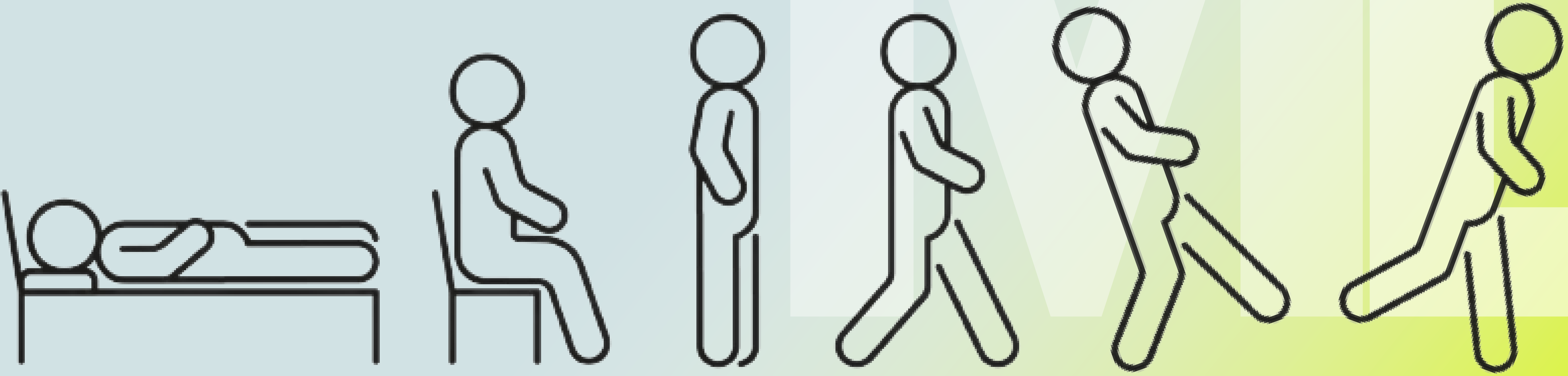
IML PROJECT

SMARTPHONE-BASED RECOGNITION OF HUMAN ACTIVITIES



PROJECT MENTOR

ATHARVA PANDEY

TEAM MEMBERS

- RUTAM RAJHANSA
- HARSH PRASAD
- RIPUNJ GUPTA
- ADITYA GIRI
- MITALI AGRAWAL

PROBLEM STATEMENT

SMARTPHONE-BASED RECOGNITION OF HUMAN ACTIVITIES

-BUILD ML MODEL TO PERFORM TASK OF HUMAN ACTION RECOGNITION USING THE INTERTIAL MEASUREMENTS FROM SMARTPHONE AS INPUTS.

UNDERSTANDING FROM PROBLEM STATEMENT

We are basically given a dataset with training data and test data. We need to use the training data to train our model and then use the test data to predict what kind of activity the user is doing.

DIFFERENT ACTIVITIES

1 WALKING

2 WALKING_UPSTAIRS

3 WALKING_DOWNSTAIRS

4 SITTING

5 STANDING

6 LAYING

7 STAND_TO_SIT

8 SIT_TO_STAND

9 SIT_TO_LIE

10 LIE_TO_SIT

11 STAND_TO_LIE

12 LIE_TO_STAND

UNDERSTANDING OUR DATASET

A GROUP OF 30 PEOPLE PERFORMED A PROTOCOL OF ACTIVITIES COMPOSED OF SIX BASIC ACTIVITIES: THREE STATIC POSTURES (STANDING, SITTING, LYING) AND THREE DYNAMIC ACTIVITIES (WALKING, WALKING DOWNSTAIRS AND WALKING UPSTAIRS).

THE EXPERIMENT ALSO INCLUDED POSTURAL TRANSITIONS THAT OCCURRED BETWEEN THE STATIC POSTURES. THESE ARE: STAND-TO-SIT, SIT-TO-STAND, SIT-TO-LIE, LIE-TO-SIT, STAND-TO-LIE, AND LIE-TO-STAND.

