Mobile Price Classification

Student 1	**180104036**
Student 2	**180104037**
Student 3	**180104043**

Project Report

Course ID: CSE 4214

Course Name: Pattern Recognition Lab

Semester: Spring 2021

Department of Computer Science and Engineering

Ahsanullah University of Science and Technology

Dhaka, Bangladesh

08 September 2022

Mobile Price Classification

Submitted by

Student 1	**180104036**
Student 2	**180104037**
Student 3	**180104043**

Submitted To

Faisal Muhammad Shah

Sajib Kumar Saha Joy,

Department of Computer Science and Engineering

Ahsanullah University of Science and Technology

Department of Computer Science and Engineering

Ahsanullah University of Science and Technology

Dhaka, Bangladesh

08 September 2022

ABSTRACT

To predict “If the mobile with given features will be Economical or Expensive” is the main motive of this research work, Predictive Model has become one of the most demanded areas of Information Technology and it has been successfully applied to solving many issues in machine learning, for example, Detecting Spam, Medical diagnosis, financial analysis, etc. This thesis describes developing a predictive model for mobile price prediction by its feature. Its related technologies have been widely used in industries. This will help industries for predicting the price of mobile trends very efficiently, it will discuss various algorithms and techniques used in predicting the mobile price with its procedure and also will be compared from another algorithm to identify the best algorithm and highest accuracy.

ACKNOWLEDGEMENT

The authors are grateful and pleased to all the researchers in this research study.

Contents

List of Figures

Chapter 1

Introduction

Price is the most productive attribute of marketing and business. The very first question in industries is about the price of items. All the industries are first worried and think “If he/she would be able to purchase something with the given specifications or not”. Machine learning provides us best techniques like supervised learning and unsupervised learning. Different tools are available for machine learning tasks like Python etc. We can use any of the classifiers like Linear Regression, KNN (K-Nearest Neighbors), and many more. Different type of feature selection algorithms is available to select only the best features and minimize the dataset. This will reduce the computational complexity of the problem. We have seven steps to make the predicting in machine learning Gathering data, Preparing that data, Choosing a model, Training, Evaluation, and Hyperparameter tuning. Prediction Mobile nowadays is one of the most selling and purchasing devices. Every day new mobiles with a new version and more features are launched. Many features are very important to be considered to estimate the price of mobile. For example, the Processor of the mobile, Battery, Size, and thickness. Internal memory, Camera pixels, video quality, Internet browsing. And so is the list of many features based upon those, the mobile price is decided. So, we will use many of the above-mentioned features to classify whether the mobile would be very economical, economical, and expensive or very expensive.

Chapter 2

Literature Reviews

2.1 Mobile Price Class prediction using Machine Learning Techniques

In paper [01] author has used only two models. One is the Decision tree classifier and another is Naïve Bayes classifier. The accuracy of these two classifiers is respectively about 70.14% and 64.83%. So, the decision tree classifier has given the best accuracy. The dataset is made from

2.2 Performance Evaluation of Different Supervised Learning Algorithms for Mobile Price Classification

In paper [02] author has implemented five models. These are Logistic Regression, Decision Tree, K – Nearest Neighbour (KNN), SVM, and Gradient Boost Algorithm. The accuracy of the models are respectively 81% for (Logistic Regression), 55% for (KNN), 82% for (decision tree), 84% for (SVM), and 90% for (gradient Boost). Gradient Boost Algorithm has given the best accuracy. Mobile price classification dataset is used.

2.3 A hybrid model for predicting Mobile Price Range using machine learning techniques

In paper [03] author has used two machine learning models are Decision tree and Random Forest and one is Hybrid Ensemble Learning. The accuracy of these two models are 83% for Decision tree and 90% for Random Forest. And the accuracy of Hybrid Ensemble Learning is about 94%. Hybrid Ensemble Learning performs best.

2.4 Mobile Price Prediction by its Features Using Predictive Model of Machine Learning

In paper [04] author has used four models. These are linear regression, k-nearest neighbors (KNN), Decision Tree, and Naïve Bayes. The accuracy of linear regression is 82%, KNN is 84%, Decision tree is 90%, and Naïve Bayes is 95%. Naïve Bayes algorithm has given the best accuracy among the all algorithm. Mobile price classification dataset is used.

