INTERNSHIP REPORT
AT INTERNSHALA TRAININGS,

Attended by

C. Naga Bhargava Reddy (18121096)

in partial fulfilment for the award of the degree of

Bachelor of Technology

In

ELECTRONICS AND COMMUNICATION

from 15.06.2020 to 27.07.2020

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

SCHOOL OF ELECTRICAL SCIENCES

HINDUSTAN INSTITUTE OF TECHNOLOGY AND SCIENCE

PADUR 603 103

HINDUSTAN
INSTITUTE OF TECHNOLOGY & SCIENCE
(DEEMED TO BE UNIVERSITY)
CHENNAI

BONAFIDE CERTIFICATE

This is to certify that the “Internship report” submitted by C. Naga Bhargava Reddy is the work done by him and submitted during 2020–2021 academic year, in partial fulfilment of the requirements for the award of the degree of Bachelor of Technology in ELECTRONICS AND COMMUNICATION, at INTERNSHALA TRAININGS.

HEAD OF THE DEPARTMENT
Dr. A.L.Vallikannu

INTERNSHIP COORDINATOR
Ms.K.Thenkumari

CERTIFICATION

INTERNSHALA TRAININGS

Certificate of Training

NAGA BHARGAVA REDDY CHANDRA,

student of HINDUSTAN INSTITUTE OF TECHNOLOGY AND SCIENCE, has successfully completed a six weeks online training on **Machine Learning** from 15th June, 2020 to 27th July, 2020. The training consisted of Introduction to Machine Learning, Python for Machine Learning, Machine Learning Life Cycle, Data Exploration and Manipulation, Build Your First Model, Evaluation Metrics, k-NN, Selecting the Right Model, Linear Regression, Logistic Regression, Decision Trees, Feature Engineering, Basics of Ensemble Models, Random Forest and Clustering modules. In the final assessment, NAGA BHARGAVA REDDY scored 84% marks. We wish NAGA BHARGAVA REDDY all the best for the future.

Sarvesh Agrawal
Founder & CEO, Internshala

Date of certification: 2020-07-09 Certificate no. : D2B4A1AD-9D23-414E-9EF1-71719A442DAE

For certificate authentication, please visit https://trainings.internshala.com/verify_certificate

TABLE OF CONTENT

1.Introduction

The platform, which was founded in 2010, started out as a WordPress blog that aggregated internships across India and articles on education, technology and skill gap. Internshala launched its online training in 2014. As of 2018, the platform had 3.5 million students and 80,000 companies.

Services: Internship matching, online training

Founder: Sarvesh Agrawal

Industry: Education, Employment

Headquarters: Gurgaon, India

Internshala is India's no.1 internship and training platform with 40000+ paid internships in Engineering, MBA, media, law, arts, and other streams. The Top Internshala Courses

- Beginner's Trading Certification.
- Learn Digital Marketing.
- Learn How to Startup.
- Learn Business Communication Skills.
- Learn Financial Modelling and Valuation.
- Learn French.
- Learn Advanced Excel.
- Business Communication Skills Training.

In this Internship, I had learnt about introduction of ml along with the ml algorithms. Such that 42 Days of internship has 32 topics and did mini projects and final assessment in the end of internship i.e., from 35th day.

2. AGENDA

Day 1 - Introduction to python and software requirements to download.
Day 2 - Python comments, variables and syntax
Day 3&4 - Python Data Types, Numbers and Casting
Day 5&6 - Python Strings, Booleans and Operators
Day 7&8 - Lists and Tuples
Day 9 - Sets and Dictionaries
Day 10 - Machine Learning Introduction
Day 11 - Life cycle of Data Scientist
Day 12&13 - Data exploration
Day 14 - Data Manipulation
Day 15 - Brief introduction to Classifications
Day 16 - Supervised learning
Day 17&18 - Unsupervised learning with an real time example
Day 19&20 - Reinforcement learning with real time example

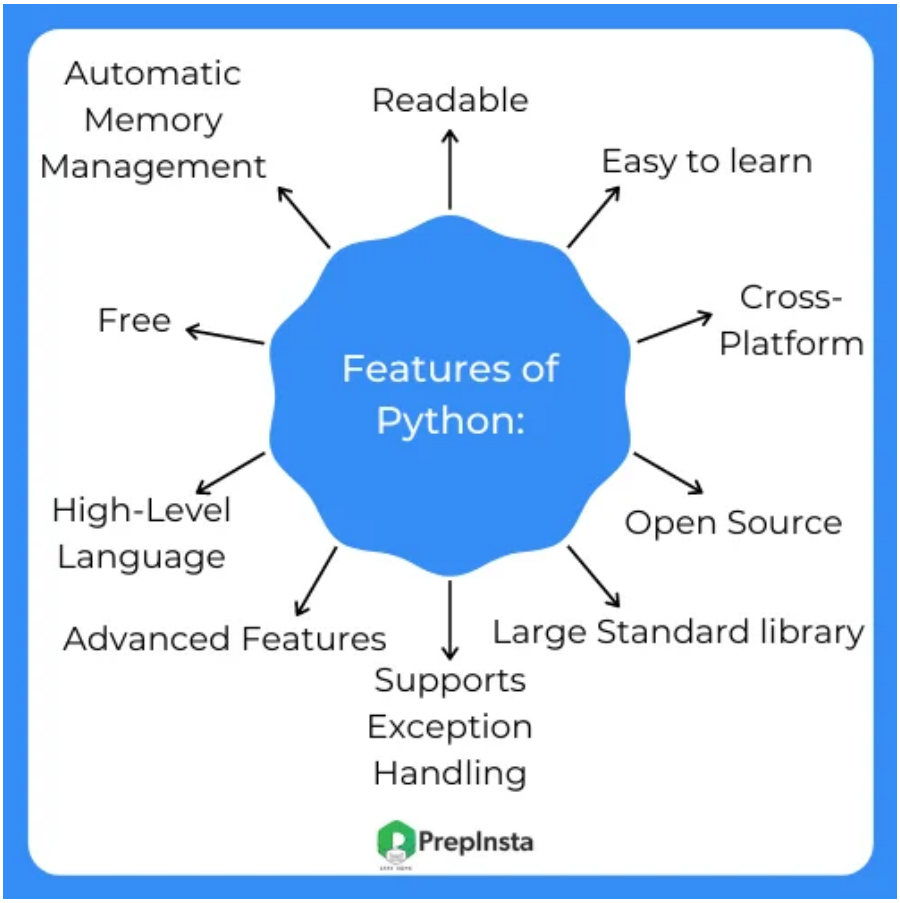
Day 21 & 22 - Linear Regression
Day 23 - Logistic Regression
Day 24 - Random Forest
Day 25 - Module Quiz
Day 26 - Ensemble models
Day 27&28 – Decision trees & knn model
Day 29 - Programming Questions from Level 0
Day 30 - Programming Questions from Level 1
Day 31 - Programming Questions from Level 2
Day 32 - Programming Questions from Level 3

