Predictive and Failure Analyzing:**Student Placement Predicter and Analyzer**

Guide:
Prof. Medha Wyavahare
E&TC Dept.

Project By:

Rishikesh Devarmani-03

Harshit Sarna-08

Nikhil Sonavane-10

Shravan Sriram-12

Shoumik Nandi-

- **Motivation for this project:**
  - The implementation of our project is aimed to understand and suggest positive changes on the placement procedures conducted by colleges.
  - We wanted to undertake a task of applying various machine learning techniques and using effective data analysis and data mining techniques and deliver a output to support a social cause.
  - With the current generation of students struggling to get employed, such a project would boost their probability of being placed in a company.
  - Helping the students to get out from the confused state and giving them a clear view of what is to be done in order to achieve the desired result.
  - It is a new ,unique and socially relevant idea , which hasn’t seen any implementation anywhere.

OUR AIM

Our aim is to consider a set of input academic related parameters
And predict the companies that a student will be or most likely will be placed in.

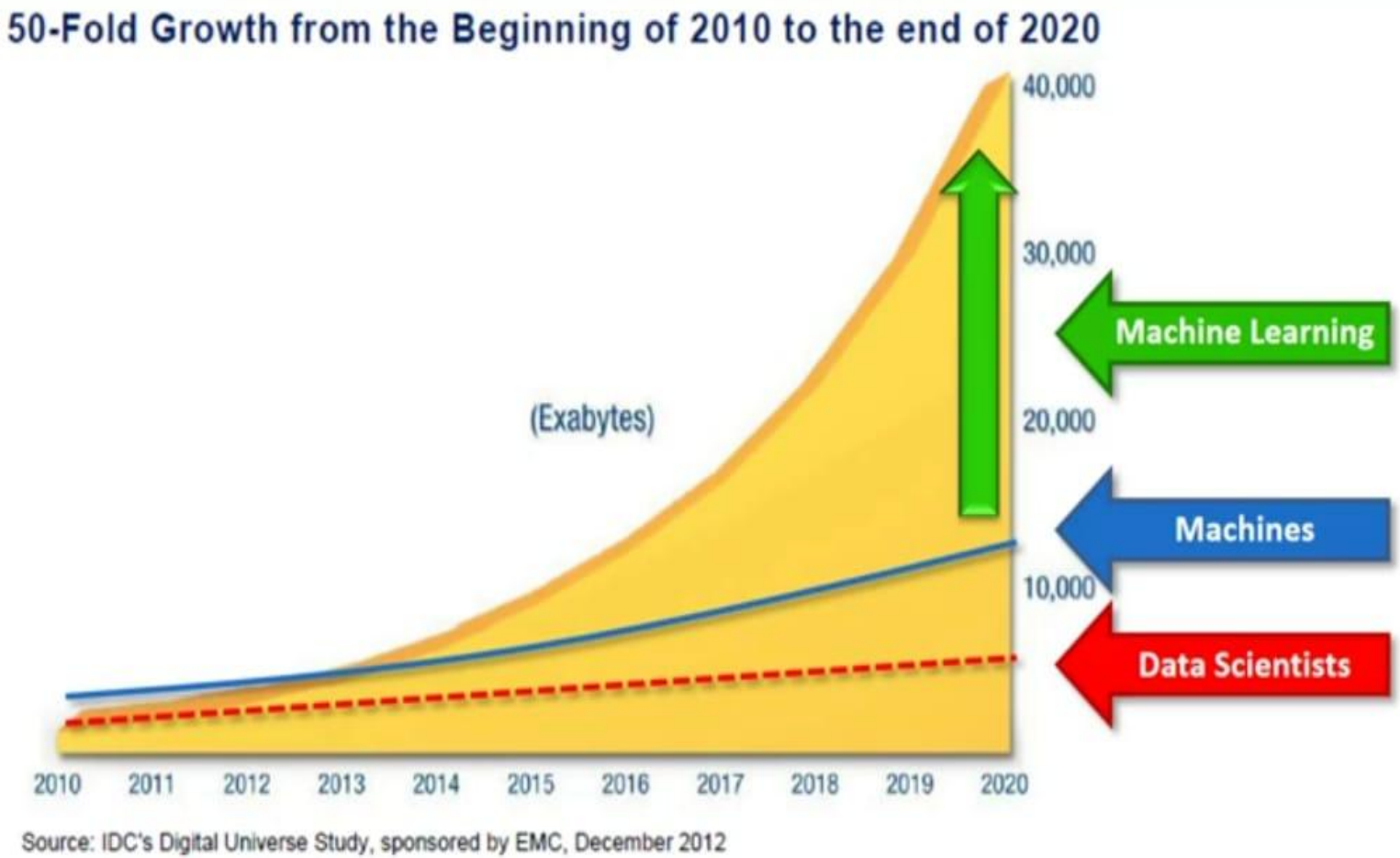
Why this domain? Why Machine Learning is the future?

2005 – 130 EXABYTES

2010 – 1,200 EXABYTES

2015 – 7,900 EXABYTES

2020 – 40,900 EXABYTES


50-Fold Growth from the Beginning of 2010 to the end of 2020
40,000
30,000
20,000
10,000
(Exabytes)
Machine Learning
Machines
Data Scientists
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
Source: IDC's Digital Universe Study, sponsored by EMC, December 2012

- **Introduction**:
  - With a exponential increase in population, the annual numbers of undergraduate degree holders have followed the same curve in the last two decades.
  - This sudden rise in demand for placements in companies and post graduation seats have resulted in stiff competition amongst graduates.
  - Companies keep an eye on candidate for his choice of subjects and technical experience that he has gained over the 4 years of his undergraduate course.
  - The domain of choice for projects, internships and the co-curricular, extra curricular activities are major plus points for an individual.

- The Student Career Analyzer assists a candidate to select a company from a select list of companies when he’s sitting for placements in his final year.
- The algorithm takes into consideration the academic qualities gained by a student as its main input parameters and when he or she enters their academic data, the algorithm would make the appropriate class of companies for which they could appear for the interviews.
- It comes under the domain of Data Analytics and Machine Learning and the algorithms are applied to achieve the best results.

This is a Predictive System!

Not a Recommender System!

The Database

Arguably the most important part in a Machine Learning Model.

Why is Google at the Top?

We are using the same Algorithms that they do.

Having a huge database is not only key point,

but the way in which data is provided to the machine is.

Increasing the accuracy is in the hands of the developer and not the machine.

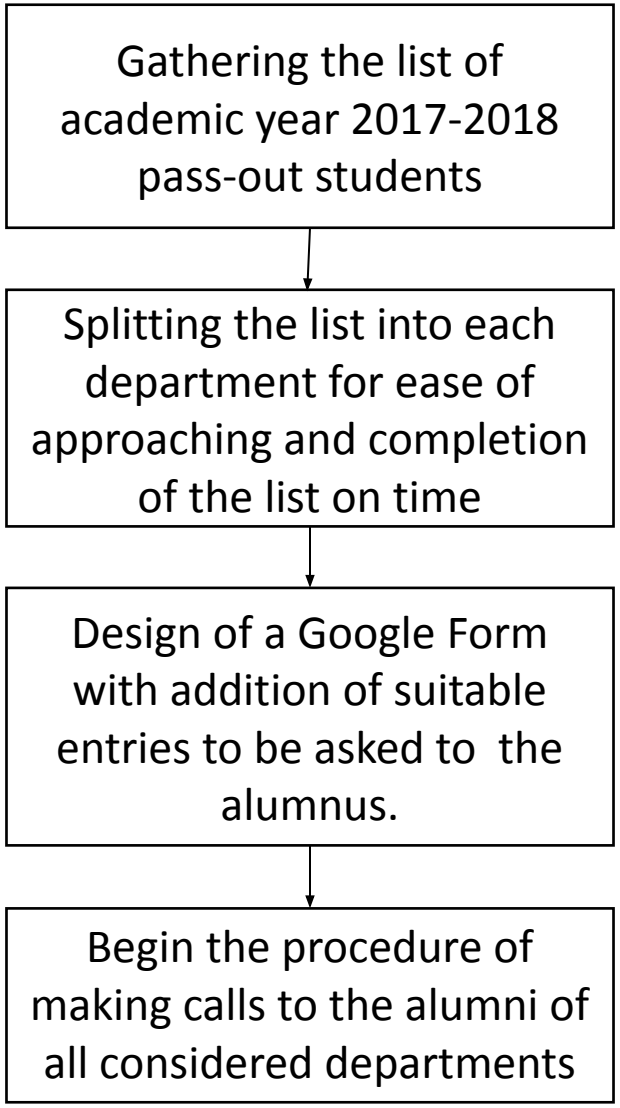
You can have two different accuracies for the same database.

