A picture containing font, text, logo, graphics

Description automatically generated

**Summer Training Report**

**on**

***Credit Card Fraud Detection using ML***

Submitted by

Aman Yadav (21CSU502)

Shivam (21CSU325)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING,

SCHOOL OF ENGINEERING AND TECHNOLOGY

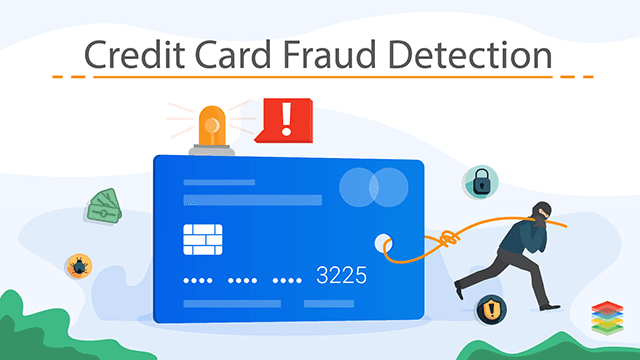
THE NORTHCAP UNIVERSITY

GURUGRAM-122017

Internship Period: 12th June 2023 TO 21st July 2023 (6 Weeks)

***Credit Card Fraud Detection***

***Using Machine Learning***



**Table of Contents**

|  |  |
| --- | --- |
| 1. | List of Figures |
| 2. | **Abstract** |
| 3. | **Introduction** |
| 4. | **Problem Statement** |
| 5. | **Social Relevance of our Project** |
| 6. | **Training Description**   * **Objective** * **Data Extraction** * **Features of Dataset** * **Data Preprocessing** * **Fixing Dataset using SMOTE** * **Data Visualization** * **Training our Dataset** * **ML Algorithms Used** * **Result** |
| 7. | **Analysis** |
| 8. | **Conclusion** |
| 9. | **Bibliography** |
| 10. | **Daily Analysis** |

**List of Figures**

* Kaggle Image (Taken From Google)
* Duplicate Values Check
* Comparing the Size of Fraudulent and Legitimate Transactions
* List of Dataset
* Over Sampling of Fraudulent Transactions Using SMOTE
* Scatterplot (Before SMOTE)
* Heatmap (Before SMOTE)
* Displaying Fraud and Legit Transactions (After SMOTE)
* Heatmap (After SMOTE)
* Taking a sample
* Decision Tree Implementation
* KNN Implementation
* Random Forest Classifier Implementation
* Confusion Matrix of Random Forest

***Abstract***

Our Summer Internship & Training Program was carried out in the The NorthCap University itself.

Over the course of our summer internship training program focused on Machine Learning, I gained invaluable insights into the multifaceted world of data science and its applications. This abstract encapsulates the key elements and takeaways from this enriching experience, both in terms of technical and non-technical aspects.

**Technical Takeaways:**

Algorithm Proficiency: Through hands-on workshops and guided projects, I honed my skills in implementing various machine learning algorithms, including Regression, Classification, Clustering, and Neural Networks. I gained an understanding of their strengths, weaknesses, and best use cases.

Data Preprocessing: I learned the critical importance of data preprocessing, from handling missing values and outliers to feature scaling and transformation. This preprocessing played a pivotal role in enhancing model accuracy and robustness.

Model Evaluation: I gained expertise in assessing model performance using metrics such as accuracy, precision, recall, and F1-score. This understanding enabled me to make informed decisions when selecting and fine-tuning models.

Feature Engineering: Exploring feature engineering techniques taught me how to extract valuable information from raw data, optimizing model performance and interpretability.

**Non-Technical Takeaways:**

Collaborative Environment: I experienced a collaborative work culture where diverse skills and perspectives converged to solve complex challenges. Interactions with senior data scientists and cross-functional teams nurtured a sense of collective learning and growth.

Company Initiatives: Observing ongoing projects and initiatives underscored the role of data-driven decision-making in guiding strategic directions. It highlighted how data science plays a pivotal role in shaping business trajectories.

In conclusion, the summer internship training program provided me with a comprehensive understanding of Machine Learning's technical intricacies and its integration within a dynamic organizational framework. The practical exposure and holistic insights garnered from this program will undoubtedly shape my future endeavors in the realm of data science and beyond.