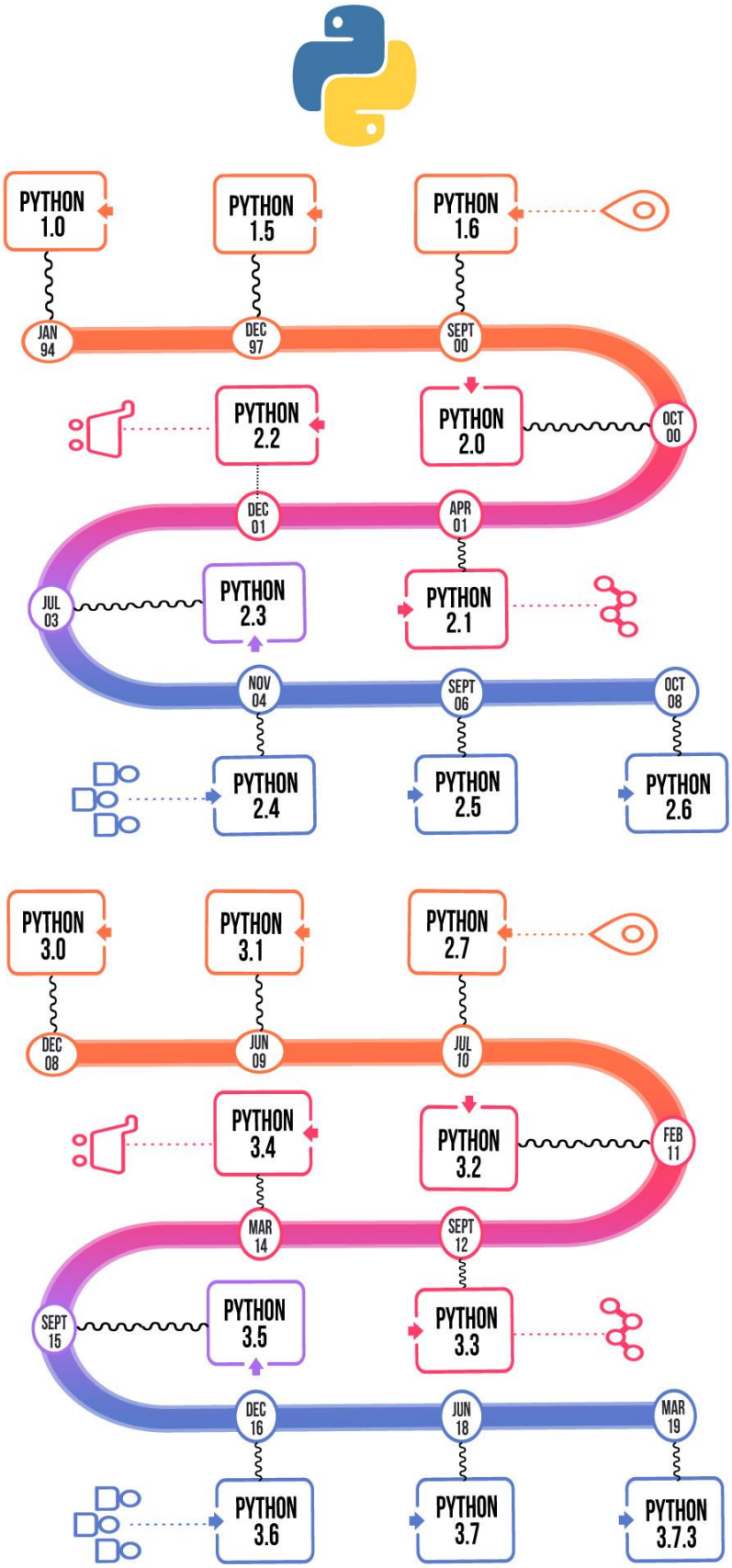
3. Topics

3.1 Python History

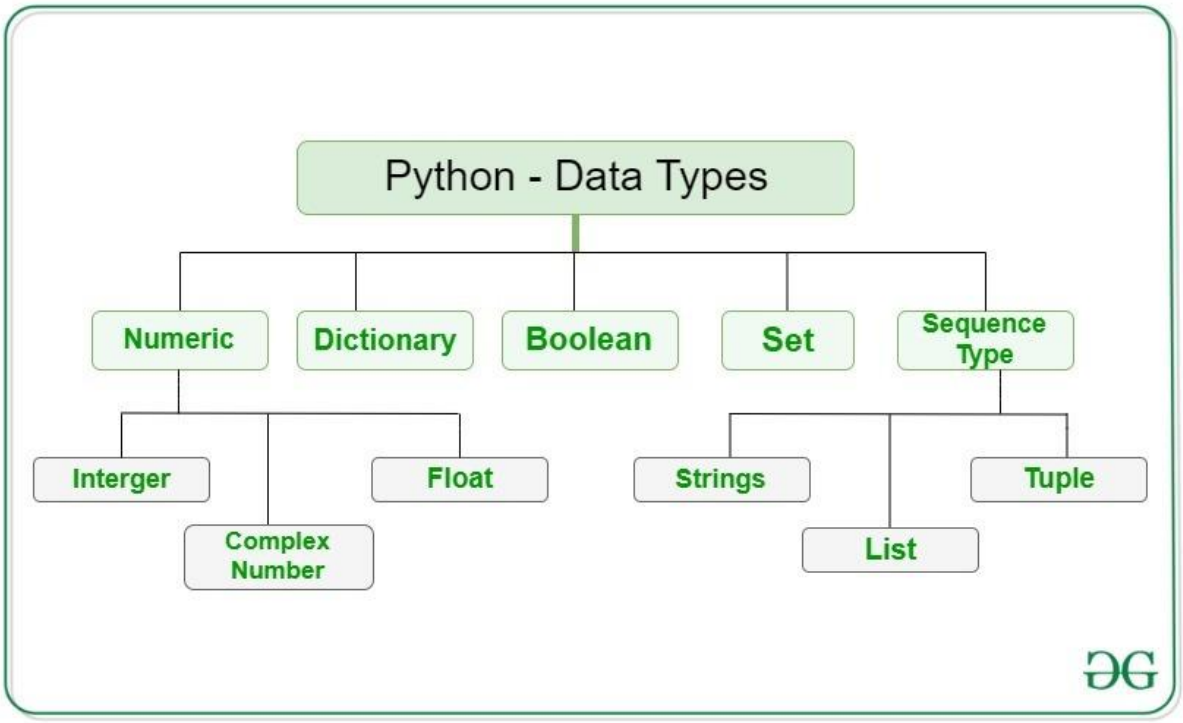
Python is a widely used general-purpose, high-level programming language. It was initially designed by Guido van Rossum in 1991 and developed by Python Software Foundation. It was mainly developed for emphasis on code readability, and its syntax allows programmers to express concepts in fewer lines of code. The two of the most used versions have to do with Python 2.x & 3.x. There is a lot of competition between the two and both of them seem to have quite a number of different fanbase.

For various purposes such as developing, scripting, generation and software testing, this language is utilised. Due to its elegance and simplicity, top technology organisations like Dropbox, Google, Quora, Mozilla, Hewlett-Packard, Qualcomm, IBM, and Cisco have implemented Python.