Chapter 3

Data Collection & Processing

In our project, We use a mobile price classification train and test data. There are 2000 samples for training and 1000 samples for testing and also 21 features.

Preprocessing helps transform data so that a better machine learning model can be built, providing higher accuracy. The preprocessing performs various functions: checking the missing values and checking duplicate value.

	battery_power	blue	clock_speed	dual_sim	fc	four_g	int_memory	m_dep	mobile_wt	n_cores	pc	px_height	px_width	ram	sc_h	sc_w	talk_time	three_g	touch_screen	wifi	price_range
0	842	0	2.2	0	1	0	7	0.6	188	2	2	20	756	2549	9	7	19	0	0	1	1
1	1021	1	0.5	1	0	1	53	0.7	136	3	6	905	1988	2631	17	3	7	1	1	0	2
2	563	1	0.5	1	2	1	41	0.9	145	5	6	1263	1716	2603	11	2	9	1	1	0	2
3	615	1	2.5	0	0	0	10	0.8	131	6	9	1216	1786	2769	16	8	11	1	0	0	2
4	1821	1	1.2	0	13	1	44	0.6	141	2	14	1208	1212	1411	8	2	15	1	1	0	1

Figure 3.1: Train Dataset

	id	battery_power	blue	clock_speed	dual_sim	fc	four_g	int_memory	m_dep	mobile_wt	n_cores	pc	px_height	px_width	ram	sc_h	sc_w	talk_time	three_g	touch_screen	wifi
0	1	1043	1	1.8	1	14	0	5	0.1	193	3	16	226	1412	3476	12	7	2	0	1	0
1	2	841	1	0.5	1	4	1	61	0.8	191	5	12	746	857	3895	6	0	7	1	0	0
2	3	1807	1	2.8	0	1	0	27	0.9	186	3	4	1270	1366	2396	17	10	10	0	1	1
3	4	1546	0	0.5	1	18	1	25	0.5	96	8	20	295	1752	3893	10	0	7	1	1	0
4	5	1434	0	1.4	0	11	1	49	0.5	108	6	18	749	810	1773	15	8	7	1	0	1

Figure 3.2: Test Dataset

3.1 Dataset Preprocessing

3.1.1 Check null values:

Here, we are checking null values in the train data because null values usually take unnecessary spaces and hamper our work by consuming time.

3.1.2 Check unique values:

Here, we are checking the unique values in the train and test data. This helped us figure out finding the unique features in the dataset.

3.1.3 Splitting:

We have used 80% data to train our model and 20% data to test our model.

3.1.4 Feature Scaling:

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range.Since the range of values of raw data varies widely, in some machine learning algorithms, objective functions will not work properly without normalization.It is performed during data pre-processing.

Chapter 4

Methodology

4.1 Proposed Model

For our project, we used four machine learning algorithms like Logistic Regression, Support Vector Machine (SVM), Random Forest (RF), XGBOOST, and Gradient Boosting Classifier(GBC). For this, we preprocessed the data and split the data for training and testing the model. For training, we used 80% of the data and the rest of the 20% for testing purposes. After training the models, we evaluated the models with testing data.