***INTRODUCTION***

During my second year as a B.Tech Computer Science student specializing in Data Science at The NorthCap University, I embarked on an enriching training program that provided a comprehensive understanding of machine learning aspects. This unique experience was facilitated within the university itself, under the expert guidance of faculty members Mrs. Srishti, Mrs. Sujata, and Mr. Nitin Malik.

**Company and Background**:

The training was carried out in collaboration with NorthCap University, where I am pursuing my degree. The university, nestled in the vibrant city of Gurugram, Haryana, has carved a niche for itself in providing exceptional education across various disciplines. With a strong emphasis on practical learning and industry-relevant skills, the university ensures that students like me are well-prepared for real-world challenges.

**Conclusion:**

The training program offered by NorthCap University, under the guidance of esteemed faculty members, provided a remarkable opportunity to bridge the gap between classroom learning and real-world application. By engaging in practical exercises and gaining insights into the intricacies of machine learning, I am now better equipped to tackle challenges and contribute effectively to the ever-evolving field of data science. This training experience has undoubtedly laid a strong foundation for my journey ahead.

***PROBLEM STATEMENT***

Credit card fraud has become a significant concern in the financial industry, leading to substantial financial losses for both financial institutions and cardholders. With the increasing prevalence of online transactions, there is a growing need for robust fraud detection systems that can accurately identify fraudulent transactions while minimizing false positives.

The aim of this project is to develop a highly accurate and efficient credit card fraud detection system using advanced machine learning techniques. The primary challenge lies in differentiating between legitimate and fraudulent transactions in real-time, as fraudulent activities are often masked by the sheer volume and complexity of legitimate transactions

***Social Relevance of the Project***

The social relevance of our Credit Card Fraud Detection project is significant and far-reaching. Here are some key points that highlight the social importance of our project:

**Financial Security:** Credit card fraud can cause substantial financial losses to individuals, leading to personal distress and financial instability. By developing an effective fraud detection system, your project directly contributes to safeguarding people's hard-earned money and financial well-being.

**Trust in Digital Transactions:** In an increasingly digital world, trust in online transactions is paramount. Credit card fraud incidents erode this trust, making people hesitant to engage in digital financial activities. A successful fraud detection system helps rebuild trust by assuring users that their transactions are being monitored and protected.

**Consumer Confidence:** With a reliable fraud detection mechanism in place, consumers are more likely to use their credit cards for transactions, both online and offline. This stimulates economic activity and contributes to a healthy and confident consumer market.

***Training Description***

**Objective**

It is vital that credit card companies can identify fraudulent credit card transactions so that customers are not charged for items that they did not purchase. Such problems can be tackled with Data Science and its importance, along with Machine Learning, cannot be overstated. This project intends to illustrate the modelling of a data set using machine learning with Credit Card Fraud Detection. The Credit Card Fraud Detection Problem includes modelling past credit card transactions with the data of the ones that turned out to be fraud. This model is then used to recognize whether a new transaction is fraudulent or not. Our objective here is to detect 100% of the fraudulent transactions while minimizing the incorrect fraud classifications. Credit Card Fraud Detection is a typical sample of classification. In this process, we have focused on analyzing and pre-processing data sets as well as the deployment of multiple anomaly detection algorithms such as Decision Tree, KNN and Random Forest Classifier on the PCA transformed Credit Card Transaction data.

**Data Extraction**

URL: <https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud>

* **We used the European Dataset available on Kaggle**
* **Dataset was extracted in the form of a csv file for easier implementation in python.**



**Features of Our Dataset**

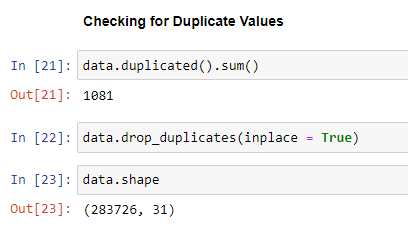
* It contains only numerical input variables which are the result of a PCA transformation.
* Features V1 to V28 are the principal components obtained with PCA.
* The only features which have not been transformed with PCA are 'Time' and 'Amount'.
* Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset.
* The feature 'Amount' is the transaction Amount.
* Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