PYTHON 1.0
PYTHON 1.5
PYTHON 1.6
JAN 94
DEC 97
SEPT 00
PYTHON 2.2
PYTHON 2.0
OCT 00
DEC 01
APR 01
JUL 03
PYTHON 2.3
PYTHON 2.1
NOV 04
SEPT 06
OCT 08
PYTHON 2.4
PYTHON 2.5
PYTHON 2.6
PYTHON 3.0
PYTHON 3.1
PYTHON 2.7
DEC 08
JUN 09
JUL 10
PYTHON 3.4
PYTHON 3.2
FEB 11
MAR 14
SEPT 12
SEPT 15
PYTHON 3.5
PYTHON 3.3
DEC 16
JUN 18
MAR 19
PYTHON 3.6
PYTHON 3.7
PYTHON 3.7.3

3.2 Data Types in Python



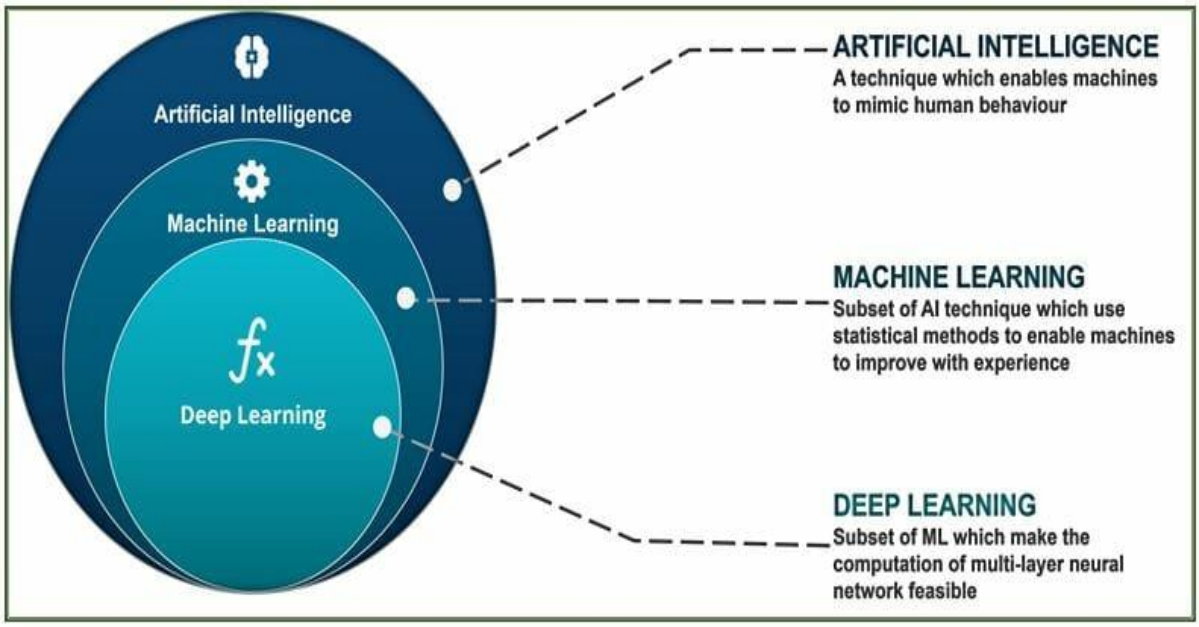
3.3 MACHINE LEARNING INTRODUCTION

Machine learning is a subfield of artificial intelligence (AI). The goal of machine learning generally is to understand the structure of data and fit that data into models that can be understood and utilized by people.