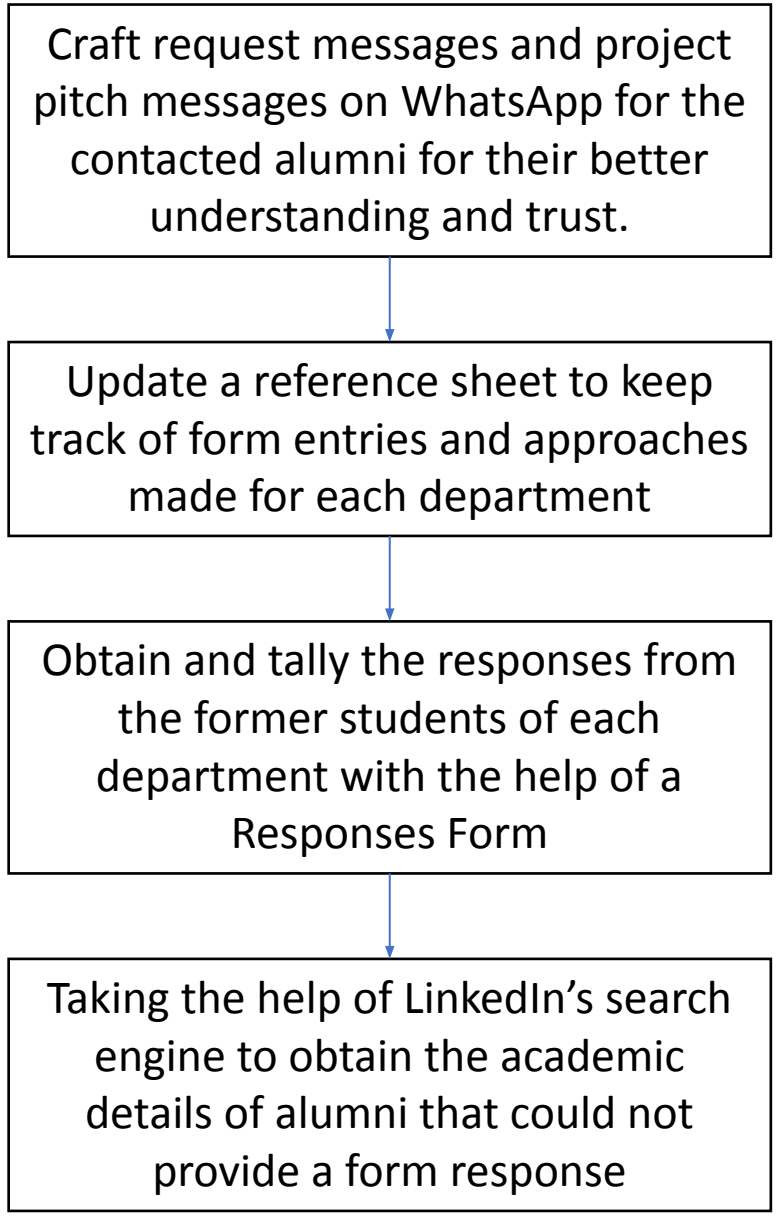
The Database

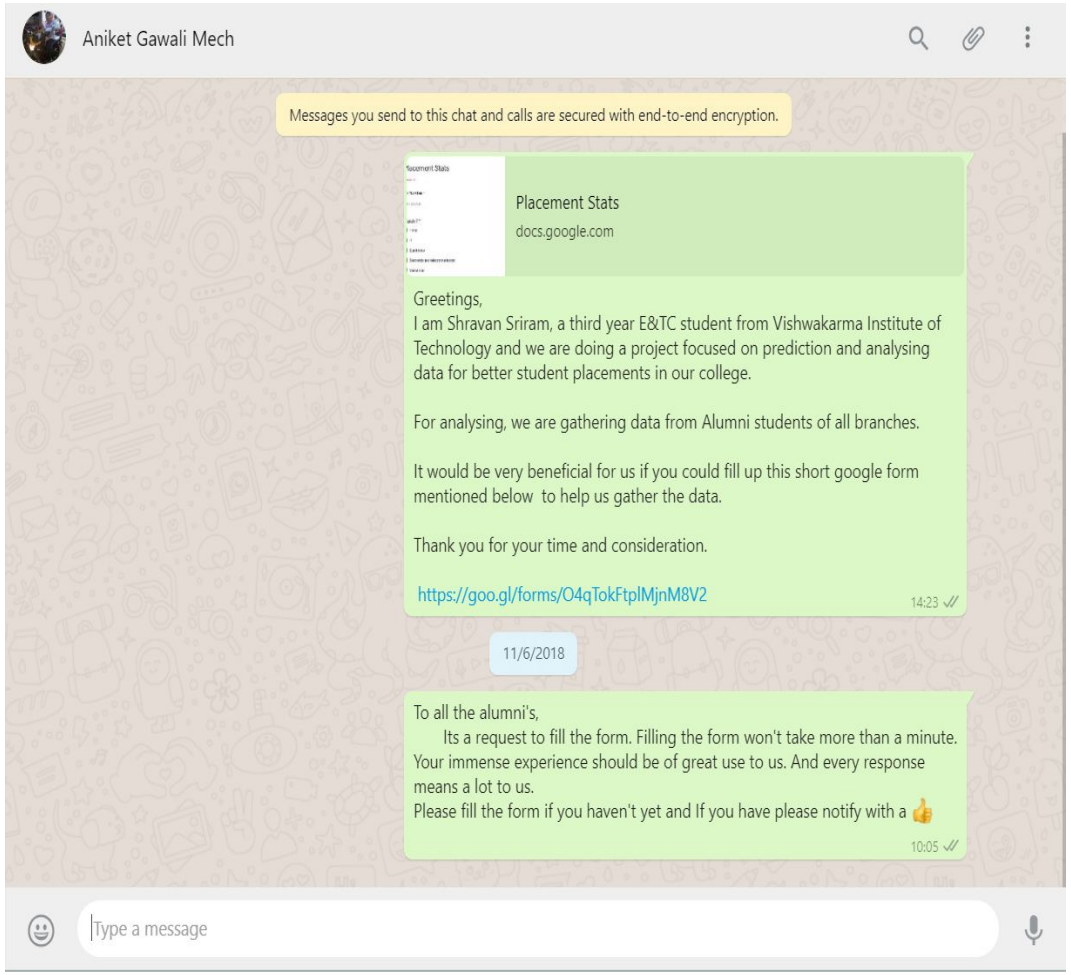
Creating the proper database consists of four steps:

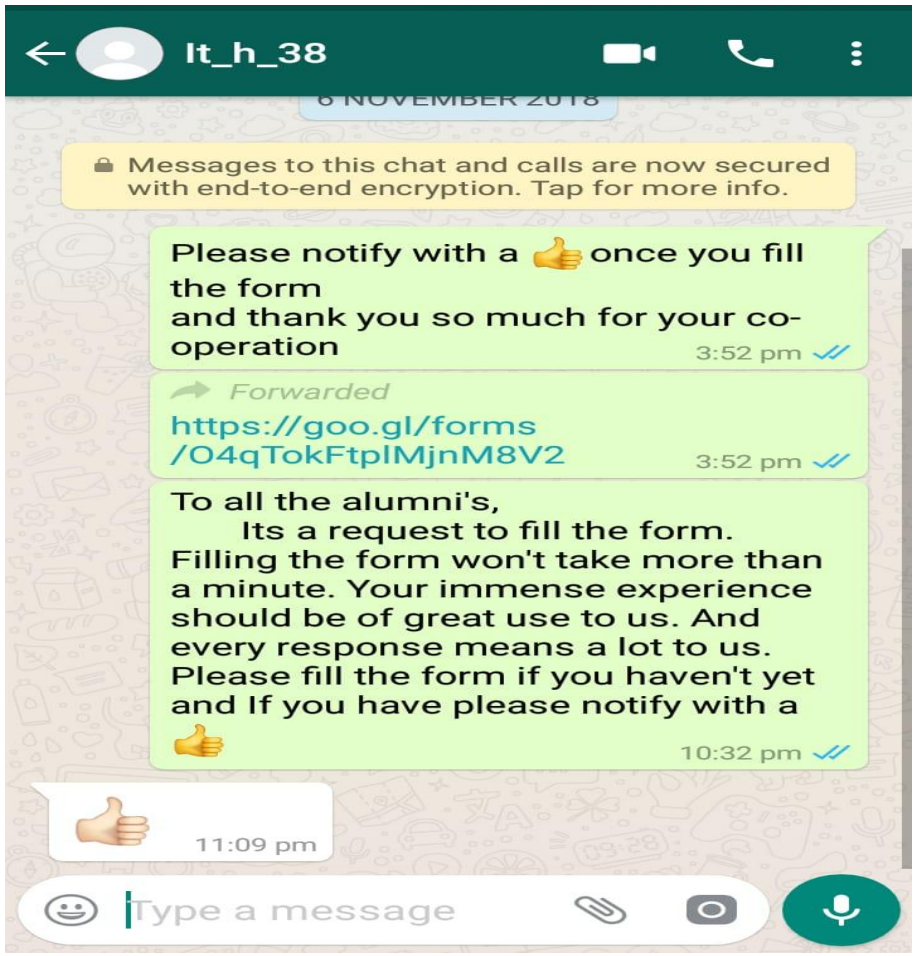
1. Data Collection
2. Data Preprocessing
3. Data Manipulation
4. Data Optimization

1. Data Collection:

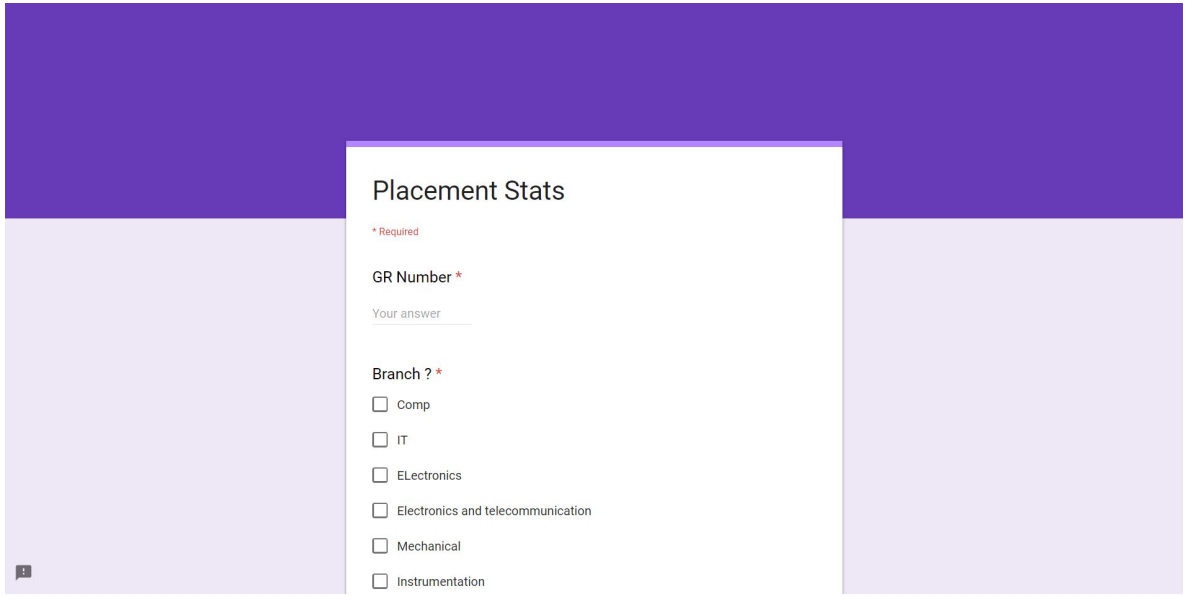


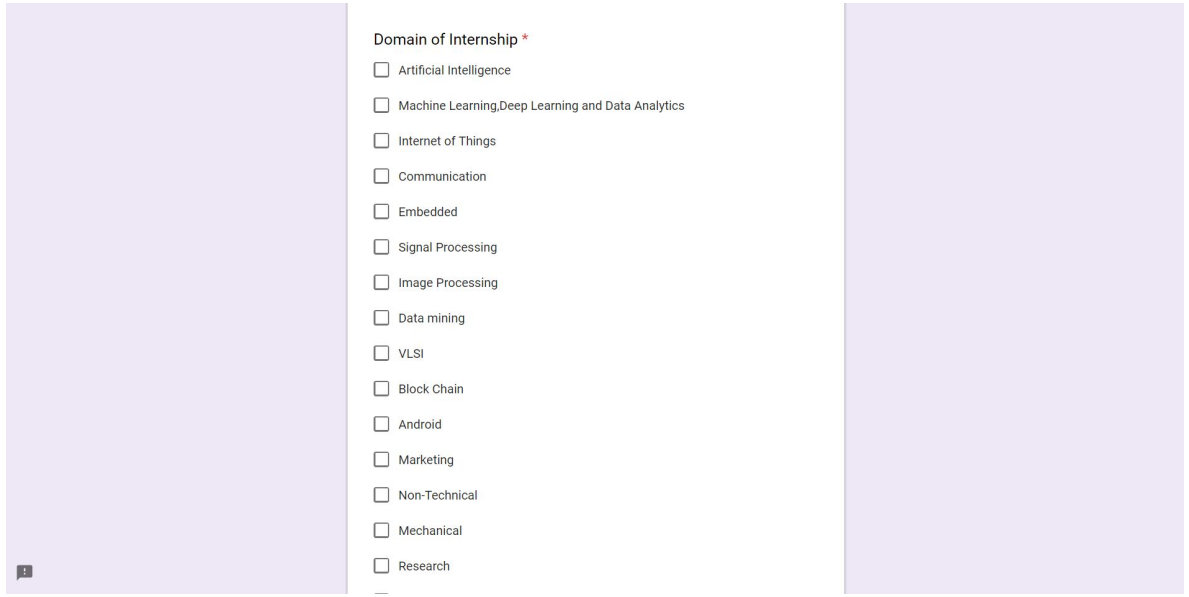

Craft request messages and project pitch messages on WhatsApp for the contacted alumni for their better understanding and trust.
Update a reference sheet to keep track of form entries and approaches made for each department
Obtain and tally the responses from the former students of each department with the help of a Responses Form
Taking the help of LinkedIn’s search engine to obtain the academic details of alumni that could not provide a form response