FOR VIEWING OUR DATA DISTRIBUTION WE HAVE USED VARIOUS PLOTS AND DIMENSIONALITY REDUCTION METHODS SUCH AS:

1)PCA
2)LDA
3)T-SNE
4)T-SNE ON LDA
5)PCA ON LDA

WE HAVE USED ALL EARLIER MODEL BUT WE CAN SEE THAT THE DATASET WE HAVE IS TIME DEPENDENT DATASET THUS WE HAVE USED THE CONCEPTS AND MODELS OF TIME SERIES ANALYSIS

WE HAVE USED SEQUENCE FEATURE SELECTION TO SELECT THE MOST ACCURATE FEATURE FOR PREDICTING RESULT

WHAT IS SEQUENCE FEATURE SELECTION?

SEQUENCE FEATURE SELECTION IS A PROCESS OF IDENTIFYING AND SELECTING RELEVANT FEATURES FROM A SEQUENCE OR TIME-SERIES DATA. IN THE CONTEXT OF MACHINE LEARNING, SEQUENCES OFTEN REFER TO DATA POINTS THAT ARE ORDERED, SUCH AS TIME-STAMPED OBSERVATIONS OR EVENTS. THE GOAL OF SEQUENCE FEATURE SELECTION IS TO CHOOSE A SUBSET OF FEATURES THAT CONTRIBUTES MOST EFFECTIVELY TO THE PREDICTIVE PERFORMANCE OF A MODEL WHILE DISREGARDING IRRELEVANT OR REDUNDANT INFORMATION. THIS PROCESS IS CRUCIAL FOR IMPROVING MODEL EFFICIENCY, REDUCING OVERFITTING, AND ENHANCING INTERPRETABILITY.
METHODS FOR SEQUENCE FEATURE SELECTION CAN INCLUDE:

1. FILTER METHODS: THESE METHODS ASSESS THE IMPORTANCE OF EACH FEATURE BASED ON STATISTICAL MEASURES OR OTHER CRITERIA. FEATURES ARE RANKED, AND A SUBSET IS SELECTED ACCORDING TO THEIR SCORES.
2. WRAPPER METHODS: THESE METHODS INVOLVE TRAINING AND EVALUATING THE MODEL WITH DIFFERENT SUBSETS OF FEATURES. THE PERFORMANCE OF THE MODEL IS USED AS FEEDBACK TO SELECT THE MOST RELEVANT FEATURES. THIS PROCESS IS COMPUTATIONALLY EXPENSIVE BUT CAN YIELD BETTER RESULTS.
3. EMBEDDED METHODS: FEATURE SELECTION IS INTEGRATED INTO THE MODEL TRAINING PROCESS. THE ALGORITHM AUTOMATICALLY SELECTS THE MOST RELEVANT FEATURES DURING TRAINING

- We used Augmented Dickey Fuller test to check that is our data stationary or not and we found out that our data is stationary from ADF test as we took only the data from lda. From KPSS test we found out that the given data is non-stationary.

WHAT IS STATIONARY DATA

A data which has following properties is called stationary:

1)Mean is constant
2)Standard Deviation is constant
3)Data is not periodic with time

WHAT IS AUGMENTED DICKEY FULLER TEST

The ADF test is an extension of the Dickey-Fuller test, designed to handle more complex time series by including additional lag terms in the regression equation. The primary objective of the ADF test is to check for the presence of a unit root in the autoregressive model.

- Null Hypothesis (H0): The null hypothesis of the ADF test is that the time series has a unit root, indicating that it is non-stationary.

- Alternative Hypothesis (H1): The alternative hypothesis is that the time series does not have a unit root, suggesting that it is stationary.

- Autoregressive Model: The ADF test involves estimating an autoregressive model of the form $\Delta y_t=\alpha+\beta t+\gamma y_{t-1}+\delta_1 \Delta y_{t-1}+\ldots+\delta_{p-1} \Delta y_{t-p+1}+\varepsilon_t$, where Δ represents differencing, y_t is the time series value at timet, and ε_t is the error term.