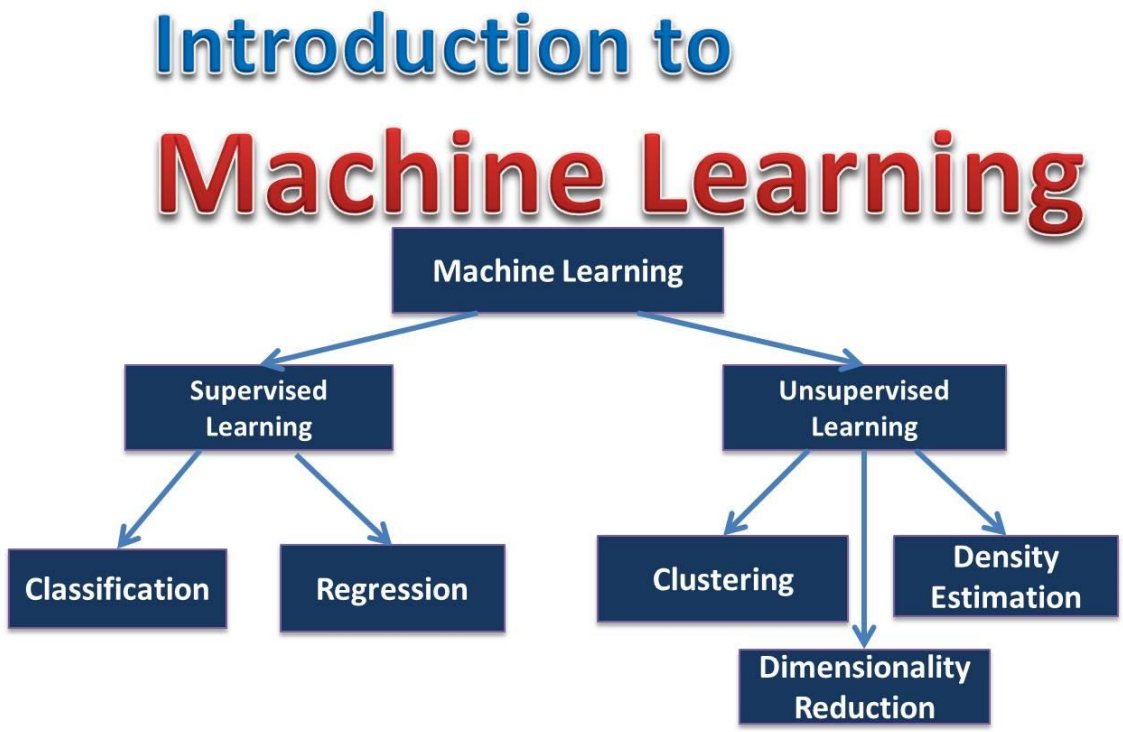
Although machine learning is a field within computer science, it differs from traditional computational approaches. In traditional computing, algorithms are sets of explicitly programmed instructions used by computers to calculate or problem solve. Machine learning algorithms instead allow for computers to train on data inputs and use statistical analysis in order to output values that fall within a specific range. Because of this, machine learning facilitates computers in building models from sample data in order to automate decision-making processes based on data inputs.

Any technology user today has benefitted from machine learning. Facial recognition technology allows social media platforms to help users tag and share photos of friends. Optical character recognition (OCR) technology converts images of text into movable type. Recommendation engines,

powered by machine learning, suggest what movies or television shows to watch next based on user preferences. Self-driving cars that rely on machine learning to navigate may soon be available to consumers.



Classification of ML:



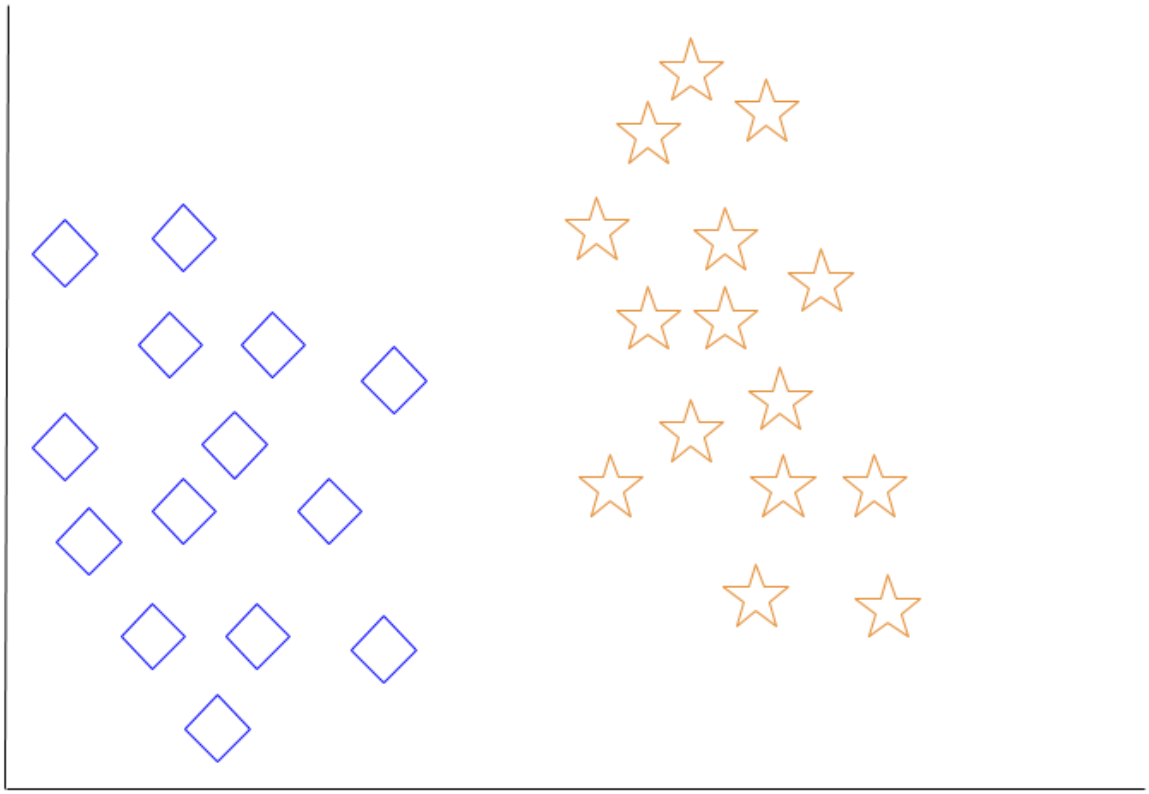
3.4 MACHINE LEARNING ALGORITHMS

3.4.1 KNN Algorithm:

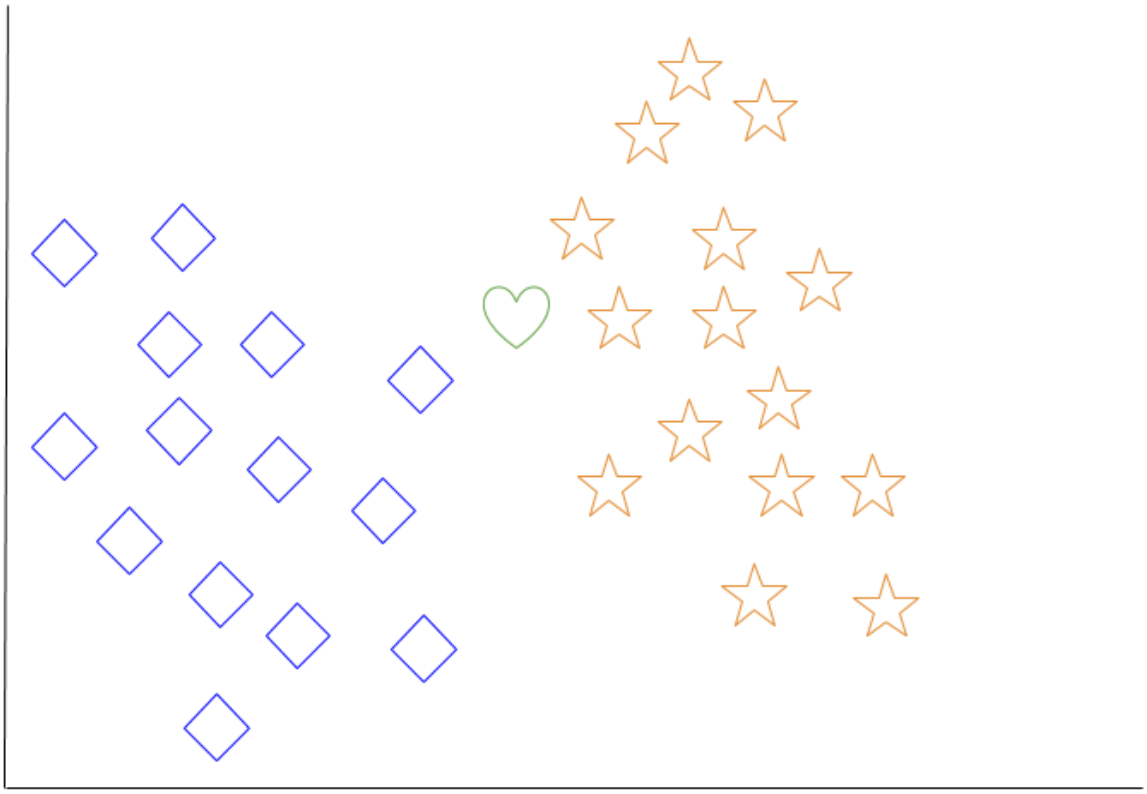
The k-nearest neighbour algorithm is a pattern recognition model that can be used for classification as well as regression. Often abbreviated as k-NN, the **k** in k-nearest neighbour is a positive integer, which is typically small. In either classification or regression, the input will consist of the k closest training examples within a space.

We will focus on k-NN classification. In this method, the output is class membership. This will assign a new object to the class most common among its k nearest neighbours. In the case of k = 1, the object is assigned to the class of the single nearest neighbour.

Let's look at an example of k-nearest neighbour. In the diagram below, there are blue diamond objects and orange star objects. These belong to two separate classes: the diamond class and the star class.