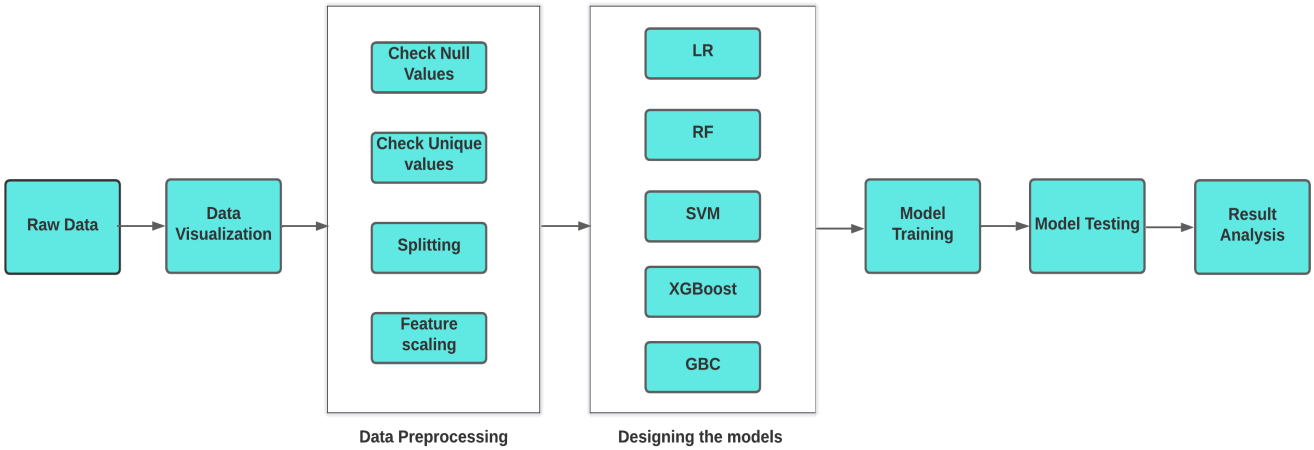

Figure 4.1: Flowchart of our project

4.2 Machine Learning Models

4.2.1 Logistic Regression :

Logistic Regression is a linear way of modeling between a dependent variable and an independent variable. Regression models dependent variables value based on independent variables. It is often used for finding out the relationship between variables and forecasting. Different regression models differ based on the kind of relationship between the dependent and independent variables. The Logistic Regression model gave an accuracy of 96.0% on the training data.

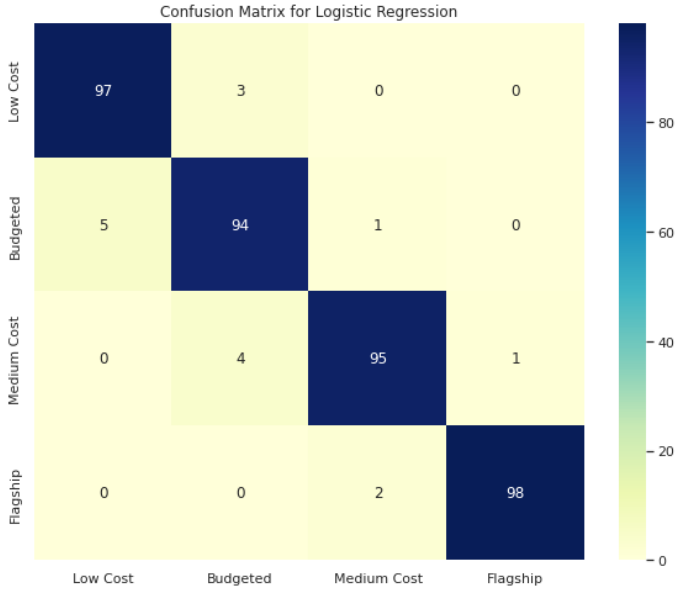

Figure 4.2: Confusion Matrix of Logistic Regression

4.2.2 Random Forest:

Random forest used to solve classification and regression problems. Random vectors are constructed ensembles of trees and decide the improved classification accuracy. Actually, while random forest is implemented to a dataset it creates multiple trees with a different score. Then it merges the trees and gives the best accuracy among them. The Random Forest model gave an accuracy of 91.25% on the training data.