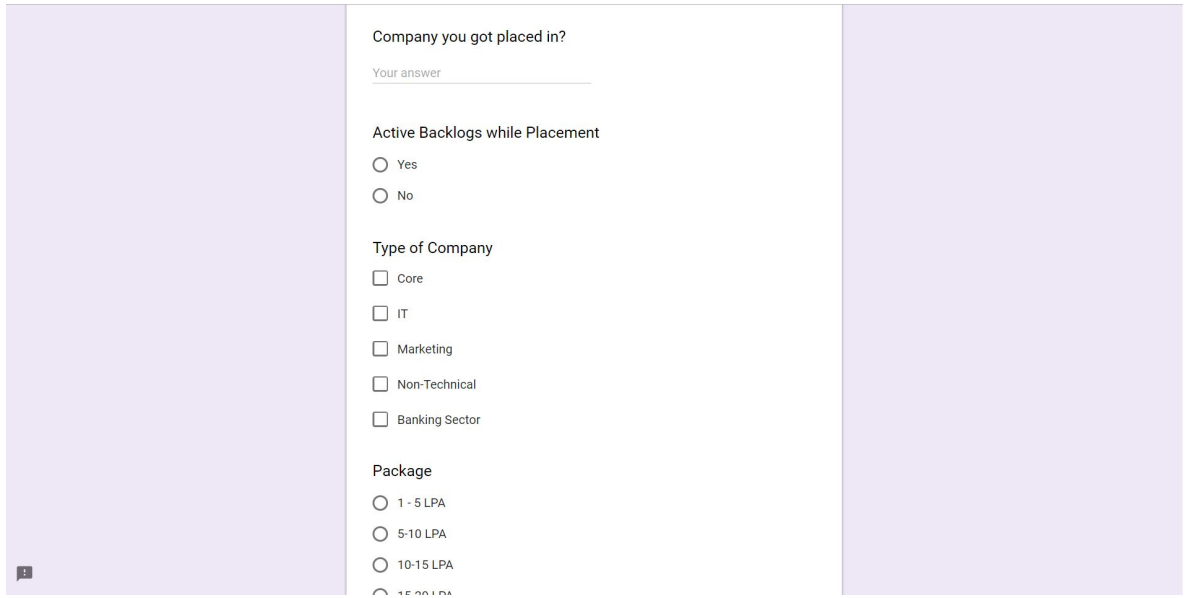


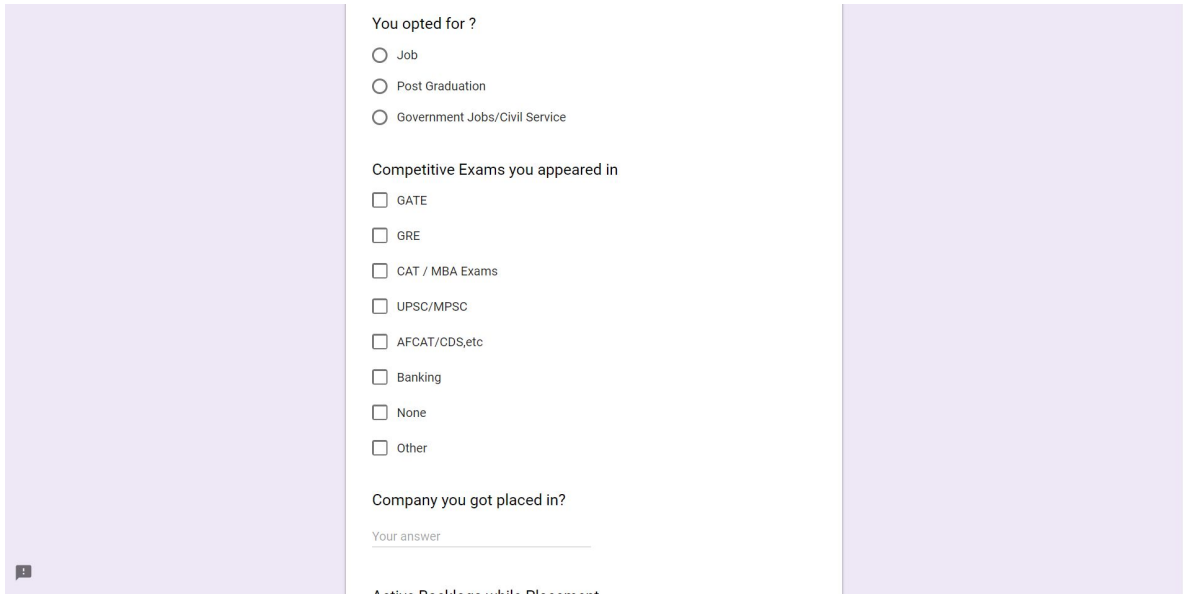


An image in reference to the message drafted on WhatsApp for the alumni of various departments









Reference images for the Google Form that had been circulated to the alumni for them to provide their response

Branch ?	Domain of Project	Domain of Internship	Courses
Electronics and telecomm	Machine Learning,Deep Learning and Data Analytics, Embedded, Signal Processing, Image Processing	Autocad	
Comp	Machine Learning,Deep Learning and Data Analytics	Other	Machine Learning,Deep
Comp	Machine Learning,Deep Learning and Data Analytics	Machine Learning,Deep L	Machine Learning,Deep
Comp	Artificial Intelligence, Machine Learning,Deep Learning and Data Analytics, Image Processing, Research	Machine Learning,Deep L	Artificial Intelligence, Ma
Comp	Machine Learning,Deep Learning and Data Analytics, Internet of Things, Block Chain, Android	Block Chain, Android	Block Chain, Android De
Comp	Machine Learning,Deep Learning and Data Analytics	Other	Machine Learning,Deep
Comp	Image Processing, Data mining, Android	Other	Android Development, V
Comp	Research	Research	
Comp	Machine Learning,Deep Learning and Data Analytics	Other	
Comp	Machine Learning,Deep Learning and Data Analytics	Internet of Things	Artificial Intelligence, Ma
Comp	Machine Learning,Deep Learning and Data Analytics	Other	C,C++, JAVA
Comp	Android	None	Android Development, V
Comp	Android	Other	Artificial Intelligence, An
Comp	Internet of Things	Other	Artificial Intelligence, C,(
Comp	Machine Learning,Deep Learning and Data Analytics	None	Artificial Intelligence, Ma
Comp	Image Processing	Other	Web Development, JAV
Comp	Machine Learning,Deep Learning and Data Analytics, Image Processing, Data mining	Machine Learning,Deep Learning and Data Analyt	
IT	Internet of Things, Image Processing, Android	None	
IT	Machine Learning,Deep Learning and Data Analytics	Other	Machine Learning,Deep
IT	Machine Learning,Deep Learning and Data Analytics	Other	Machine Learning,Deep
IT	Research	Research	Machine Learning,Deep
Comp	Internet of Things	Other	Artificial Intelligence, Int
Comp	Artificial Intelligence, Machine Learning,Deep Learning and Data Analytics, Android	None	Artificial Intelligence, Ma
Comp	Machine Learning,Deep Learning and Data Analytics	Non-Technical, Research	Artificial Intelligence, Ma
Comp	Machine Learning,Deep Learning and Data Analytics, Internet of Things, Image Processing, Data mining	Other	Artificial Intelligence, We
Comp	Machine Learning,Deep Learning and Data Analytics, Internet of Things, Image Processing	Other	Artificial Intelligence, We

Reference image for the data that was collected from the responses of alumni in excel format

- The process of data collection by call was successful in generating 320 responses out of around 550 alumnus from the considered departments.
- The remaining alumni out of 550, were unable to respond to the form due to various reasons;
- LinkedIn, one of the most renowned websites in terms of communication and connection between students, individuals and companies was utilized to search for the data of the alumnis, who couldn’t fill the Google form due to a variety of reasons.
- Every graduated student from any field or stream or domain does tend to have an updated Resume and experience list on his or her LinkedIn account and we were able to successfully generate data of 120 students.
- These two operations combined resulted us in a strong number of 440 responses with which we took the next step forward of processing the data.

GR no.	Branch	Domain Of Projects	Domain Of Internship	Courses Done	Job or PG	Company placed in	Active Backlogs	Type of Company	PG Exams appeared for	Package	CPI	Extra-Cirricular Activities

2. Data Preprocessing

- We have many Categorical variables in the database and it would cause problems if we used them in the equations.
- The machine will take only integer values.
- So we need to convert these categorical values to integer values.
- This is where we use two functions:
  i. **LabelEncoder**
  ii. **OneHotEncoder**

1. LabelEncoder

CLASS **A**
CLASS **B**

1
2

But here, the system will interpret that Class B is better than Class A, which is not the case. For us these classes are just two categories and nothing else

2. OneHotEncoder

1
2

1	**2**
1	**0**
0	**1**

3. Data Manipulation

- Data Manipulation is done in every Machine Learning Model.
- It can increase the accuracy of the prediction tremendously and also makes the computation easier for the system.
- Data manipulation cannot be done by the machine.
- It doesn’t have a specified process like other Data Creating parts. It depends on the model.
- In our case, it was the toughest part to do in the Data Creating section.
- In our Model, we have done Data Manipulation , on two of our input variables and our output.

1. Domain of Projects
   i. After seeing the patterns and similarities of the domain of projects, these projects were grouped into groups of 2 or 3.
   ii. And if the student has done any of the project from that domain, it has been assigned as **1** and if not , then **-1**.

ML/DL/DA
AI/BC/DM
VLSI/SP/COMM
IP/Embed
IOT/Android
RES/Analysis
TOM/Mech
Gears/GT

2. Domain of Projects

i. In this the internships are grouped seeing the pattern and then encoded.

ii. For eg., Machine Learning is encoded as-> ML

Artificial Intelligence as-> AI and so on…

3.Company Classes

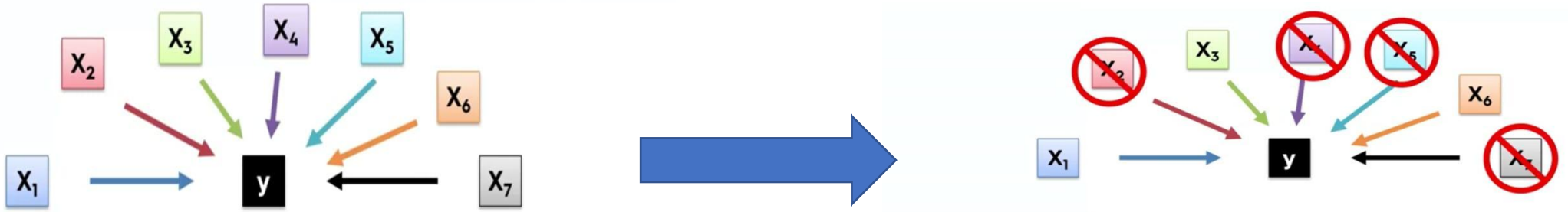
i. We had a list of companies along with their CTC's that came for college placement for the academic year 2017-2018.

ii. These companies were then divided into 6 different classes(A,B,C,D,E,F) depending upon CTC ranges.

iii. Each class of companies was further divided into sub classes , using the criteria of Branch. For eg., Class B->Branch E&tc.

4. Data Optimization

- Not all the input parameters affect the output.



So why to keep these variables?