- ADF Test Statistic: The test produces a t-statistic, and the decision to reject or not reject the null hypothesis depends on whether this test statistic is less than the critical value corresponding to a chosen significance level.

- Interpretation: If the p-value associated with the test statistic is less than the significance level (commonly set at 0.05), the null hypothesis is rejected, suggesting that the time series is stationary. If the p-value is greater than the significance level, the null hypothesis is not rejected, indicating non-stationarity.

After we have found out that the data is stationary we applied multiple models such as:

- AUTO REGRESSION MODEL(AR)
- AUTOREGRESSIVE INTEGRATED MOVING AVERAGE(ARIMA)
- AUTOREGRESSIVE MOVING AVERAGE(ARMA)
- MOVING AVERAGE(MA)

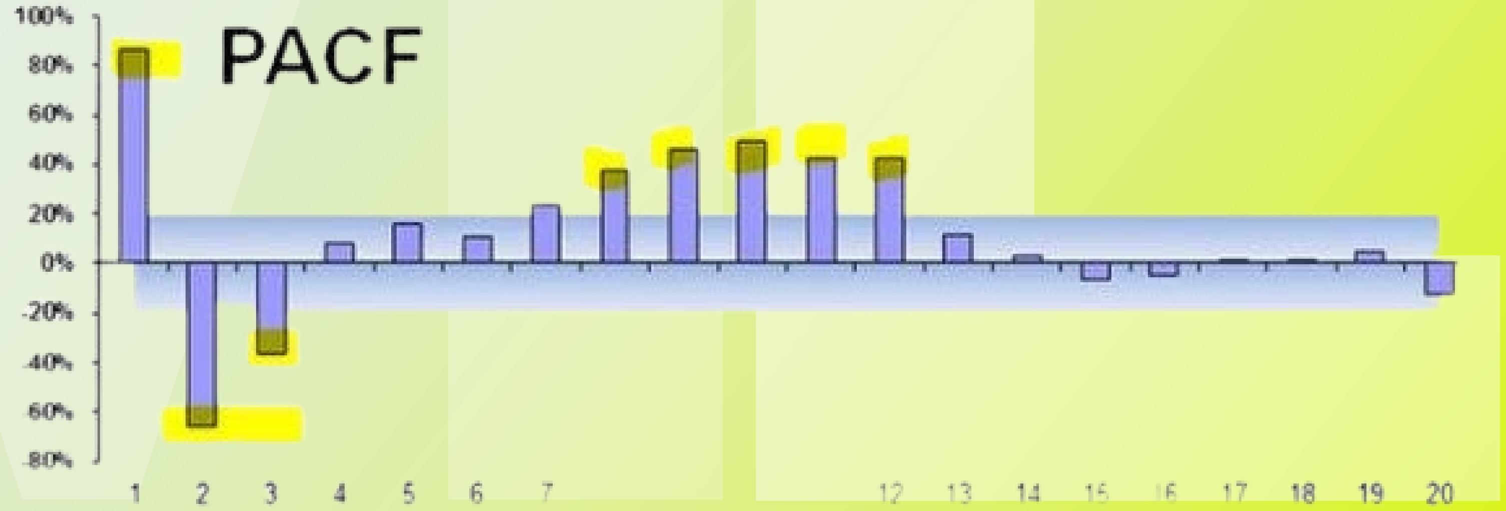
AUTO REGRESSION MODEL

- Auto Regression Model is a time series model which tell the current output by making prediction form the (t-i) output where t is time and i is the past number.

- We use PACF graph to tell which (t-i) output has more impact on the current output and we choose this by finding that output which has value greater than the threshold value.

WHAT IS PACF(PARTIAL AUTOCORRELATION FUNCTION)

PACF measures the correlation between a time series and its past values at a specific lag, excluding the contributions from shorter lags. It essentially represents the direct influence of a past observation on the current observation, without the intermediate influence of other lags.



