IIITD

MACHINE LEARNING
PROJECT

5 Takeaways

1. ML module projects are designed to have a detailed hands on to integrate theoretical knowledge with actual practical implementations.

2. ML module projects are designed to enable you as a learner to workon real time industry scenarios, problems and data sets.

3. ML module projects are designed to enable you simulating the designed solution using ML techniques onto python technology platform.

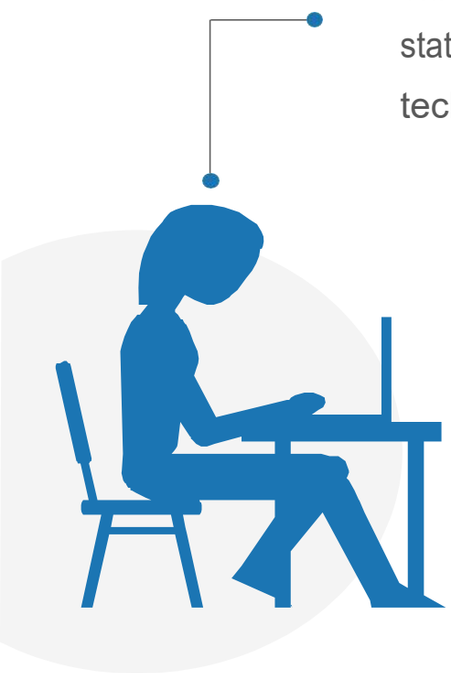
4. ML module projects are designed to be scored using a predefined rubric based system.

5. ML module projects are designed to enhance your learning above and beyond. Hence, it might require you to experiment, research, self learn and implement.

IIITD

MACHINE LEARNING PROJECT

UNSUPERVISED LEARNING



This project consists of industry based dataset and problem statement which can be solved using unsupervised learning techniques.

TOTAL SCORE 20

PROJECT BASED

TOTAL SCORE 20

DOMAIN: Mobile

BUSINESS CONTEXT:

- A key challenge for Mobile App businesses is to analyze the trend in the market to increase their sales/usage.
- We have access to the user's demographic characteristics, geo-location, and mobile device properties. This grouping can be done by applying different criteria like user's data, their age group, phone brand compatibility and so on.
- The machine learning clustering algorithms can provide an analytical method to cluster user segments with similar interests/habits.This will help App/mobile providers better understand and interact with their subscribers.

DATA DESCRIPTION:

- **events.csv** - Event data has an event id, location detail (lat/long), and timestamp, when the user is using an app on his device
- **gender_age.csv** - Details of users age & gender
- **phone_device.csv** - Device ids, brand, and models name. Here the brands names are in Chinese, you can convert it to english using google for better understanding but we will not do it here.

  Few important conversions are as listed below:

  - 三星 samsung
  - 天语 Ktouch
  - 海信 hisense
  - 联想 lenovo
  - 爱派尔 ipair
  - 一加 oneplus
  - 诺基亚 nokia
  - 华硕 asus
  - 夏新 panosonic
  - 锤子 hammer

PROJECT OBJECTIVE:

- We will be clustering the users into groups by selected features that significantly distinguish different brands from each other and understand which factors are responsible for making the clusters

STEPS TO THE PROJECT: [TOTAL SCORE: 20 POINTS]

1. Import the required libraries and load the data: [Score: 1 point]
   - Load the required libraries and read the dataset.
   - Check the first few samples, shape, info of the data and try to familiarize yourself with different features

2. Data cleansing and Exploratory data analysis: [Score: 8 point]
   - Check if there are any duplicate records in the dataset? If any drop them.
   - Merge the data into a single data-frame.
   - Check for missing values in each column of the dataset? If it exists, handle them accordingly.
   - Check the statistical summary for the numerical and categorical columns and write your findings.
   - Perform the data visualization on the dataset to gain some basic insights about the data.
   - Encode the categorical variables in the dataset.
   - Drop irrelevant columns like 'timestamp','event_id','device_id' etc from the dataset.

3. Data Preparation for model building: [Score: 1 point]
   - Standardize the data, so that the values are within a particular range.

4. Principal Component Analysis and Clustering: [Score: 8 point]
   - Apply PCA on the above dataset and determine the number of PCA components to be used so that 90-95% of the variance in data is explained by the same.
   - Apply K-means clustering and segment the data. (You may use original data or PCA transformed data)
     a. Find the optimal K Value using elbow plot for K means clustering.
     b. Build a K means clustering model using the obtained optimal K value from the elbow plot.
     c. Compute silhouette score for evaluating the quality of the K means clustering technique.
   - Apply Agglomerative clustering and segment the data. (You may use original data or PCA transformed data)
     a. Find the optimal K Value using dendrogram for Agglomerative clustering.
     b. Build a Agglomerative clustering model using the obtained optimal K value observed from dendrogram.
     c. Compute silhouette score for evaluating the quality of the Agglomerative clustering technique.
     (Hint: Take a sample of the dataset for agglomerative clustering to reduce the computational time)

5. Conclusion : [Score: 2 point]
   - Perform cluster analysis by doing bi-variate analysis between cluster label and different features and write your conclusion on the results.

" Put yourself in the shoes of an actual "

DATA SCIENTIST

THAT's YOU

Assume that you are working at the company which has received the above problem statement from internal/external client. Finding the best solution for the problem statement will enhance the business/ operations for your organization/project. You are responsible for the complete delivery. Put your best analytical thinking hat to squeeze the raw data into relevant insights and later into an AIML working model.