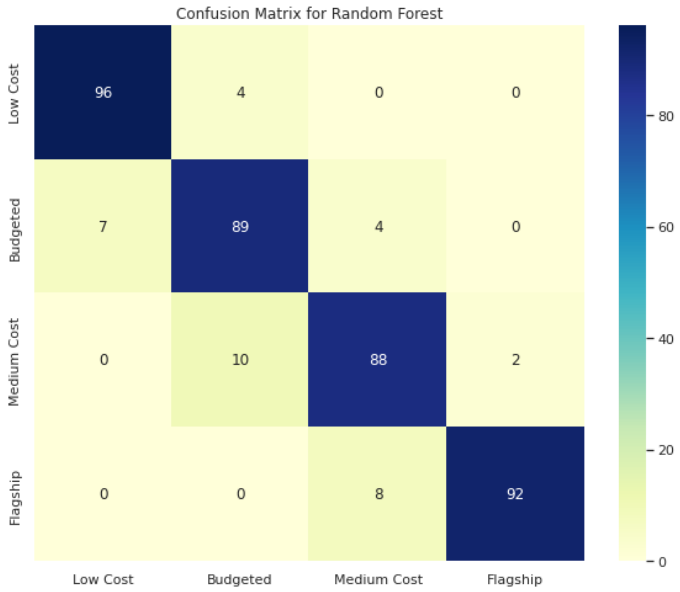

Figure 4.3: Confusion Matrix of Random Forest

4.2.3 Support Vector Machine (SVM):

In ML, SVM are supervised learning models associated with learning algorithms that inspect data used for classification and regression analysis. Determined a settled of activity cases, each show as going to one or the new of two gatherings, in SVM preparing calculation develops a model that apportions new cases to one gathering or the other, making it a non-probabilistic parallel direct classifier. The Support Vector Machine model gave an accuracy of 96.25% on the training data.

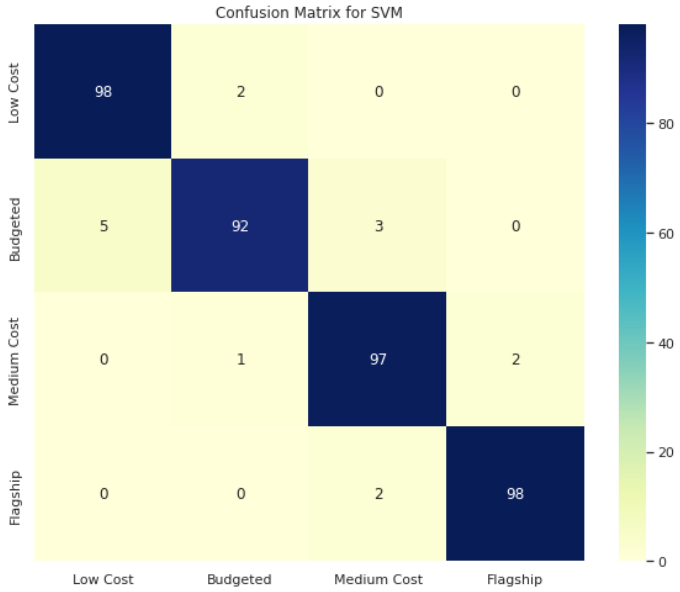

Figure 4.4: Confusion Matrix of Support Vector Machine

4.2.4 XGBoost (XGB):

XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. In prediction problems involving unstructured data (images, text, etc.) artificial neural networks tend to outperform all other algorithms or frameworks. However, when it comes to small-to-medium structured/tabular data, decision tree-based algorithms are considered best-in-class right now. The XGBoost model gave an accuracy of 92.0% on the training data.

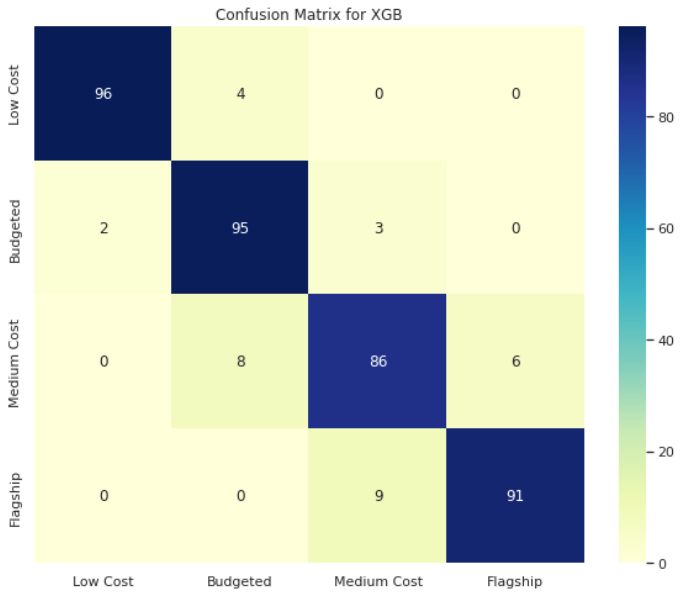

Figure 4.5: Confusion Matrix of XG Boost

4.2.5 Gradient Boosting Classifier(GBC):

When the target column is continuous, use Gradient Boosting whereas when it is a classification problem, we use Gradient Boosting Classifier. The only difference between the two is the "Loss function". The objective here is to minimize this loss function by adding weak learners using gradient descent. Since it is based on loss function hence for regression problems, we'll have different loss functions like Mean squared error (MSE) and for classification, we will have different log-likelihood.The Gradient Boosting model gave an accuracy of 91.25% on the training data.