WHAT IS ARIMA(AUTOREGRESSIVE INTEGRATED MOVING AVERAGE)

ARIMA is an acronym that stands for Auto-Regressive Integrated Moving Average. It is a class of model that captures a suite of different standard temporal structures in time series data.
IN THIS TUTORIAL, WE WILL TALK ABOUT HOW TO DEVELOP AN ARIMA MODEL FOR TIME SERIES FORECASTING IN PYTHON.
AN ARIMA MODEL IS A CLASS OF STATISTICAL MODELS FOR ANALYZING AND FORECASTING TIME SERIES DATA. IT IS REALLY SIMPLIFIED IN TERMS OF USING IT, YET THIS MODEL IS REALLY POWERFUL.
THE PARAMETERS OF THE ARIMA MODEL ARE DEFINED AS FOLLOWS:
P: THE NUMBER OF LAG OBSERVATIONS INCLUDED IN THE MODEL, ALSO CALLED THE LAG ORDER.
D: THE NUMBER OF TIMES THAT THE RAW OBSERVATIONS ARE DIFFERENCED, ALSO CALLED THE DEGREE OF DIFFERENCE.
q: The size of the moving average window, also called the order of moving average

WHAT IS MA(MOVING AVERAGE)

A **MOVING AVERAGE** IS A METHOD USED TO SMOOTH OUT VARIATIONS IN DATA BY CALCULATING THE AVERAGE OF SUCCESSIVE SUBSETS OF DATA POINTS WITHIN A SPECIFIED WINDOW OR PERIOD.IT HELPS IN IDENTIFYING UNDERLYING TRENDS OR PATTERNS WITHIN THE SERIES BY REDUCING THE IMPACT OF RANDOM FLUCTUATIONS OR NOISE.

THERE ARE 3 TYPES OF MA:

1)**SIMPLE MOVING AVERAGE (SMA)**- IT'S CALCULATED BY TAKING THE AVERAGE OF A FIXED NUMBER ('N') OF THE MOST RECENT DATA POINTS.

2)**CUMULATIVE MOVING AVERAGE (CMA)**-THE CMA IS THE UNWEIGHTED MEAN OF PAST VALUES TILL THE CURRENT TIME.

3)**EXPONENTIAL MOVING AVERAGE (EMA)**- IT GIVES MORE WEIGHT TO RECENT OBSERVATIONS WHILE EXPONENTIALLY DECREASING THE WEIGHT OF OLDER OBSERVATIONS.

WHAT IS ARMA(AUTOREGRESSIVE MOVING AVERAGE)

ARMA is a class of statistical models commonly used in time series analysis and forecasting. It is widely employed in the field of machine learning for handling and predicting sequential data, where the order and temporal dependencies of observations matter.

The ARMA model is often denoted as ARMA(p, q), where "p" represents the order of the autoregressive component, and "q" represents the order of the moving average component. By adjusting these orders, the model can be tailored to different time series patterns.