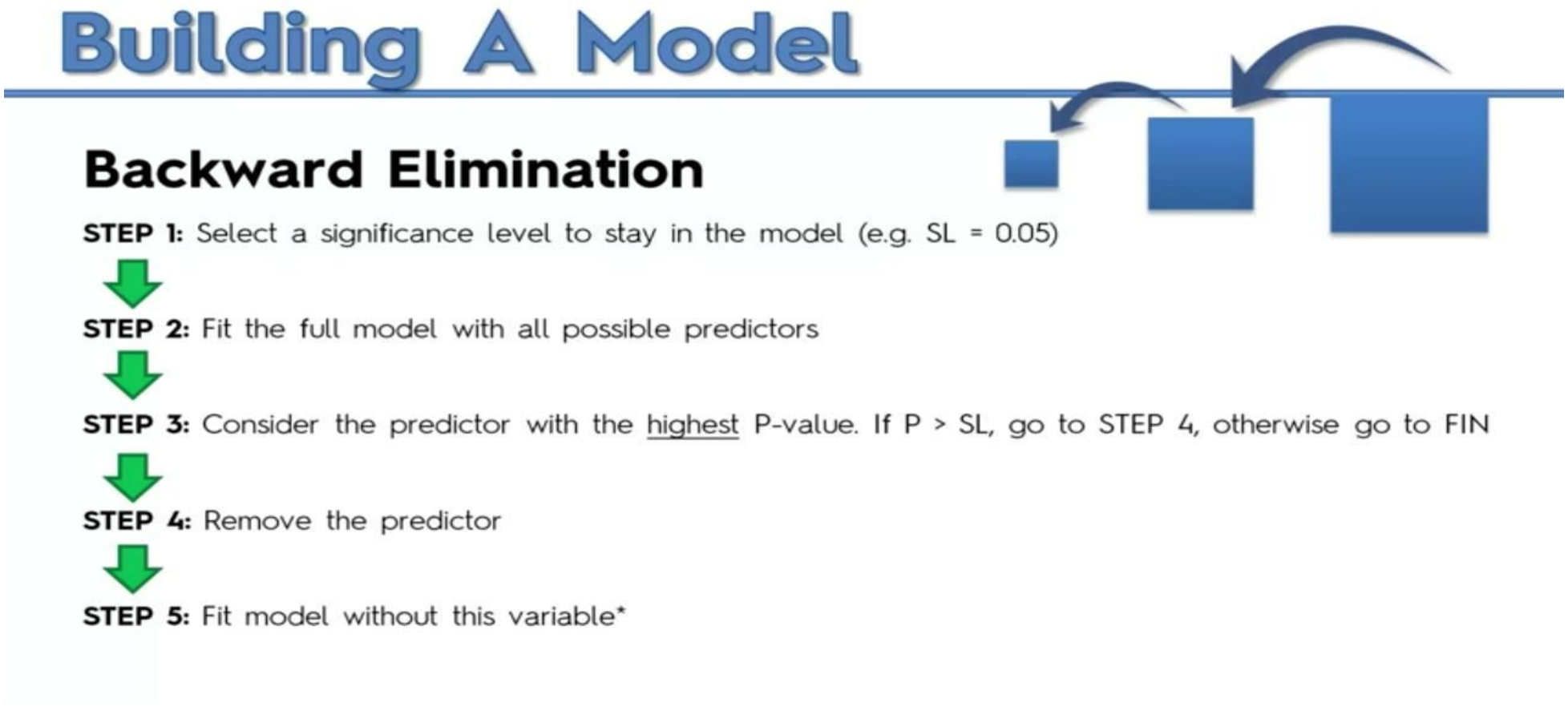
- It’s better to discard them and the reason to that is:
  i. Garbage in=Garbage out . If you add unnecessary data in your model, the model will predict wrong outputs and will have a bad prediction accuracy.
  ii. Computation becomes easier.

Now to build the Model, we use any of the following 5 methods:

1. All-in
2. Backward Elimination
3. Forward Selection
4. Bidirectional Elimination
5. Score Comparison

Here we will be only talking about Backward Elimination

- Backward Elimination:



Using this from our original database, three of the input parameters were removed as they did not significantly affect the output. These three parameters are:

i. Courses Completed
ii. Post Graduation Exams Appeared for
iii. Extra-Cirricular Activities

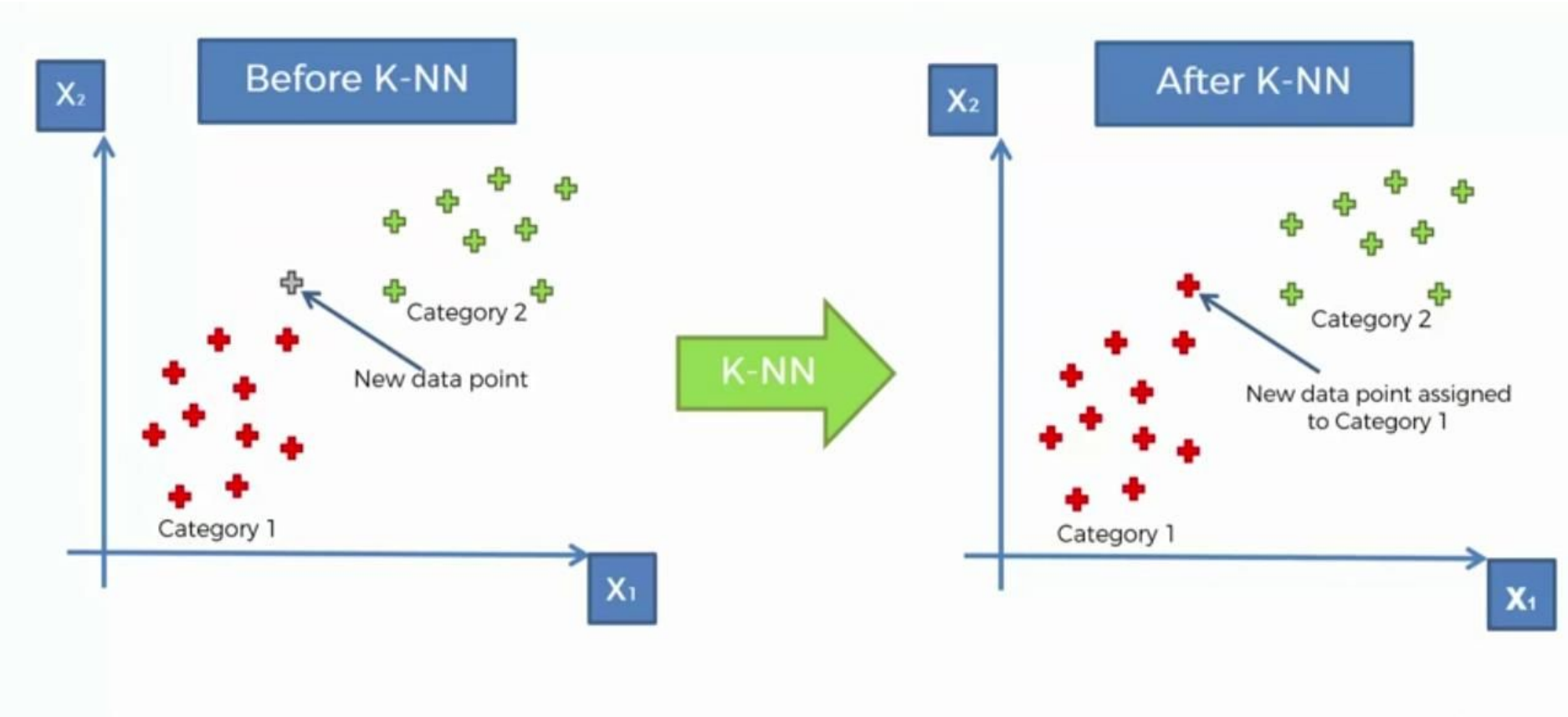
Test and Training set split

- The size of the training and test set is user specified
- Usually the best is 25% ->Test set and 75% -> Training set.
- The elements for Test set and Training set are chosen randomly by the machine.

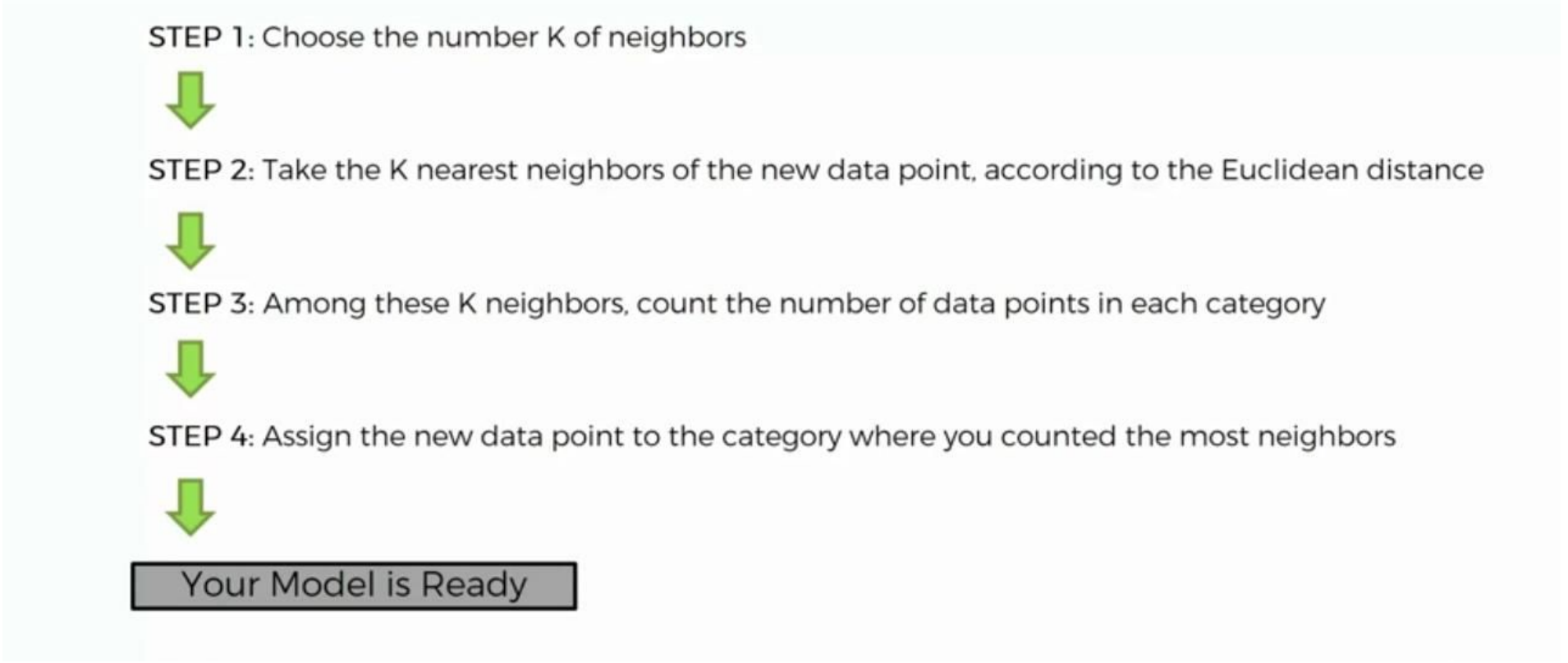
Algorithms

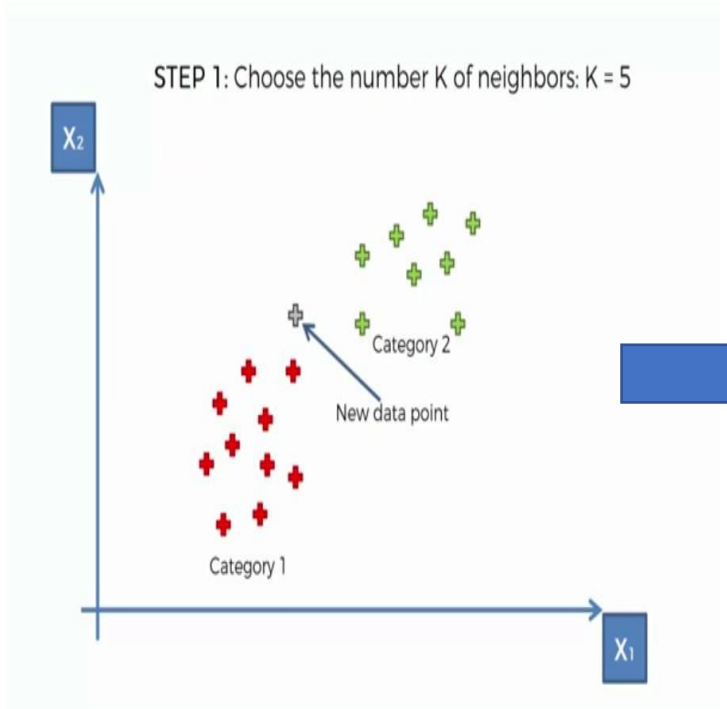
Now we have cleaned, optimized, processed and completed our data and hence now we will apply the algorithms on the data and get results .