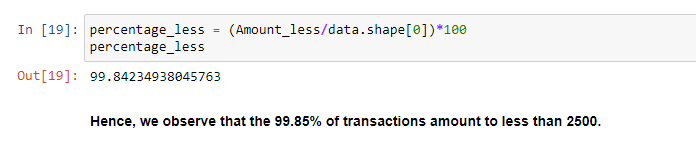
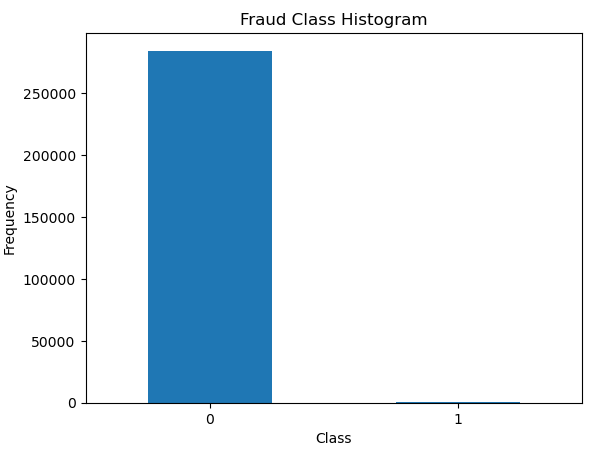
**Data Preprocessing**

The Dataset that we chose from Kaggle was already preprocessed i.e.

* No Missing Values
* No Impossible Values

**Dealing with Duplicate Values**



**Although, our data is highly imbalanced**

**Fixing our Dataset using SMOTE**

SMOTE (Synthetic Minority Over-sampling Technique) is a popular technique used in machine learning to address class imbalance in datasets. It is particularly useful when dealing with imbalanced datasets where the number of instances in the minority class (in this case, fraudulent transactions) is significantly lower than the number of instances in the majority class (non-fraudulent transactions).