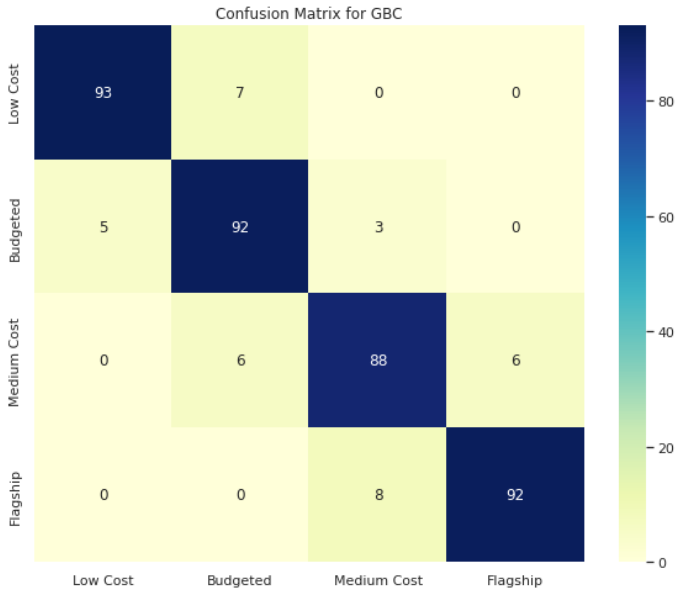

Figure 4.6: Confusion Matrix of Gradient Boosting Classifier

Chapter 5

Experiments and Results

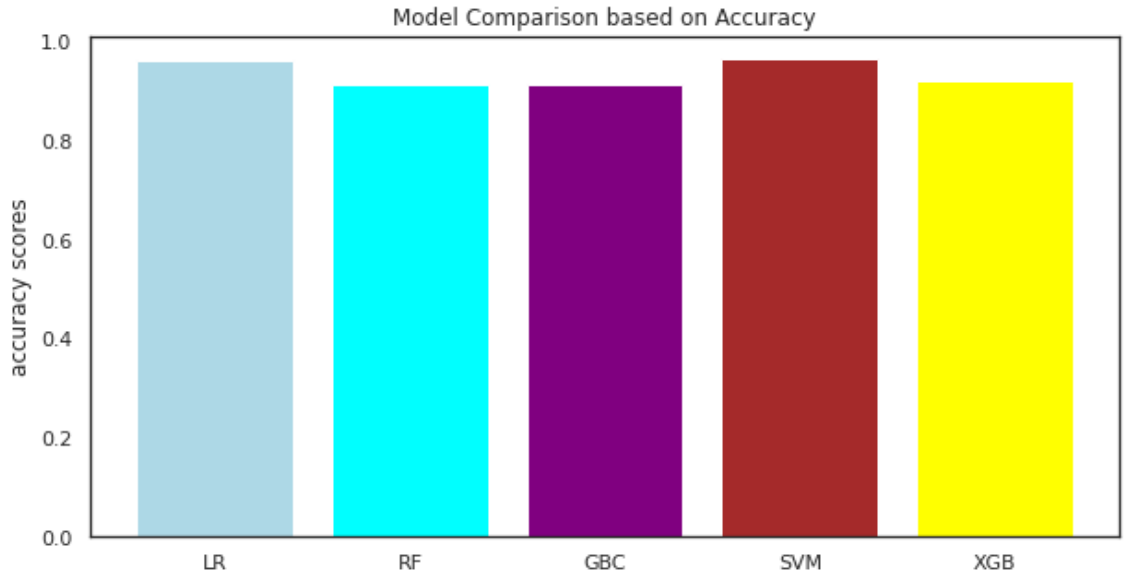

Figure 5.1: Accuracy of each model on Test Dataset

From the Fig:5.1, we can see Random Forest, Logistic Regression, Random Forest, Support Vector Machine, XGBoost and , Gradient Boosting Classifier models have corresponding accuracy of 96.0% , 91.25% , 96.25% , 92.0% , 91.25%. We can see that Support Vector Machine(SVM) performs better than all other models on classifying mobile price.