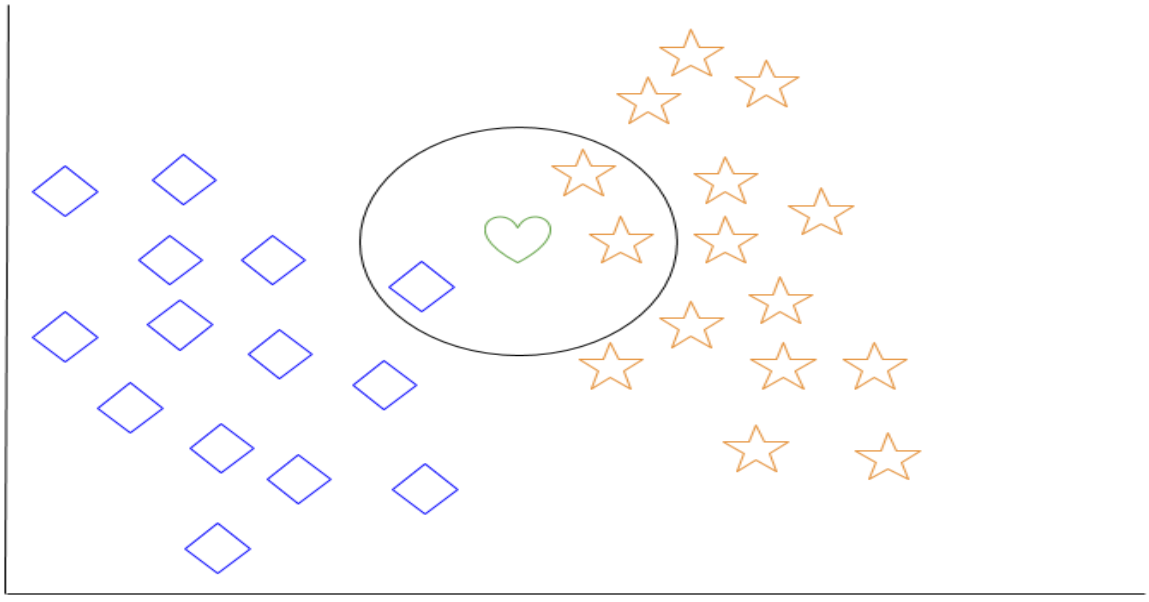


When a new object is added to the space in this case a green heart, we will want the machine learning algorithm to classify the heart to a certain class.



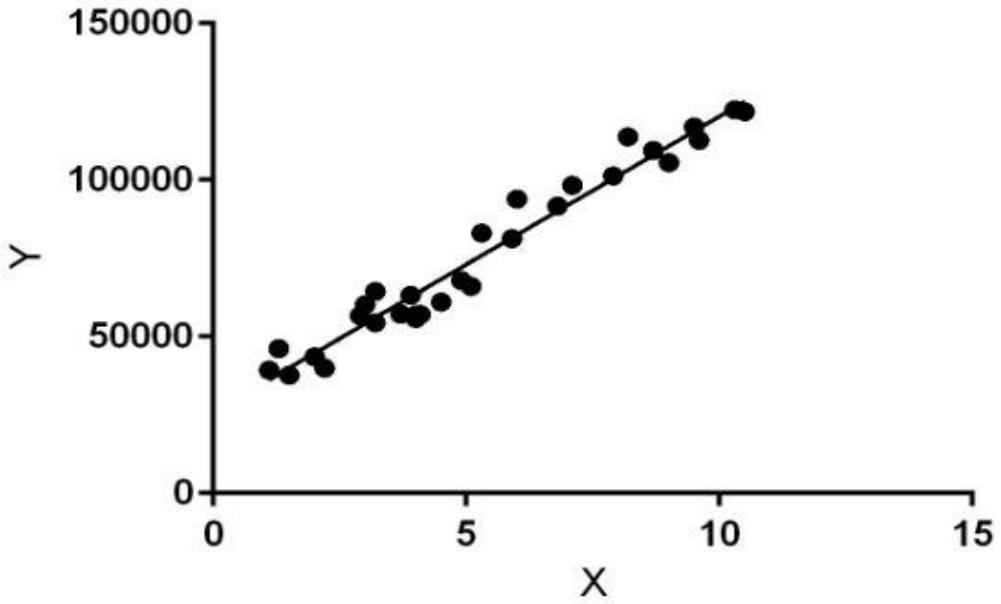
When we choose k = 3, the algorithm will find the three nearest neighbours of the green heart in order to classify it to either the diamond class or the star class.

In our diagram, the three nearest neighbours of the green heart are one diamond and two stars. Therefore, the algorithm will classify the heart with the star class.



3.4.2 LINEAR REGRESSION

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.



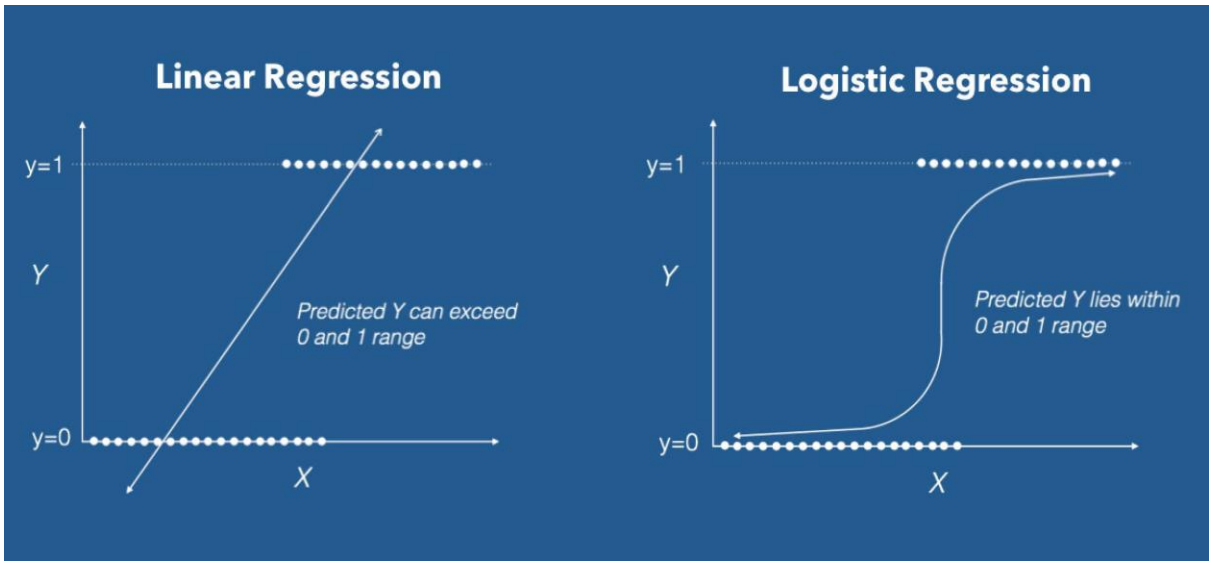
Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model.