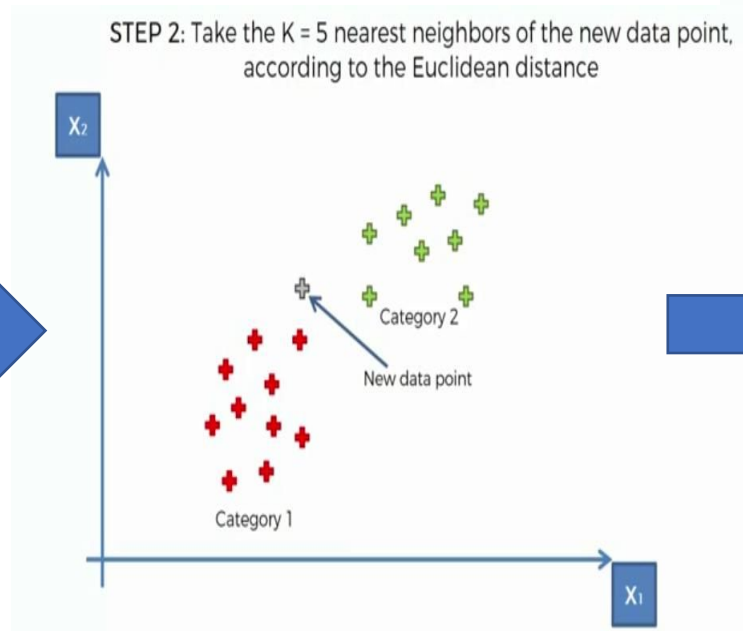
1.K-Nearest Neighbor Algorithm

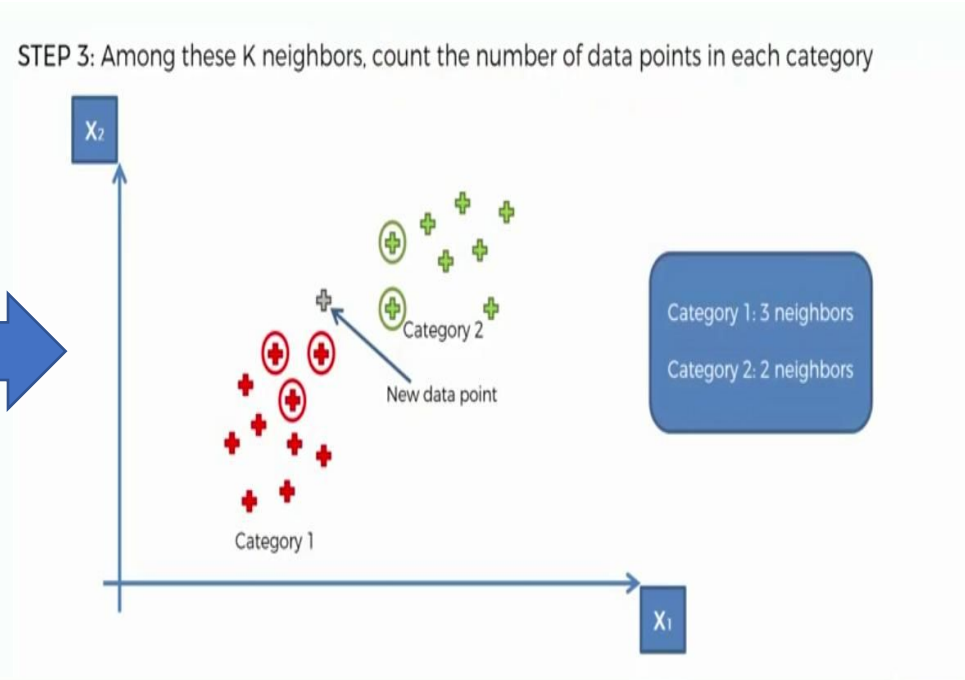


The steps to do K-NN Classification is

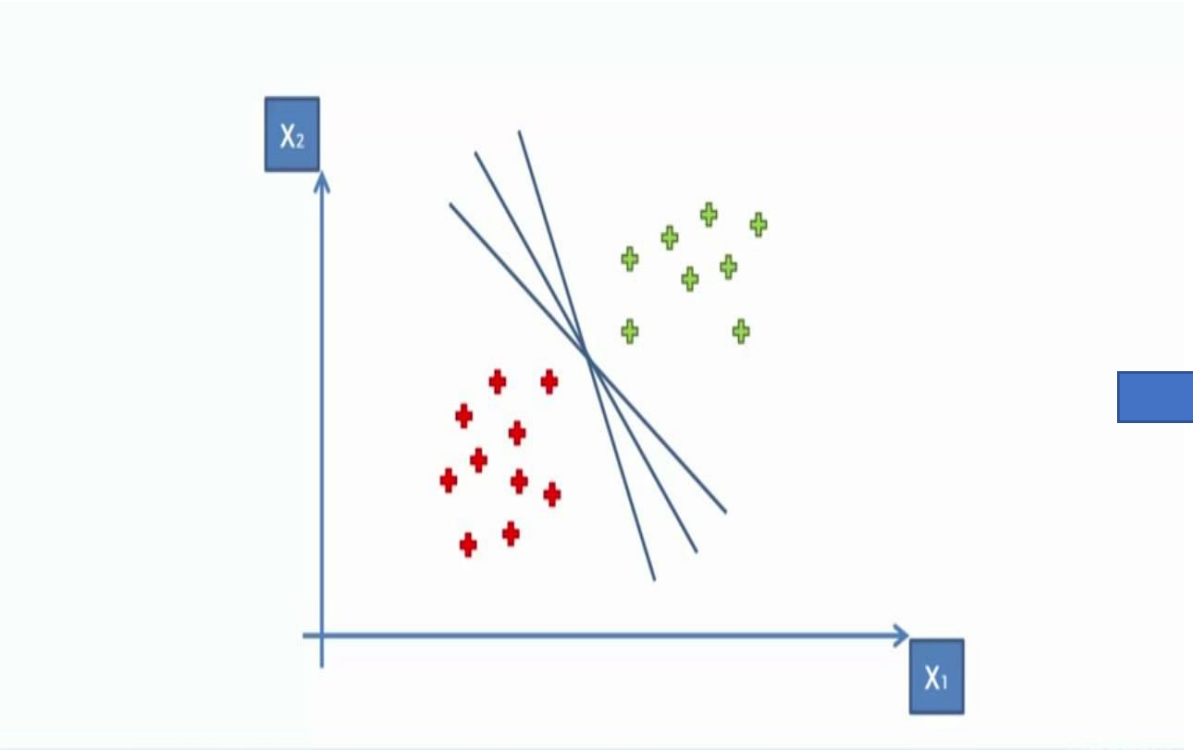


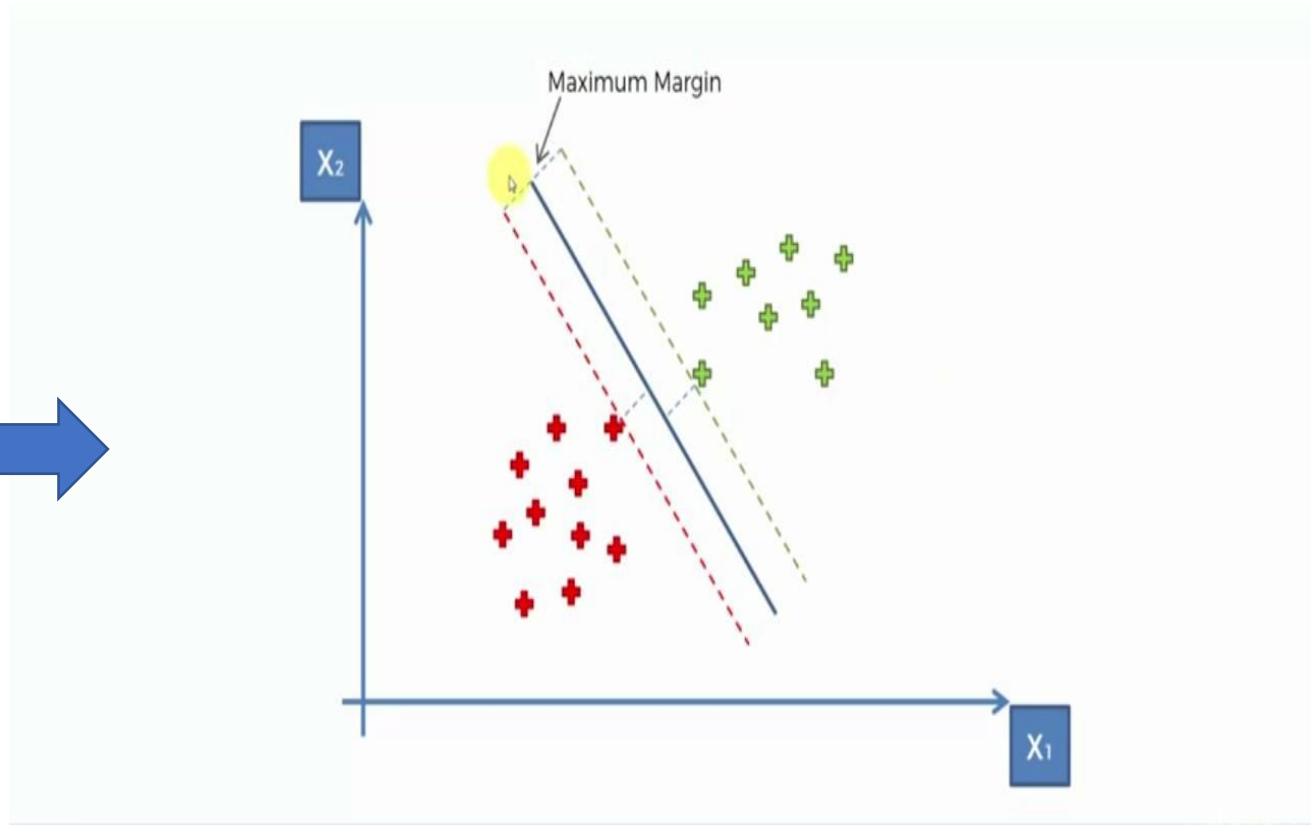

STEP 1: Choose the number K of neighbors: K = 5
X2
Category 2
New data point
Category 1
X1


STEP 2: Take the K = 5 nearest neighbors of the new data point, according to the Euclidean distance
X2
Category 2
New data point
Category 1
X1