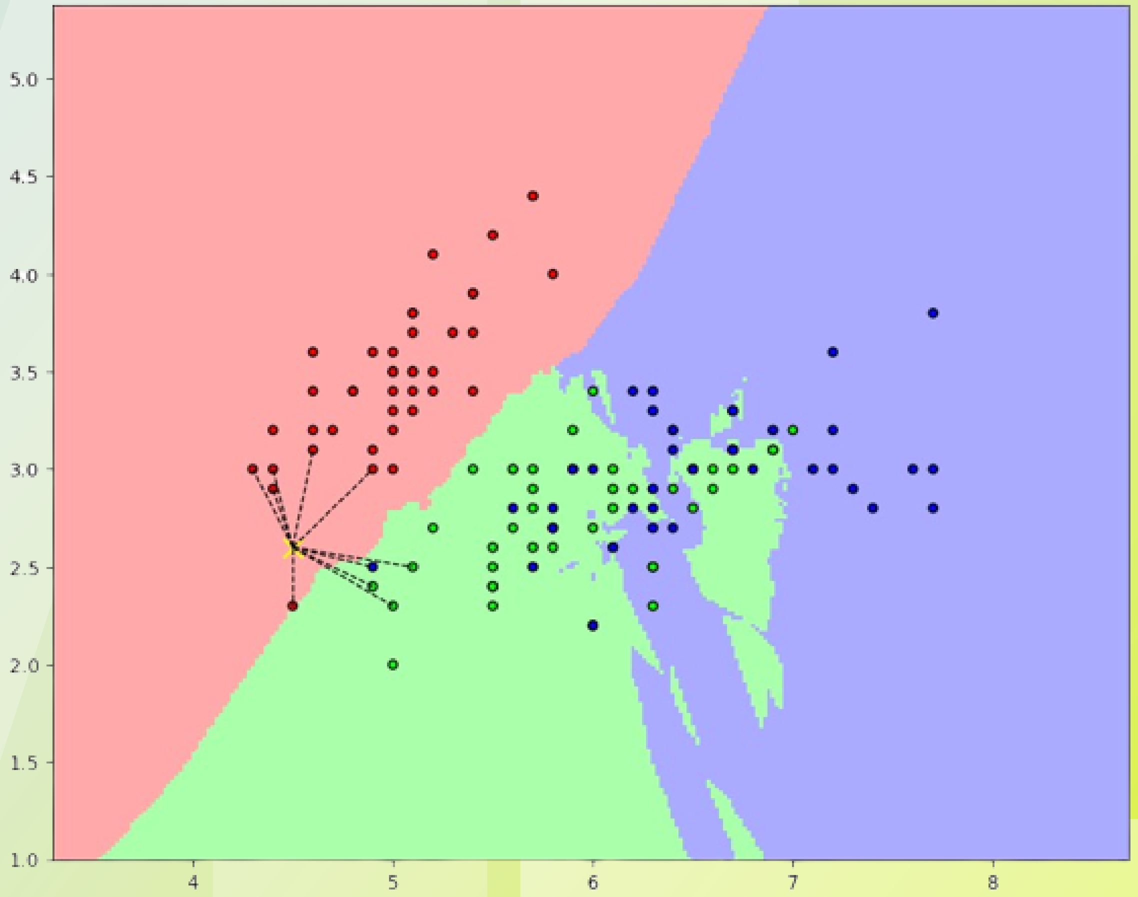
ARMA models are useful for capturing short- and long-term patterns in time series data, making them valuable tools in applications such as finance, economics, and environmental science. It's worth noting that ARMA models assume stationarity in the time series data, meaning that statistical properties such as mean and variance remain constant over time. If stationarity is not satisfied, data transformation or differencing may be applied to meet this assumption.

PROPOSED MODEL-1 KNN(K-NEAREST NEGHBORS)

- Considering the problem statement, we believe that K-nearest neighbors (KNN) could be an appropriate selection. KNN is a supervised learning model commonly employed for classification tasks.