$$y = \theta_1 + \theta_2.x$$

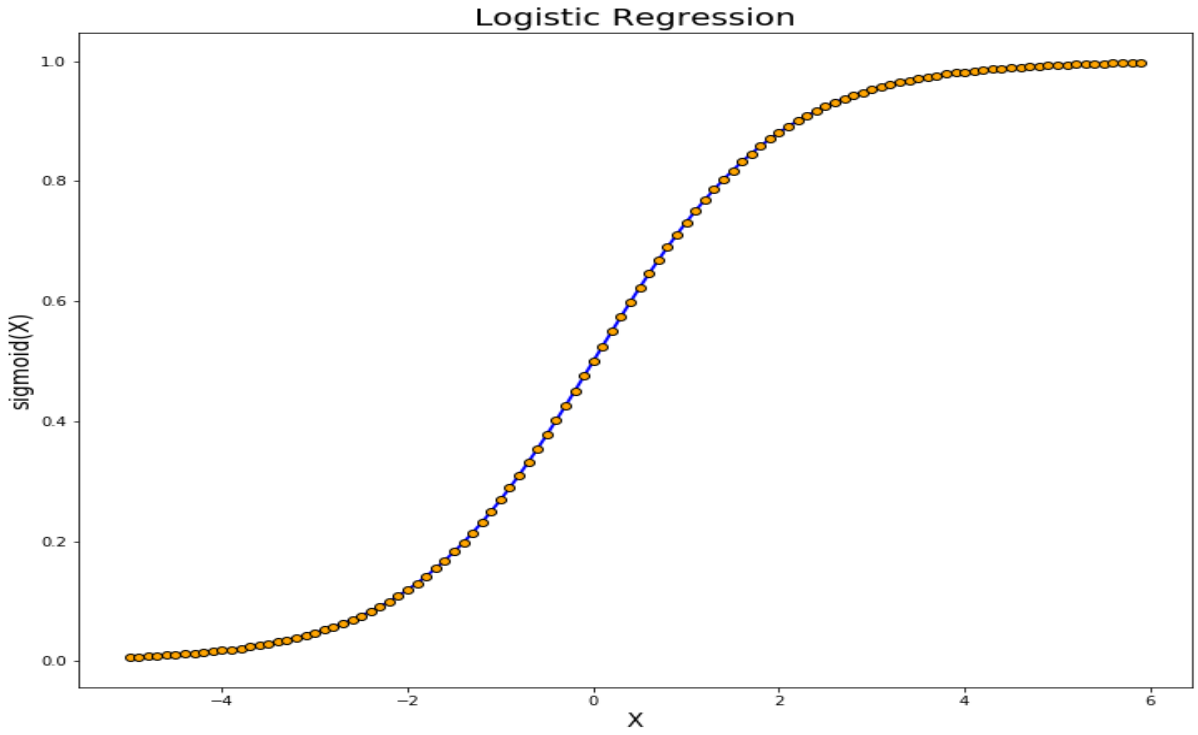
3.4.2 LOGISTIC REGRESSION

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression).

It is used in statistical software to understand the relationship between the dependent variable and one or more independent variables by estimating probabilities using a logistic regression equation. This type of analysis can help you predict the likelihood of an event happening or a choice being made.



In statistics, the logistic model is used to model the probability of a certain class or event taking place, such as the probability of a team winning, of a patient being healthy, etc. This can be extended to model several classes of events such as determining whether an image contains a cat, dog, lion, etc.



3.4.3 DECISION TREES

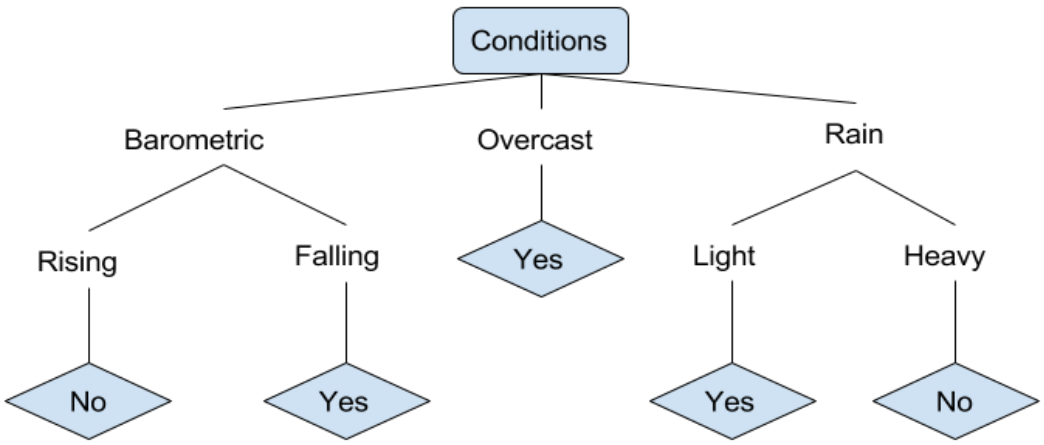
For general use, decision trees are employed to visually represent decisions and show or inform decision making. When working with machine learning and data mining, decision trees are used as a predictive model. These models map observations about data to conclusions about the data's target value.

The goal of decision tree learning is to create a model that will predict the value of a target based on input variables.

In the predictive model, the data's attributes that are determined through observation are represented by the branches, while the conclusions about the data's target value are represented in the leaves.

When "learning" a tree, the source data is divided into subsets based on an attribute value test, which is repeated on each of the derived subsets recursively. Once the subset at a node has the equivalent value as its target value has, the recursion process will be complete.