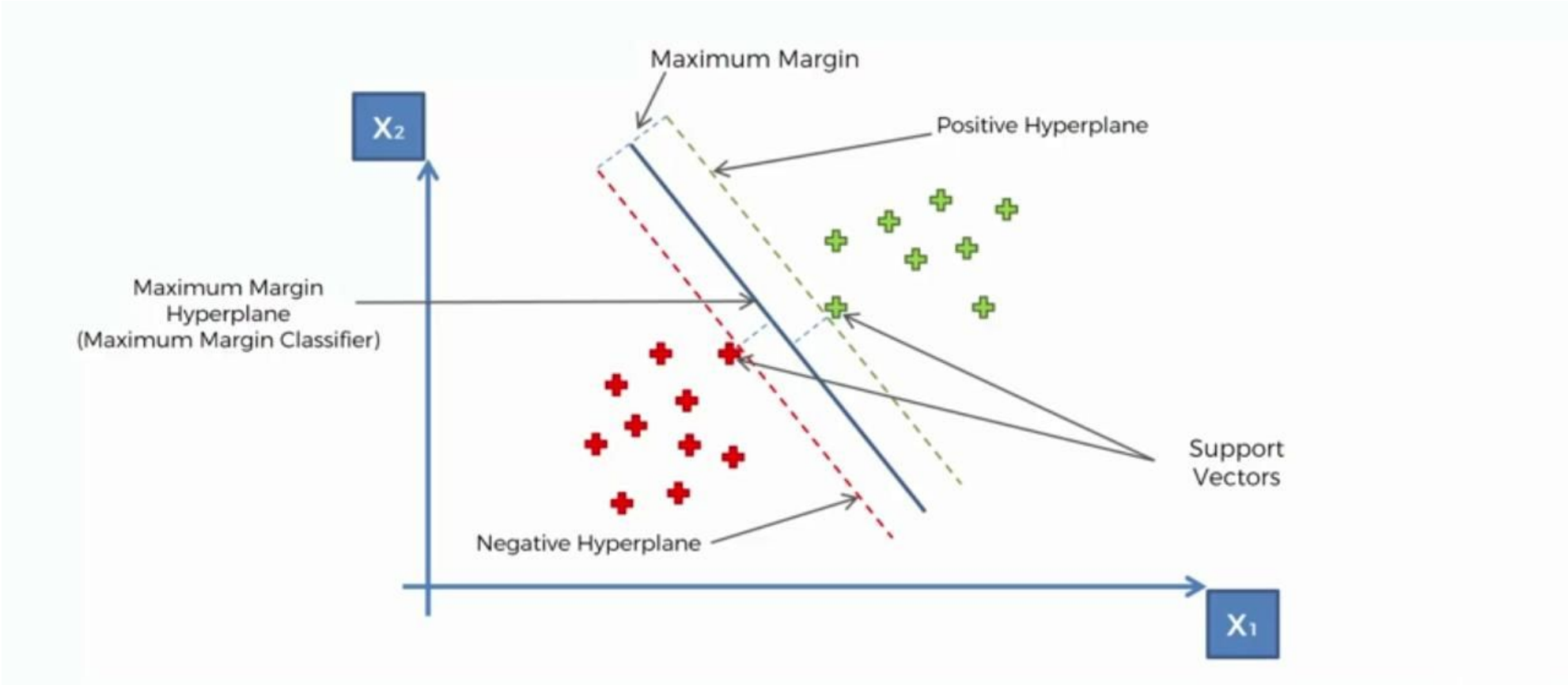

STEP 3: Among these K neighbors, count the number of data points in each category
X2
Category 2
New data point
Category 1
X1
Category 1: 3 neighbors
Category 2: 2 neighbors

2.Support Vector Machine(SVM)