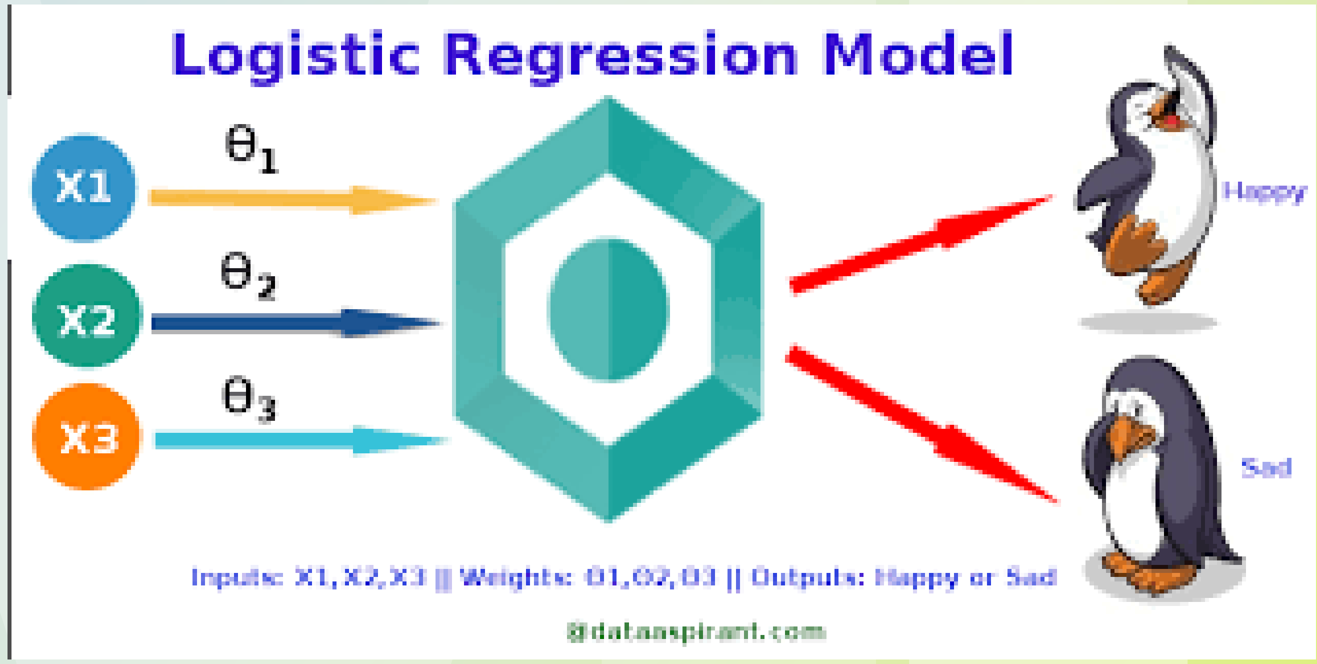
WHY KNN?

- Given a new data point, K-NN identifies the K-nearest neighbors from the training set based on a distance metric, usually Euclidean distance.
- The class label that is most common among these neighbors is assigned to the new data point as the predicted output.



PROPOSED MODEL-2 LOGISTIC REGRESSION

- Logistic Regression is a machine learning model that is primarily used for binary classification tasks, where the goal is to predict one of two possible outcomes (e.g., yes/no, 0/1). However, it can also be extended to multi-class classification.

