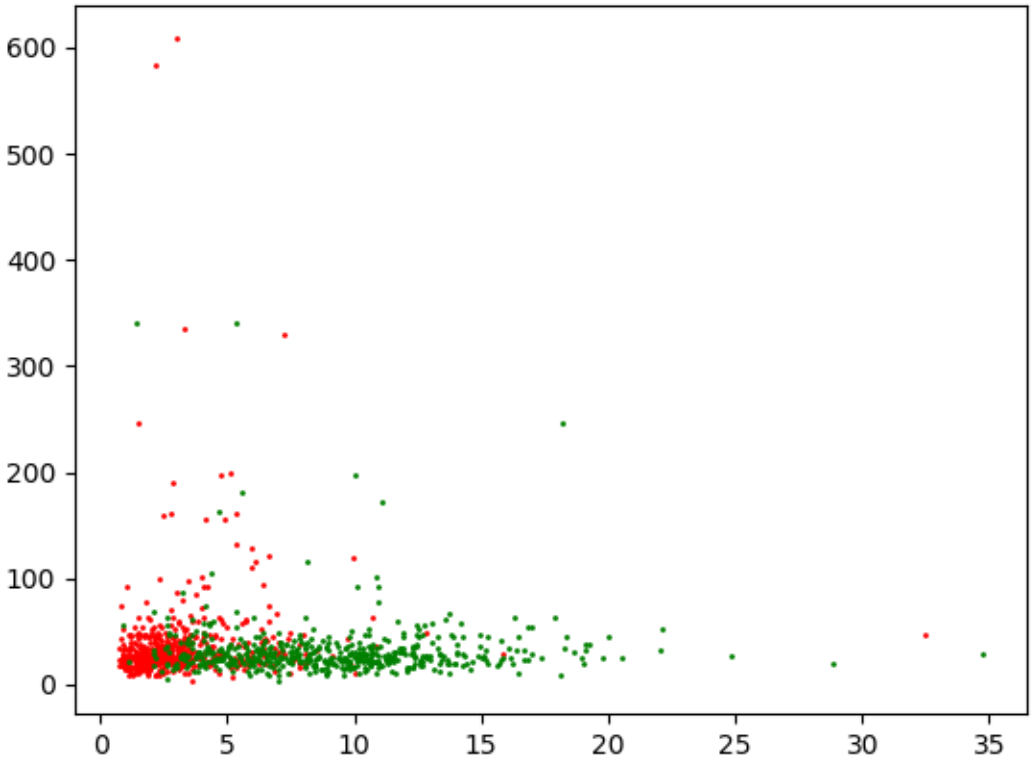


Testing Accuracy Plot:

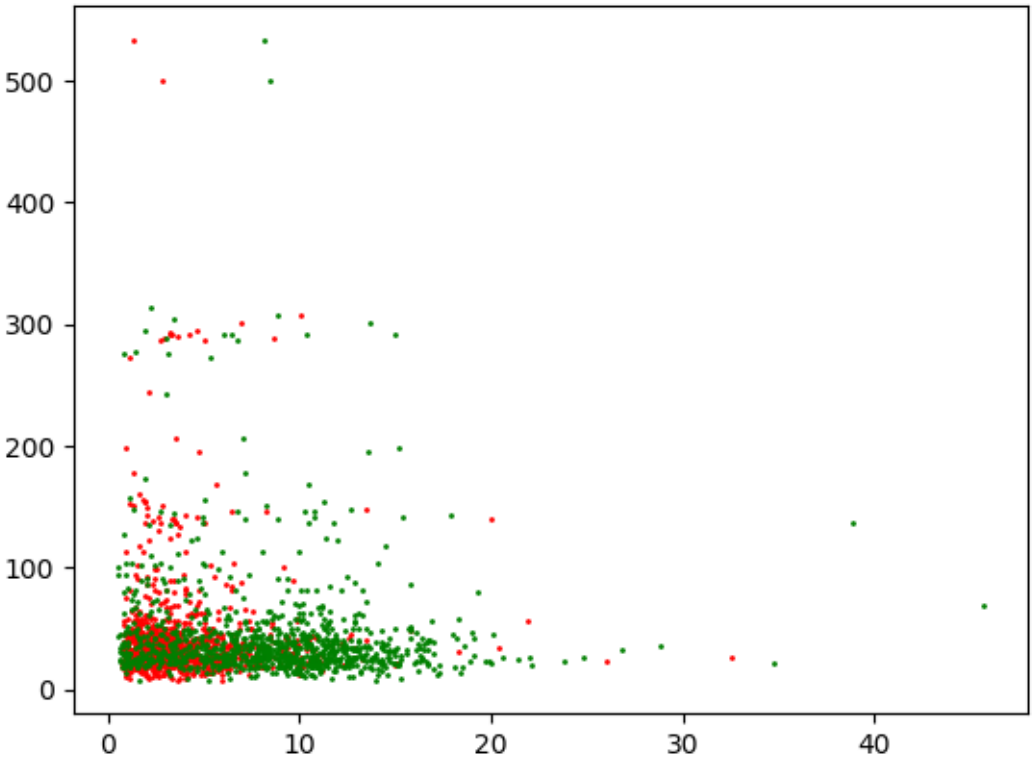


Distribution of data for none used features:

Standard Deviation of (Distance vs. Angular Velocity) 2 weeks



Standard Deviation of (Distance vs. Angular Velocity) 4 weeks



Conclusion:

Given project enriches us with the knowledge of using one class support vector machine to authenticate the user based on his mouse log data. With progress of the project we learnt to analyze and choose some useful feature out of bunch of features available to increase the accuracy of our classifier.

Final Results:

Validation accuracy : 85%

Testing Accuracy : 87%

References:

- https://www.researchgate.net/figure/Classification-of-data-by-support-vector-machine-SVM_fig8_304611323 - fig 1
- https://ieeexplore.ieee.org/document/7860231
- Python Library Documentation
- https://www.youtube.com/watch?v=kcQ0RzLfB_0&list=PLbRMhDVUMngc7NM-gDwcBzIYZNFSK2N1a&index=13
- https://en.wikipedia.org/wiki/Radial_basis_function
- https://en.wikipedia.org/wiki/Support-vector_machine