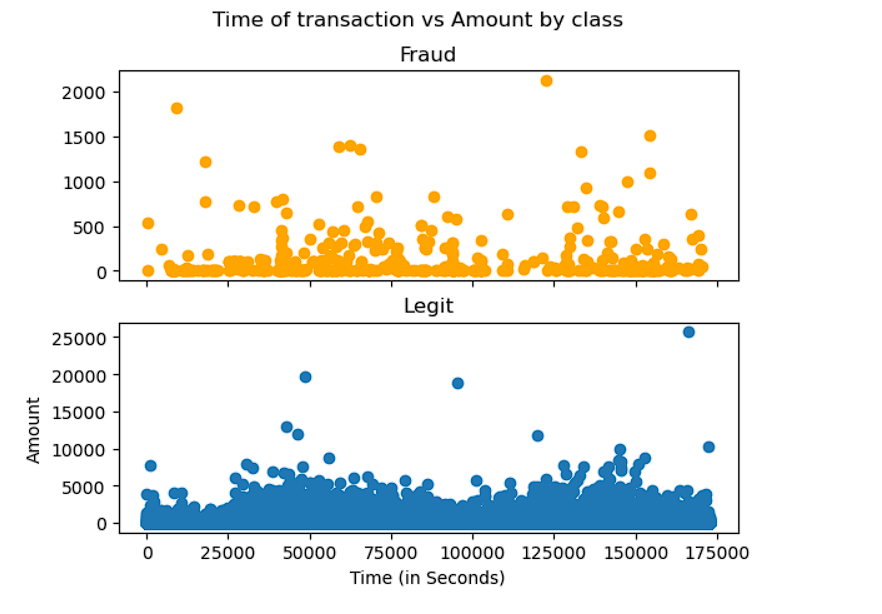


**Now our Dataset is fit to be tested for Model Training**

**Data Visualization**

* **Before SMOTE**

1. **Scatter Plot of Fraud and Legit Transactions**

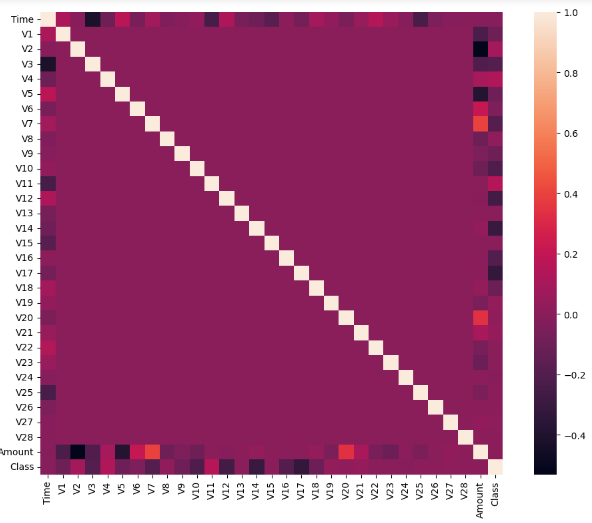


**Insights:**

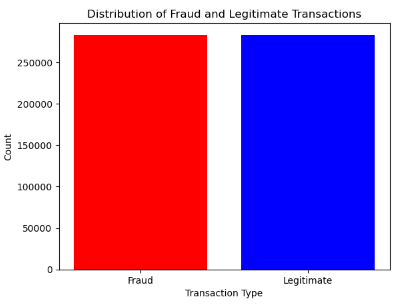
- It can be observed that the fraud transactions are generally not above an amount of 2500.

- It can also be observed that the fraud transactions are evenly distributed about time.

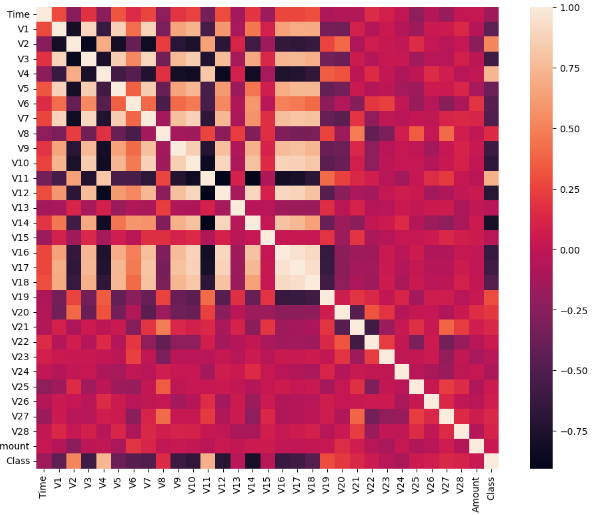
**Correlation Matrix / Heatmap:**

****

* In the Heatmap we can clearly see that most of the features do not correlate to other features but there are some features that either has a positive or a negative correlation with each other. For example, V2 and V5 are highly negatively correlated with the feature called Amount. This gives us a deeper understanding of the Data available to us.
* **After SMOTE**



**The Heatmap**



* **SMOTE (Synthetic Minority Over-sampling Technique) is a popular technique used in machine learning to address class imbalance in datasets. It is particularly useful when dealing with imbalanced datasets where the number of instances in the minority class (in this case, fraudulent transactions) is significantly lower than the number of instances in the majority class (non-fraudulent transactions).**
* Using SMOTE, we oversampled our fraudulent transactions to remove the inconsistency / imbalanced of our data.
* We also shuffled our data and applied frac = 0.2 to take 20% out from the original data for training and testing our various algorithms.

**Training our Model**

* After oversampling using Smote we shuffled our dataset (to remove biasness) and took a sample of 20% out for the final training of our data.

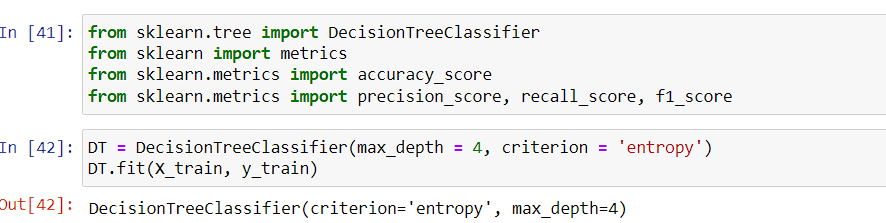


* We’ve split out training data and testing data in 80% and 20%

**ML Algorithms Used:**

* **Decision Tree Algorithm**

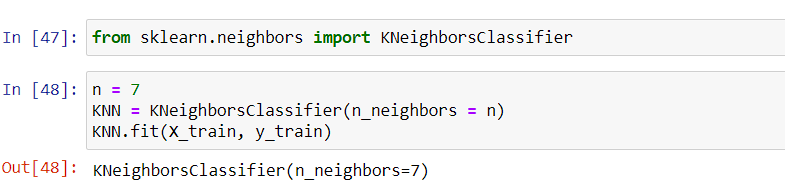
A Decision Tree is a versatile supervised machine learning algorithm used for both classification and regression tasks. It works by partitioning the feature space into segments based on the values of input features. The goal is to create a tree-like structure of decision nodes and leaf nodes, where each internal node represents a decision based on a feature, and each leaf node represents the predicted outcome.