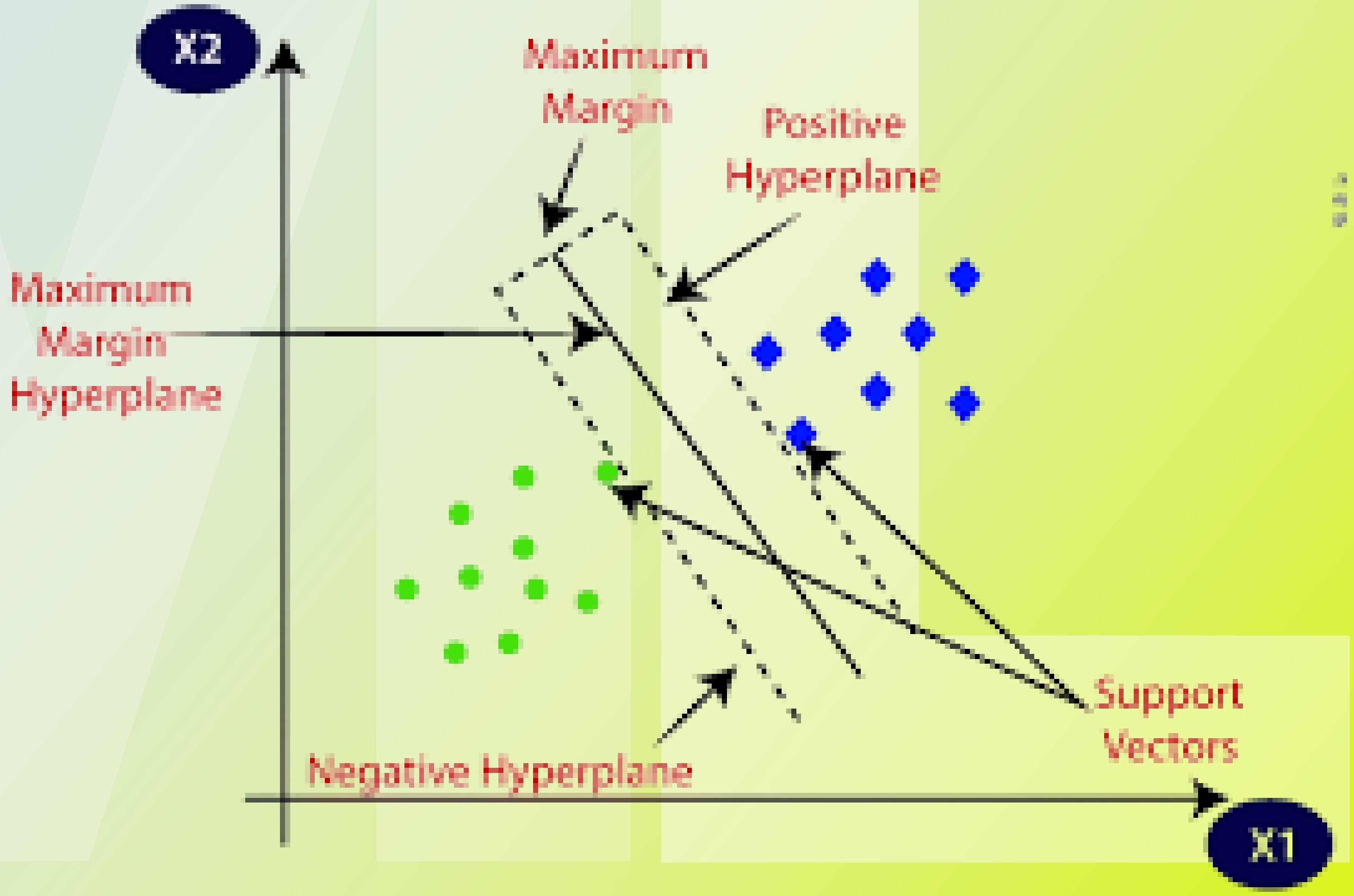

Logistic Regression Model
X1
X2
X3
θ_1
θ_2
θ_3
Happy
Sad
Inputs: X1,X2,X3 || Weights: θ1,θ2,θ3 || Outputs: Happy or Sad
@dataaspirant.com

WHY LOGISTIC REGRESSION?

- Logistic regression can be a suitable choice for smartphone-based recognition of human activities due to its interpretability, efficiency, and adaptability for both binary and multi-class classification tasks.

PROPOSED MODEL-3 SUPPORT VECTOR MACHINE

A Support Vector Machine (SVM) is a supervised machine learning algorithm used for classification and regression tasks. It works by finding a hyperplane that best separates data points into distinct classes in a way that maximizes the margin between the classes. Here's a brief explanation of the key concepts:

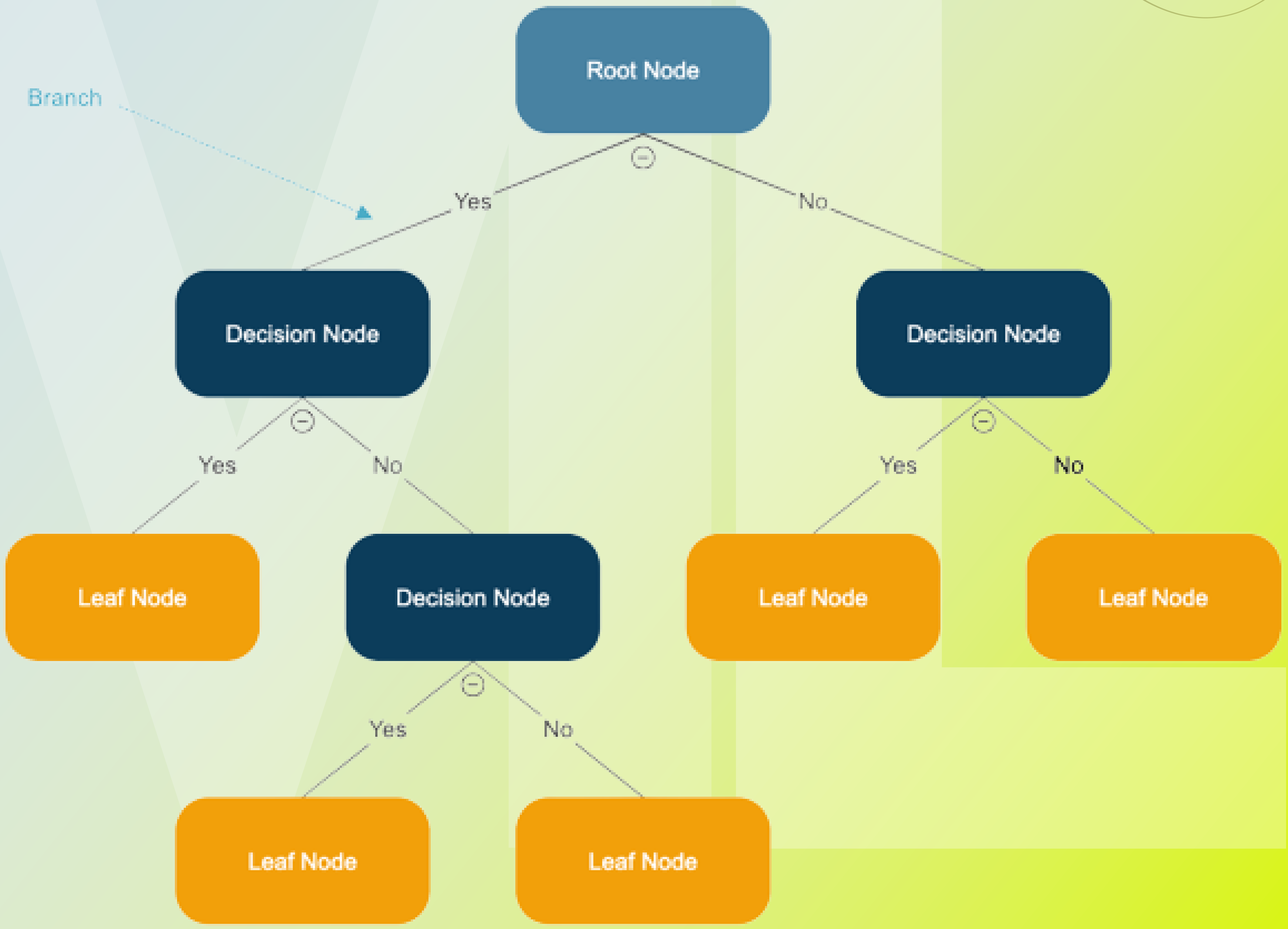


WHY SVM?

- Third model which we decided is SVM.
- SVM is the classification algorithm used to classify new data point using the hyperplane and the margin.
- SVM aims to find the optimal hyperplane that best separates different classes in the data. In the context of classification, this hyperplane maximizes the margin between the two classes.

PROPOSED MODEL-4 DECISION TREE CLASSIFIER

- A decision tree is a supervised machine learning algorithm used for both classification and regression tasks. It models decisions or predictions using a tree-like structure, where each internal node represents a feature or attribute, each branch represents a decision or condition, and each leaf node represents an outcome or class label.
- The decision tree is built through a recursive process that selects the best features and conditions to split the data until a stopping criterion is met, such as a maximum tree depth or a minimum number of data points in a node. Decision trees are interpretable, easy to visualize, and can handle both categorical and numerical data.



WHY DECISION TREE?

For classification, the leaves of the Decision Tree represent class labels. The majority class in a leaf node is the predicted class.

We finally relied on our Time Series Analysis models as it was time series data

THANK YOU