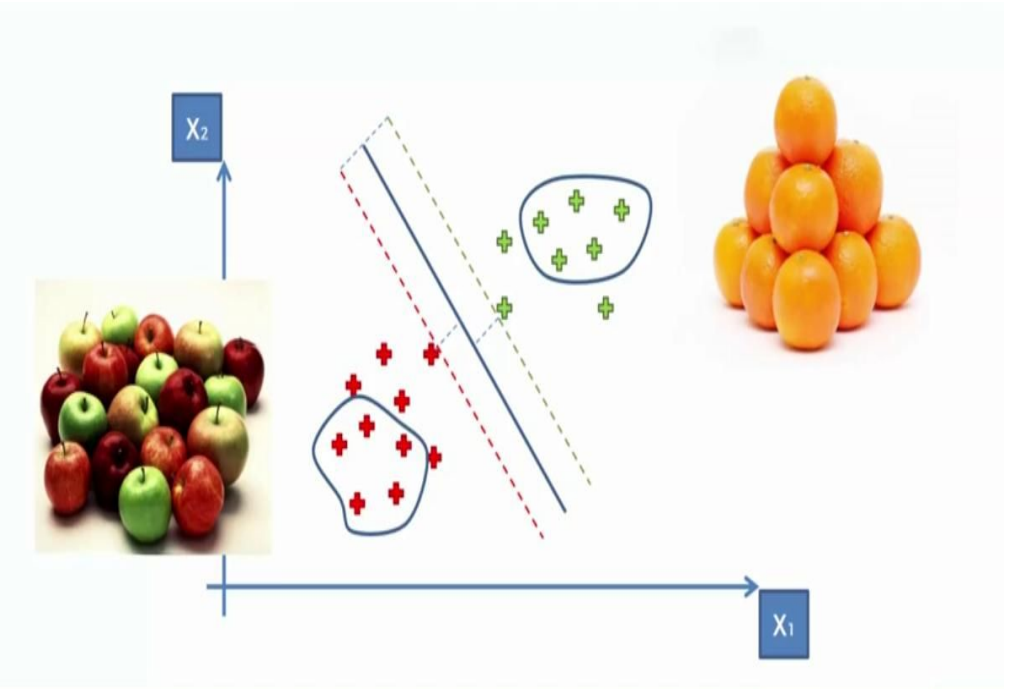


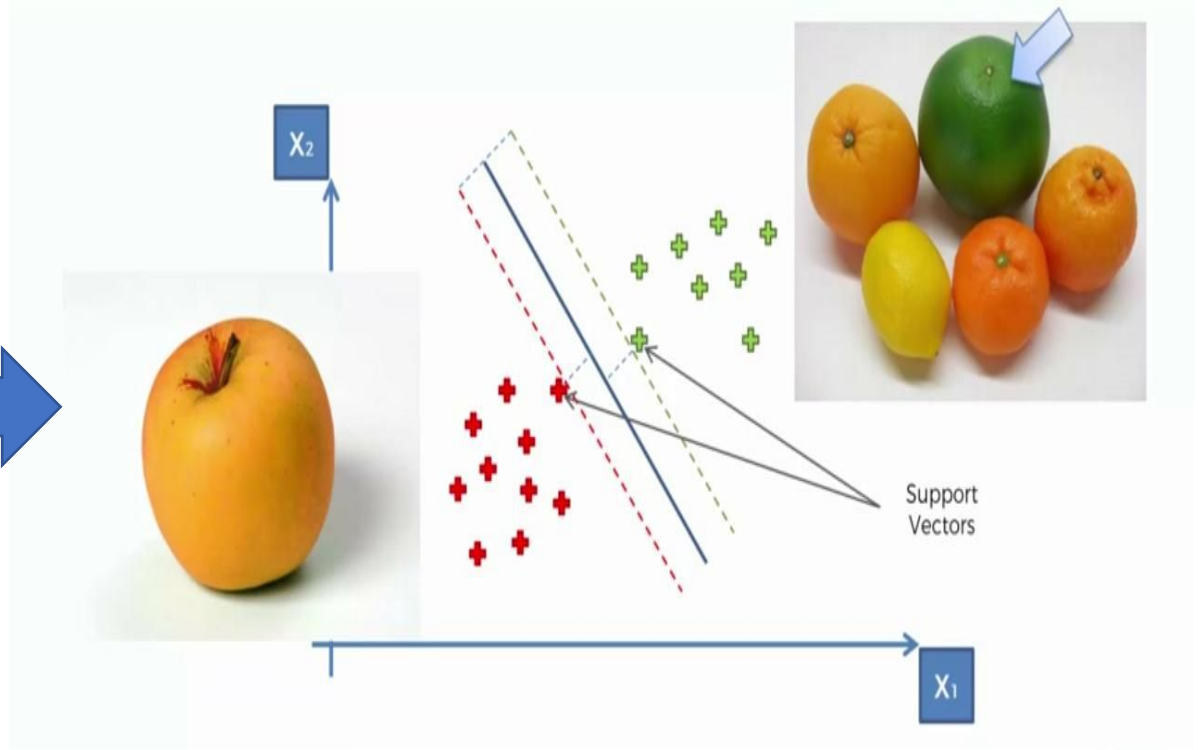


Specified Definitions in SVM



Example

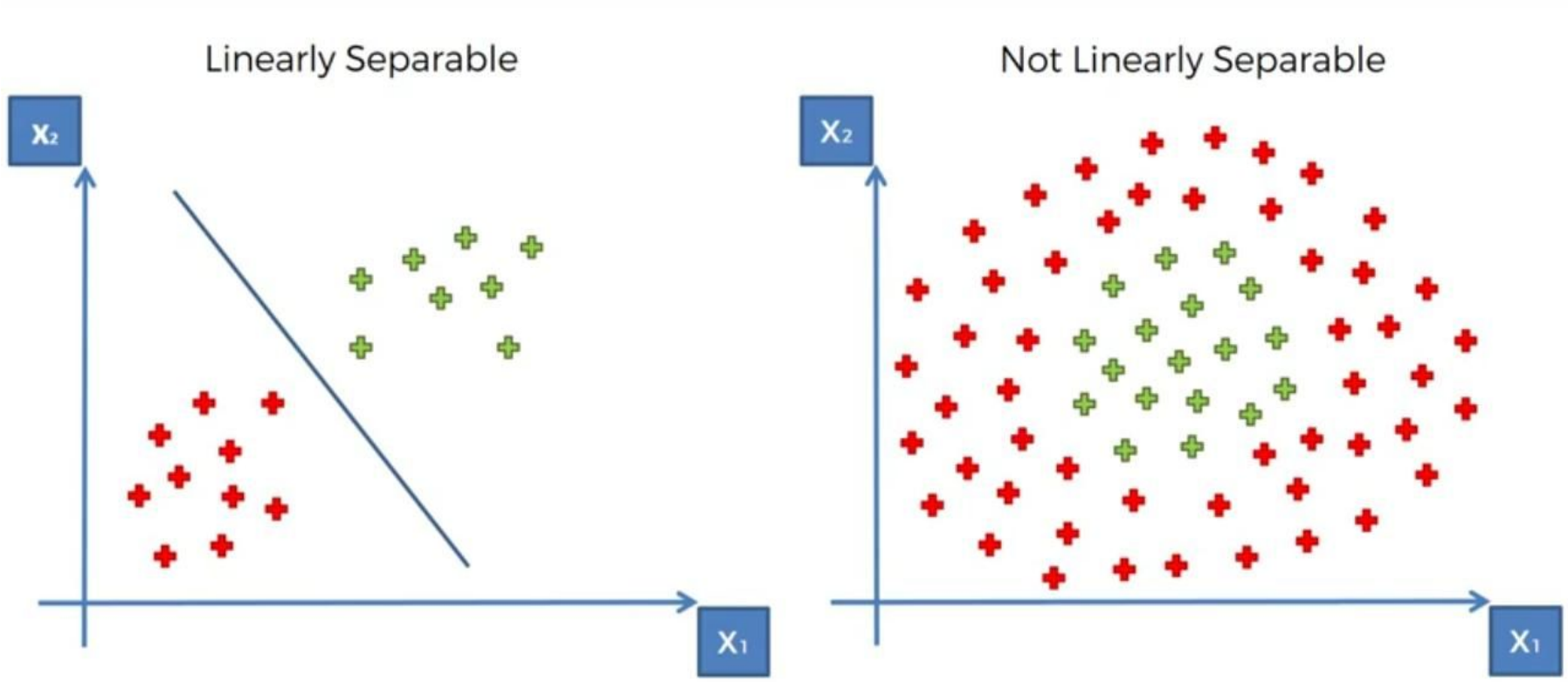




Category1:Apples
Category2:Oranges

Finding the apple closest
to an orange
And vice versa

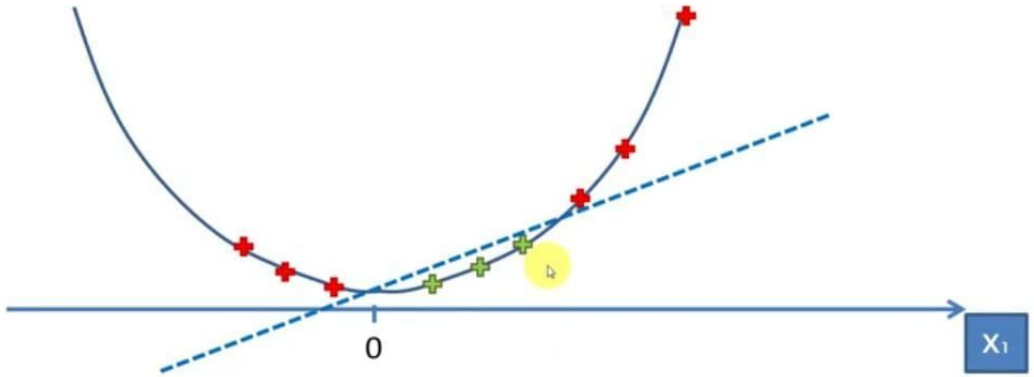
3.Kernel SVM



Mapping to Higher Dimension

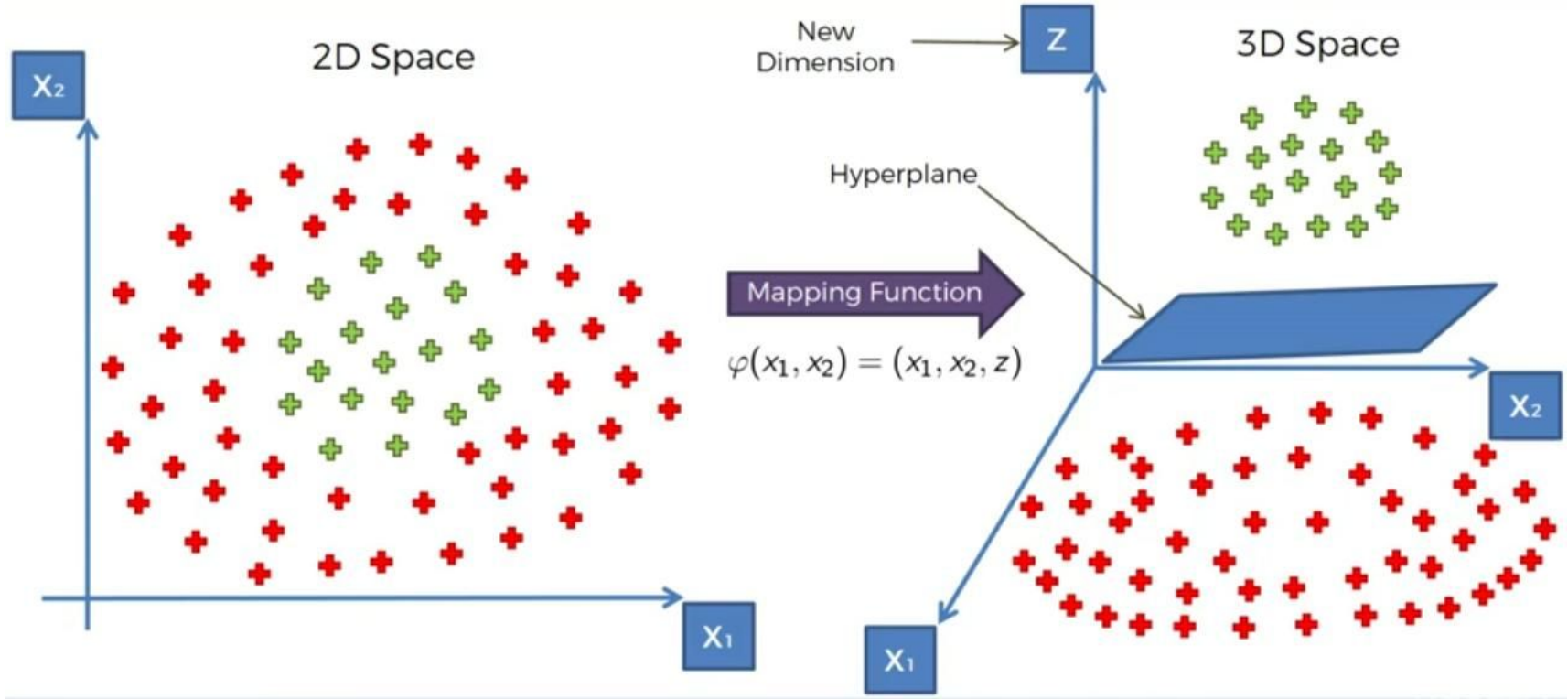
f=x-5

f=(x-5)^2

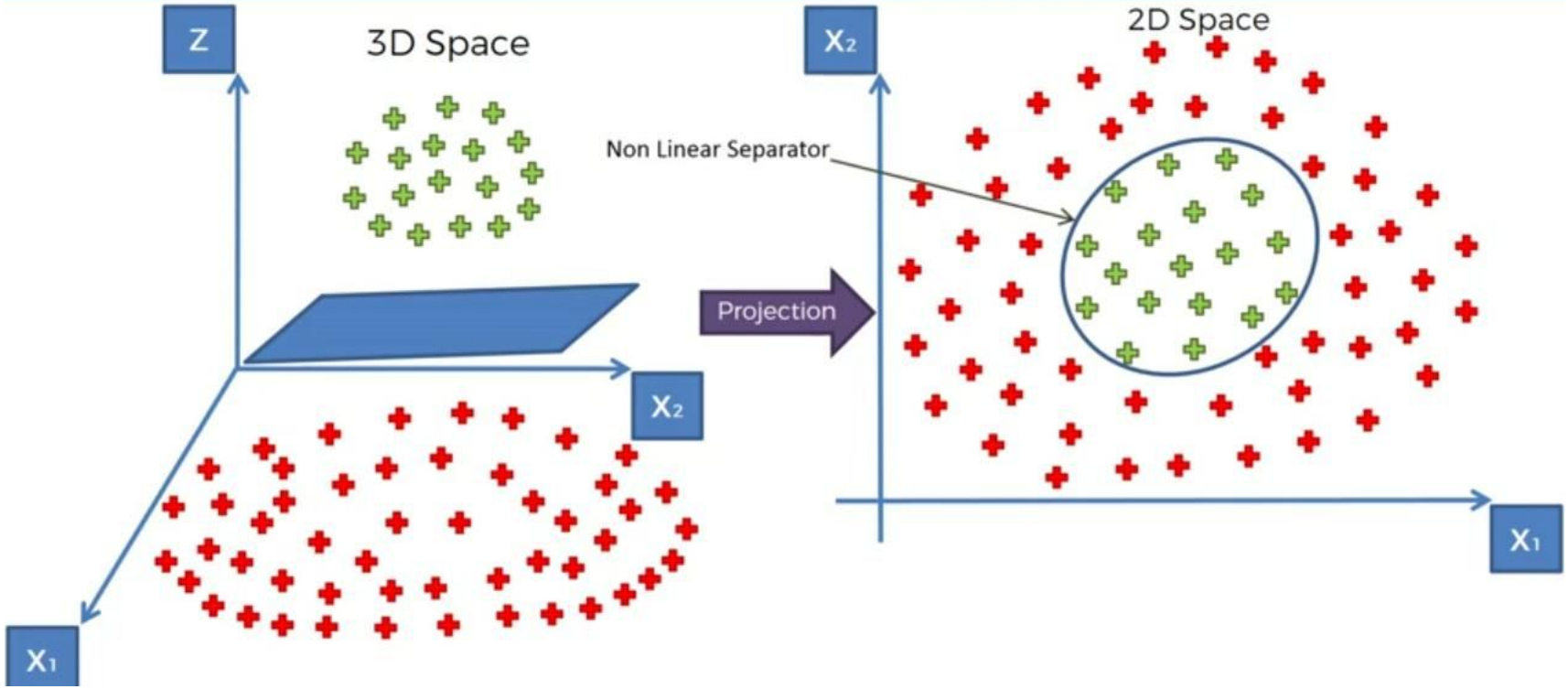


1-Dimension to 2-Dimension

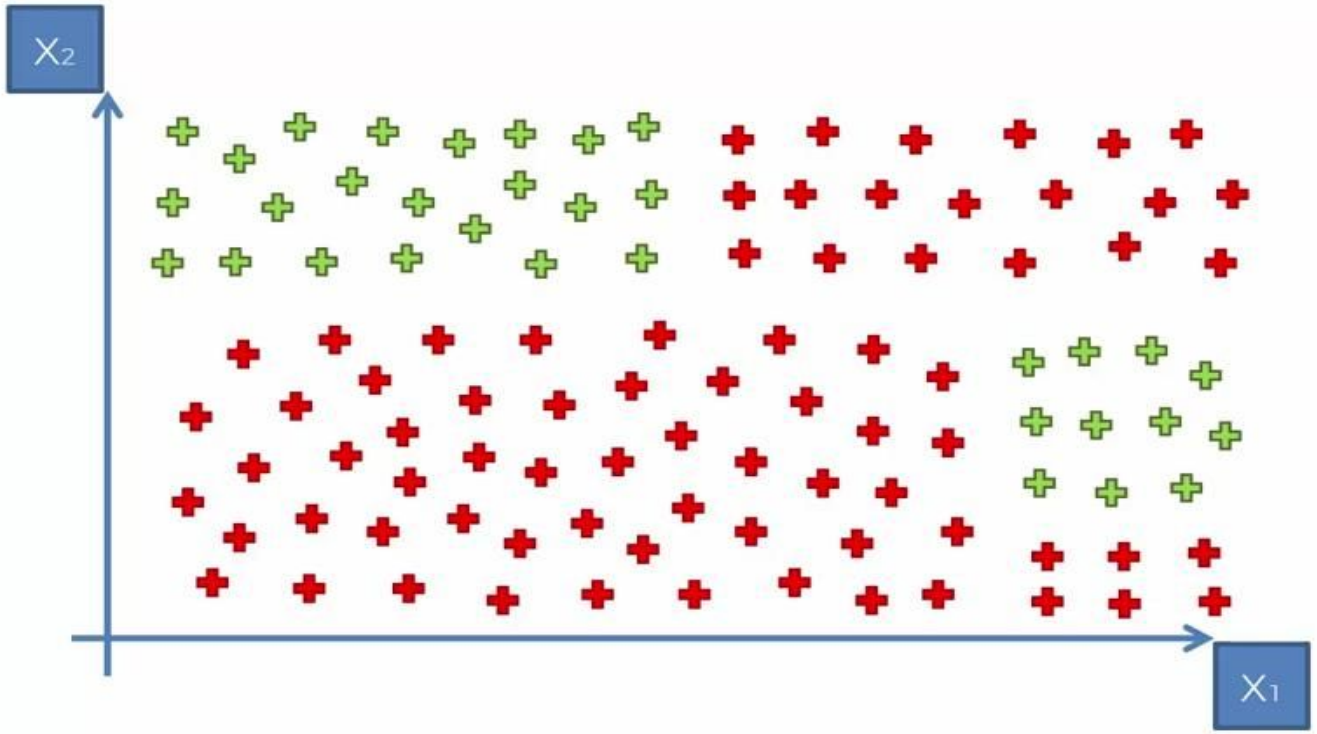
Mapping to Higher Dimension



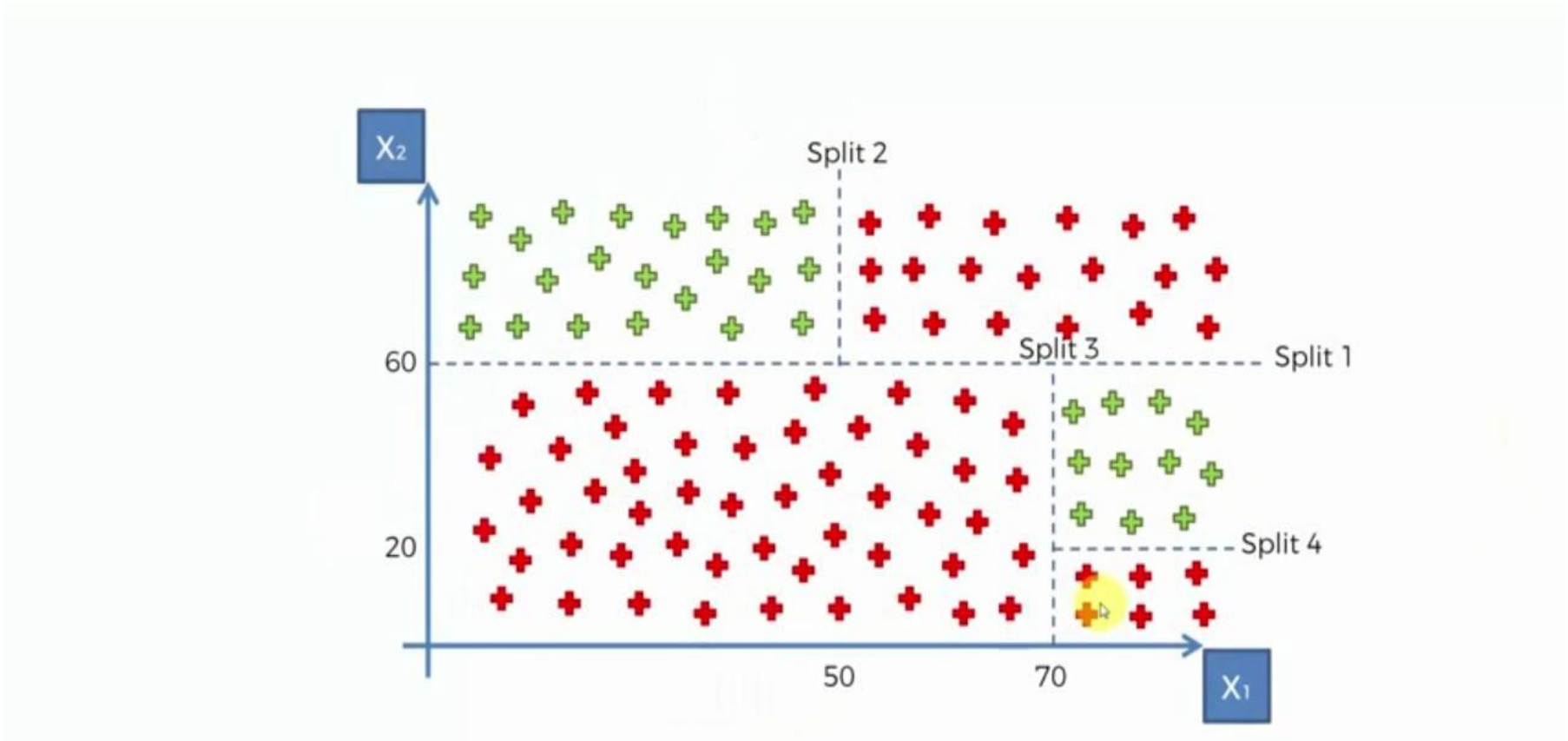
Mapping to Lower Dimension again to get the final result



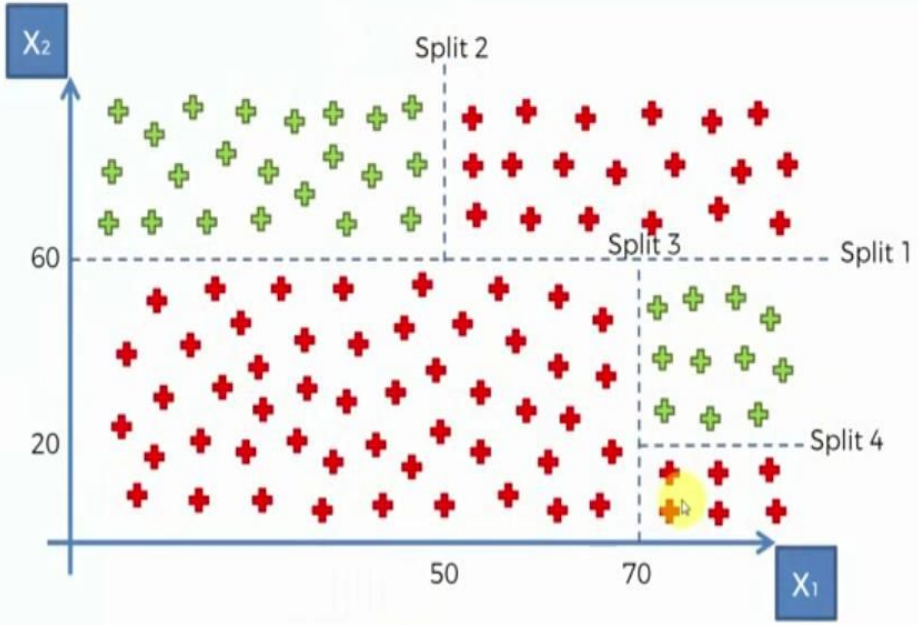
4.Decision Tree Classification

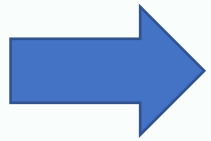


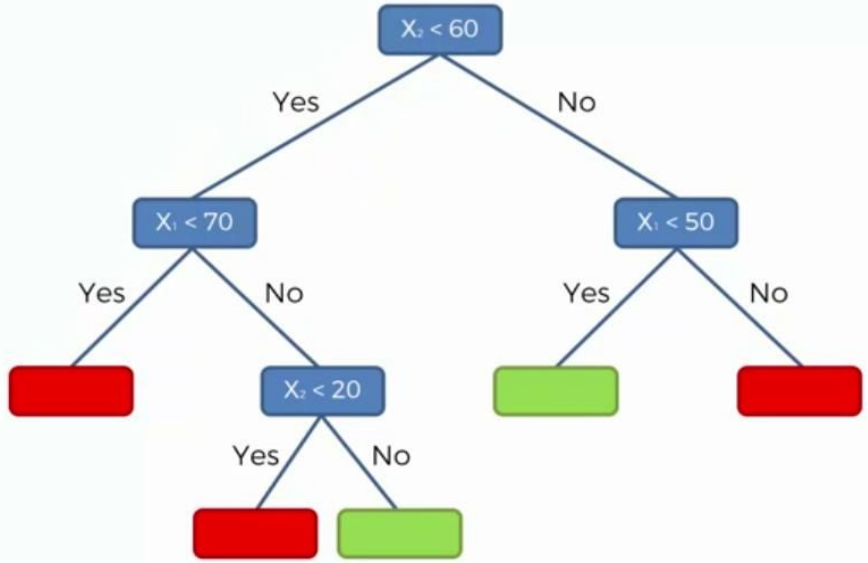
Splitting into different Classes



Converting the Splitted plot into the Decision Tree







4.Random Forest Classification

Random Forest Classifier works based on the idea of Ensemble Learning.