Accuracy: 0.9593

* **K Nearest Neighbors (KNN)**

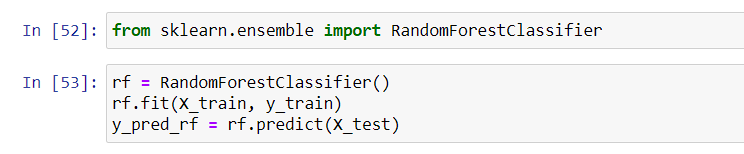
K-Nearest Neighbors is a simple and intuitive classification algorithm that classifies a data point based on the class of its k-nearest neighbors in the feature space. The algorithm doesn't learn explicit decision boundaries; instead, it makes predictions based on the majority class among the k-nearest neighbors.

****

Accuracy: 0.8857

* **Random Forest Classifier**

Random Forest is an ensemble learning algorithm that combines multiple Decision Trees to improve the overall predictive accuracy and control overfitting. Each tree in the forest is trained on a bootstrap sample of the data, and the final prediction is determined by aggregating the predictions of individual trees.

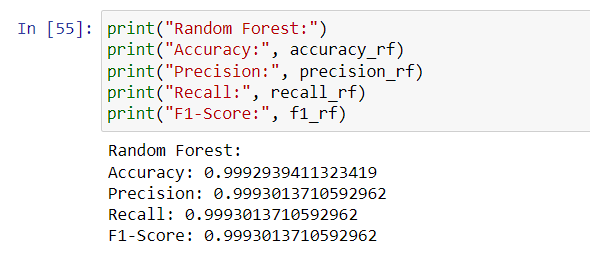


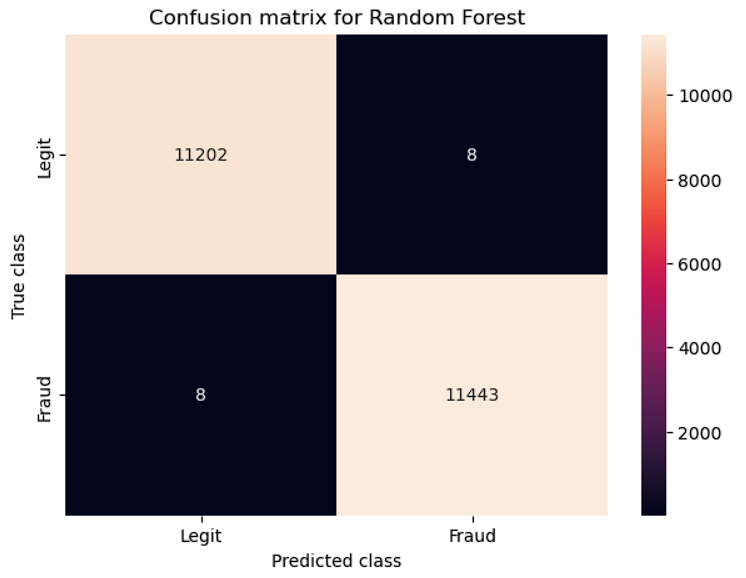
Accuracy: 0.9992

**Result**

So with our Random Forest Classifier, we got the best accuracy out of all

i.e. 99.92%



**Confusion Matrix**

***Analysis***

my participation in a real-world Data Science training program focused on machine learning has greatly contributed to my growth as an aspiring engineer. The insights gained – from the practical application of machine learning techniques to the development of a holistic problem-solving mindset – have widened my perspective of engineering's role in addressing real-world challenges.

**Technical Insights:**

Machine Learning Expertise: The specialized training focused on machine learning provided me with in-depth insights into the application of algorithms, predictive modeling, and data analysis. Engaging in hands-on projects enabled me to translate theoretical concepts into practical solutions, fostering a profound understanding of algorithmic functioning.

Real-world Application: Through practical exposure, I learned to bridge the gap between theory and real-world challenges. By employing machine learning techniques to solve complex problems, I understood the engineering profession's role in developing innovative solutions that impact industries.

Data Manipulation and Interpretation: My training honed my ability to manipulate and interpret large datasets, a fundamental skill in engineering. This equipped me to process raw data into meaningful insights, mirroring the way engineers convert raw materials into functional structures.