PLEASE NOTE

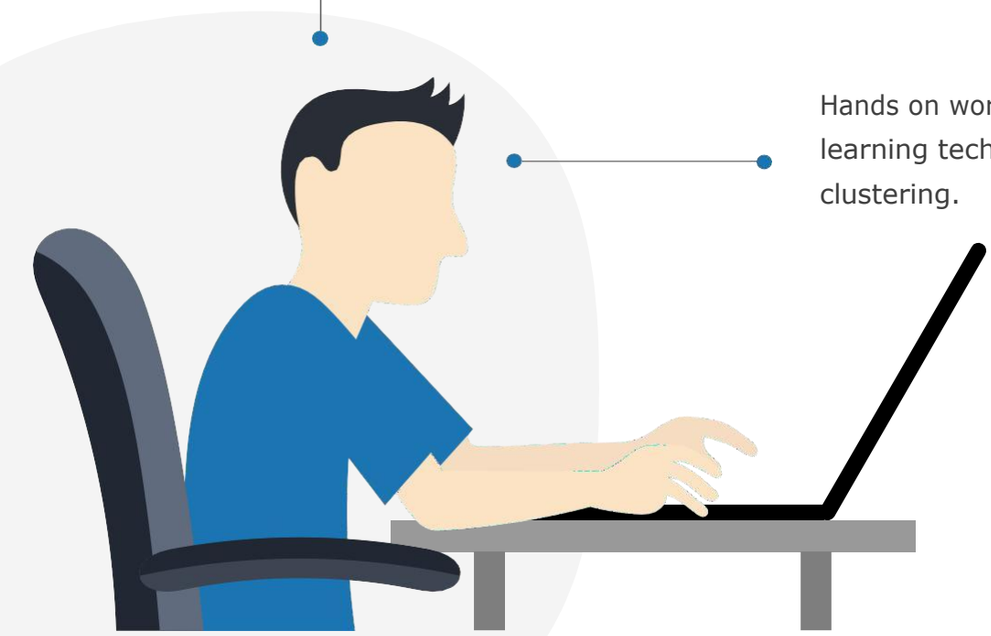
Designing a data driven decision product typically traces the following process:

Data and insights:

Warehouse the relevant data. Clean and validate the data as per the the functional requirements of the problem statement. Capture and validate all possible insights from the data as per the the functional requirements of the problem statement. Please remember there will be numerous ways to achieve this. Sticking to relevance is of utmost importance. Pre-process the data which can be used for relevant Machine learning model.

LEARNING OUTCOME

Hands on experience on data design, data cleansing, data warehousing and data manipulation required to drive an ML model.



Hands on working knowledge on Unsupervised learning techniques such as PCA and clustering.

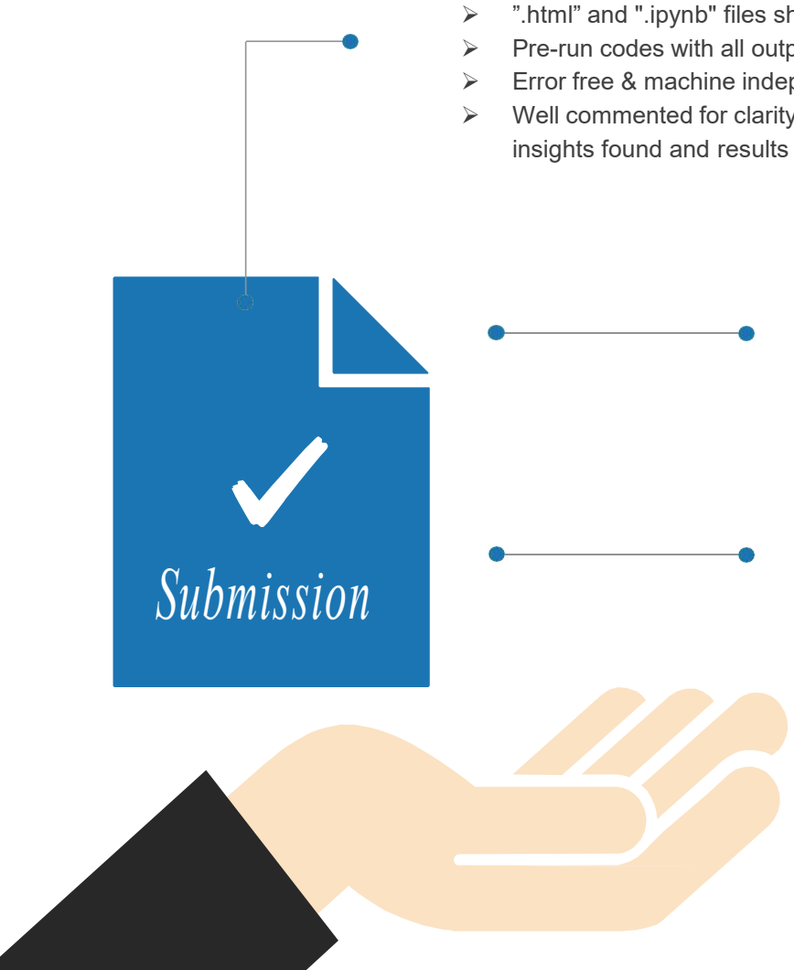
Deep dive onto applying PCA on the given data and designing a clustering model.

Hands-on experience on evaluating the goodness of the clustering models based on the appropriate performance metrics.

IMPORTANT POINTERS

Project should be submitted as a single ".html" and ".ipynb" file. Follow the below best practices where your submission should be:

- ".html" and ".ipynb" files should be an exact match.
- Pre-run codes with all outputs intact.
- Error free & machine independent i.e. run on any machine without adding any extra code.
- Well commented for clarity on code designed, assumptions made, approach taken, insights found and results obtained.



Project should be submitted on or before the deadline given by the program office.

Project submission should be an original work from you as a learner. If any percentage of plagiarism found in the submission, the project will not be evaluated and no score will be given.

HAPPY LEARNING