- Ensemble Learning:-

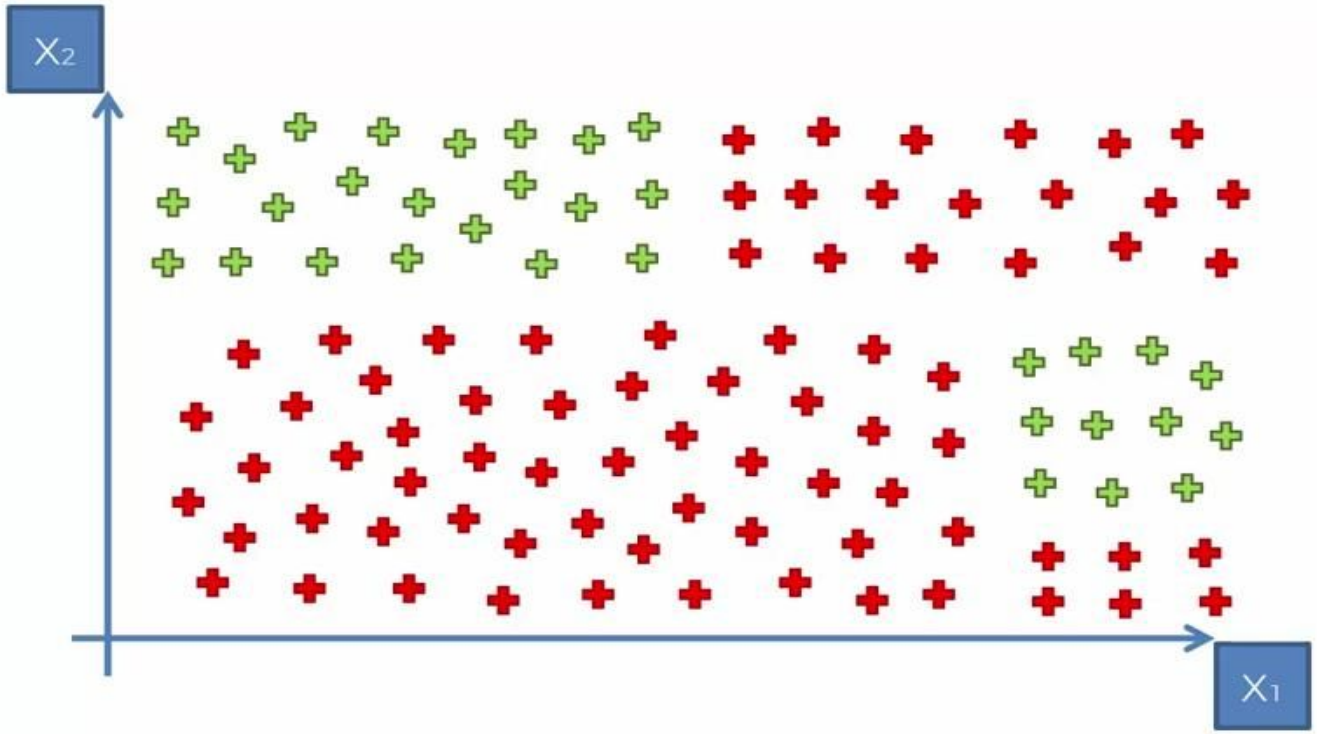
  Ensemble Learning is a Machine Learning algorithm which is created using multiple Machine Learning algorithms.

  Now these Multiple Algorithms could be same algorithms or different.

  This improves the accuracy of the entire system, as it is the combination of multiple algorithms.

  In Random Forest Classification we use Decision Tree algorithm multiple times.

It works best on the same kind of pattern, on which Decision Trees work best.



Steps for Implementing Random Forest Classifier.

STEP 1: Pick at random K data points from the Training set.

STEP 2: Build the Decision Tree associated to these K data points.

STEP 3: Choose the number Ntree of trees you want to build and repeat STEPS 1 & 2

STEP 4: For a new data point, make each one of your Ntree trees predict the category to which the data points belongs, and assign the new data point to the category that wins the majority vote.