**Non-Technical Observations:**

Team Collaboration: Immersed in the training environment, I observed the value of cohesive teamwork. Witnessing professionals from diverse backgrounds collaborate on projects emphasized how collective efforts drive innovation in engineering solutions.

* I learnt about the basics of AIML in this training program. I worked on a real world project and it really enhanced my view on this vast topic of Machine Learning.  
  My primary strengths include a being a good listener, disciplined, believes in teamwork and I do possess some leadership qualities as well.

I did well in understanding the data, doing the exploratory data analysis, data visualization and data training and Testing.

I think I could’ve done better with fixing the highly imbalanced data and the biasness of

Legitimate transactions to get a more accurate result. I’d gladly do more research on this

topic

***CONCLUSION***

Participating in the summer training has significantly advanced my understanding of the engineering profession, both in technical and organizational aspects. The experience offered practical insights into real-world applications and collaboration dynamics, enhancing my knowledge and skills.

**Key Takeaways:**

* Adaptability: The training emphasized adapting to changing scenarios and learning new tools quickly. This mirrors engineering's need to keep up with evolving technologies.
* Effective Communication: Clear communication within the team and with supervisors was crucial for project success. This skill is equally essential in presenting engineering solutions coherently.
* Problem Solving: The training encouraged a systematic approach to problem-solving, a cornerstone of engineering. This approach ensures thorough analysis and effective solutions.

In conclusion, the summer training was an invaluable experience that magnified my understanding of the engineering profession. It blended technical prowess with organizational insights, fostering my growth as a well-rounded aspiring engineer.

***BIBLIOGRAPHY***

Problem Statement for our project “Credit Card Fraud Detection” was taken from:

<https://www.geeksforgeeks.org/machine-learning-projects/>

We took our dataset from Kaggle:

<https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud>

Sites used for further information on a few topics:

* JavaTPoint
* Geeksforgeeks

|  |  |  |
| --- | --- | --- |
| Project Daily Task | | |
| Date | **Day** | **Work** |
| 13th June, 2023 | Tuesday | Overview of AI and ML Concepts |
| 14th June, 2023 | Wednesday | Lecture on Data Exploration |
| 15th June, 2023 | Thursday | Lecture on Python: Data Structures |
| 16th June, 2023 | Friday | Lecture on Data Preprocessing |
| 19th June, 2023 | Monday | Principles of Supervised Learning |
| 20th June, 2023 | Tuesday | Expert Lecture on Web Development by Animcode CEO |
| 21st June, 2023 – 23rd June, 2023 | Wednesday - Friday | Animcode Interview |
| 26th June, 2023 | Monday | Lecture on Functions in python |
| 27th June, 2023 | Tuesday | Lecture on Decision Tree Algorithm |
| 28th June, 2023 | Wednesday | Lecture on Application of Decision Tree in python |
| 29th June, 2023 | Thursday | Introduction to Feature Selection |
| 1st July, 2023 | Saturday | Lecture on Underfitting and Overfitting |
| 3rd July, 2023 | Monday | Introduction to Classification & Underfitting and Overfitting in Python |
| 4th July, 2023 | Tuesday | Lecture on Greedy and Non – Greedy Quantifiers |
| 5th July, 2023 | Wednesday | Lecture on Feature Selection and Its Types |
| 6th July, 2023 | Thursday | Lecture on Data Visualization tools and important libraries |
| 10th July, 2023 | Monday | Lecture on Boundary Matchers and Splitting in Python |
| 11th July, 2023 | Tuesday | Lecture on KNN and Apriori Algorithm |
| 12th July, 2023 | Wednesday | Evaluation 1st of our Project Problem Statement |
| 13th July, 2023 | Thursday | Finalization of Project Details and how are we deciding to proceed |
| 14th July, 2023 | Friday | Presenting our Dataset to our esteemed Faculty |
| 17th July, 2023 | Monday | Presentation of work done to our Faculty and asked for suggestions |
| 18th July, 2023 – 19th July, 2023 | Tuesday - Wednesday | Free Day (To work on Improving our project and rectify and mistakes) |
| 20th July, 2023 | Thursday | Final Evaluation by our college faculty |
| 25th July, 2023 | Tuesday | Final Evaluation of our Project by Vdoit Experts |