Let's look at an example of various conditions that can determine whether or not someone should go fishing. This includes weather conditions as well as barometric pressure conditions.

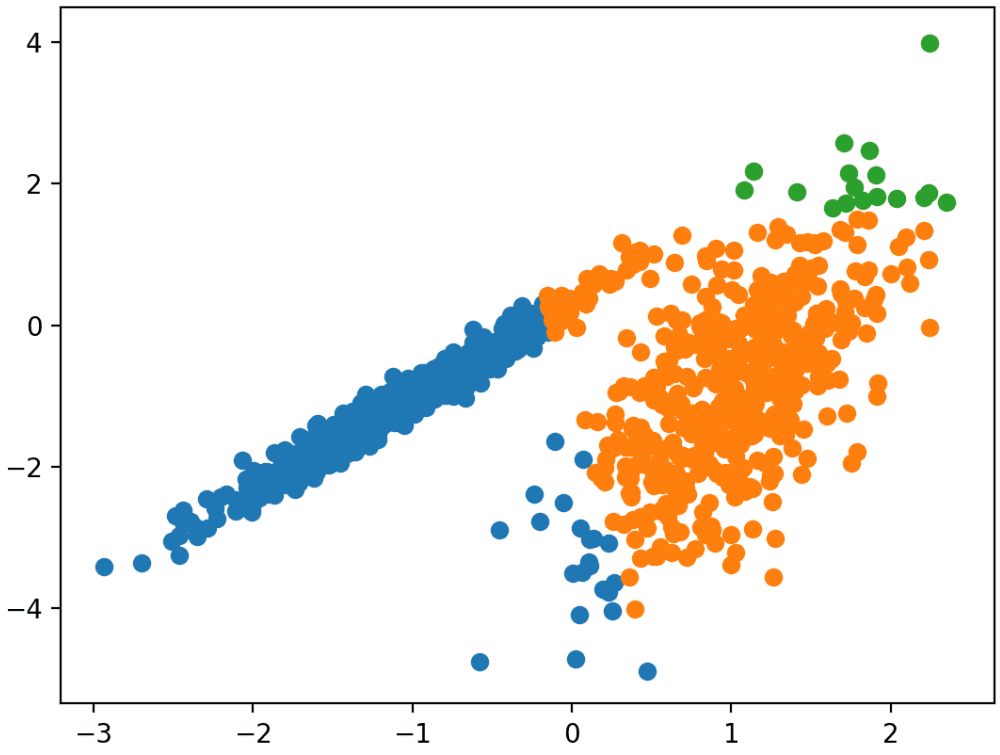


In the simplified decision tree above, an example is classified by sorting it through the tree to the appropriate leaf node. This then returns the classification associated with the particular leaf, which in this case is either a Yes or a No. The tree classifies a day's conditions based on whether or not it is suitable for going fishing.

A true classification tree data set would have a lot more features than what is outlined above, but relationships should be straightforward to determine. When working with decision tree learning, several determinations need to be made, including what features to choose, what conditions to use for splitting, and understanding when the decision tree has reached a clear ending.

3.4.4 CLUSTERING MODULES

Cluster analysis, or clustering, is an unsupervised machine learning task. It involves automatically discovering natural grouping in data. Unlike supervised learning (like predictive modelling), clustering algorithms only interpret the input data and find natural groups or clusters in feature space.



4. PROJECT

CREDIT CARD FRAUD DETECTION

4.1 Project Explanation

The problem statement chosen for this project is to predict fraudulent credit card transactions with the help of machine learning models.

In this project, we will analyse customer-level data which has been collected and analysed during a research collaboration of Worldline and the Machine Learning Group.

The dataset is taken from the Kaggle Website and it has a total of 2,84,807 transactions, out of which 492 are fraudulent. Since the dataset is highly imbalanced, so it needs to be handled before model building.

Understanding and Defining Fraud

Credit card fraud is any dishonest act and behaviour to obtain information without the proper authorization from the account holder for financial gain. Among different ways of frauds, Skimming is the most common one, which is the way of duplicating of information located on the magnetic strip of the card. Apart from this, the other ways are:

- Manipulation/alteration of genuine cards
- Creation of counterfeit cards
- Stolen/lost credit cards
- Fraudulent telemarketing

Data Dictionary

The dataset can be download using this link

The data set includes credit card transactions made by European cardholders over a period of two days in September 2013. Out of a total of 2,84,807 transactions, 492 were fraudulent. This data set is highly unbalanced, with the positive class (frauds) accounting for 0.172% of the total transactions. The data set has also been modified with Principal Component Analysis (PCA) to maintain confidentiality. Apart from ‘time’ and ‘amount’, all the other features (V1, V2, V3, up to V28) are the principal components

obtained using PCA. The feature 'time' contains the seconds elapsed between the first transaction in the data set and the subsequent transactions. The feature 'amount' is the transaction amount. The feature 'class' represents class labelling, and it takes the value 1 in cases of fraud and 0 in others.

4.2 Solution Approach

1. Data understanding and exploring

2. Data cleaning

- Handling missing values
- Outliers' treatment

3. Exploratory data analysis

- Univariate analysis
- Bivariate analysis

4. Prepare the data for modelling

Check the skewness of the data and mitigate it for fair analysis

- Handling data imbalance as we see only 0.172% records are the fraud transactions 5. Split the data into train and test set
- Scale the data (normalization)

6. Model building

- Train the model with various algorithm such as Logistic regression, SVM, Decision Tree, Random Forest, XG Boost etc.
- Tune the hyperparameters with Grid Search Cross Validation and find the optimal values of the hyperparameters

7. Model evaluation

- As we see that the data is heavily imbalanced, Accuracy may not be the correct measure for this particular case
- We have to look for a balance between Precision and Recall over Accuracy. We also have to find out the good ROC score with high TPR and low FPR in order to get the lower number of misclassifications.