Classification Report for SVM

	low cost	budgeted	medium cost	flagship
Accuracy	98.21	97.22	97.96	98.97
Precision	95.15	96.84	95.10	98.00
Recall	98.00	92.00	97.00	98.00
F-score	96.55	94.36	96.04	98.00

From the above scenario, we see that the price range of four classes is low cost, budgeted, medium cost, and flagship. The flagship has high accuracy, precision, recall, and f-score.

5.1 ROC Curve

The receiver operating characteristic curve (ROC curve) is a graph that displays how well a classification model performs across all categorization levels. True Positive Rate and False Positive Rate are the two parameters plotted on this curve. It demonstrates the trade-off between specificity and sensitivity (or TPR) (1 – FPR). A better performance is shown by classifiers that provide curves that are closer to the top-left corner.

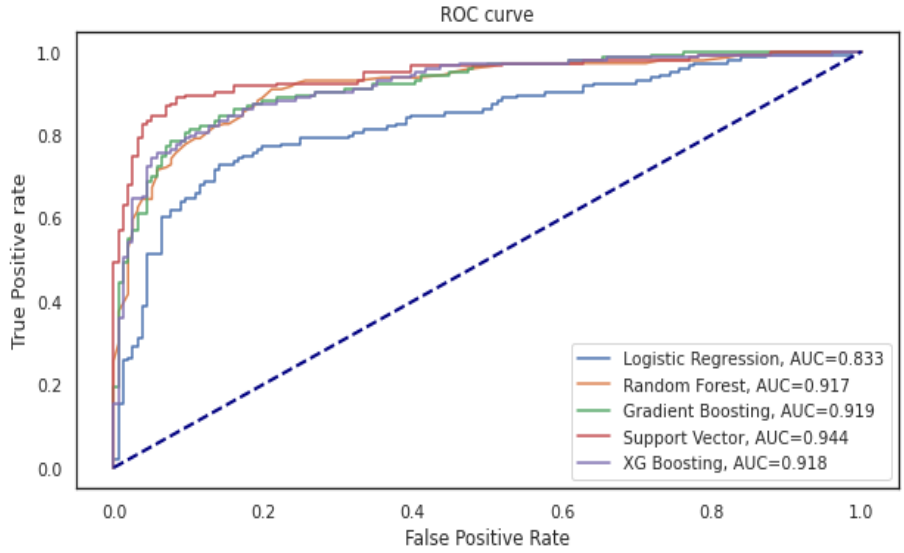

Figure 5.2: Receiver Operating Characteristic(ROC) Curve

Chapter 6

Future Work and Conclusion

6.1 Conclusion

In this project we have used mobile price classification dataset and experimented with five machine learning algorithms like Logistic Regression, Support Vector Machine (SVM),Random Forest Classifier, Gradient Boosting Classifier(GBC) and XGBoost(XGB). We trained and evaluated the stated model and found out SVM model performed better than other models on classifying mobile price range from the dataset. The SVM model achieved an accuracy of 96.25% which is very impressive. From this result, we could correctly predict the mobile price and classify it properly.

6.2 Future Work

To achieve maximum accuracy and predict more accurately, more and more instances should be added to the data set. And selecting more appropriate features can also increase the accuracy. So data set should be large and more appropriate features should be selected to achieve higher accuracy.
Software or Mobile app can be developed that will predict the market price of any newly launched product.

References

[1] Muhammad Asim, Zafar Khan, "*Mobile Price Class prediction using Machine Learning Techniques*", 29 March 2018, ResearchGate.

[2] Keval Pipalia, Rahul Bhadja, "*Performance Evaluation of Different Supervised Learning Algorithms for Mobile Price Classification*", June 2020, IJRASET.

[3] Hoque, Ahsanul, Sakib, Khan, Asif, Shakir, Shanjoy, Shutradhar, Abu, MD, Saleh, Washim, Akram, Md Badruzzaman, Khaleed Been, Biplop, "*A hybrid model for predicting Mobile Price Range using machine learning techniques*", January 2022, ACM.

[4] Akash A Gupta, Dr. Suhasini Vijaykumar, "*Mobile Price Prediction by its Features Using Predictive Model of Machine Learning*", February 2020, ResearchGate.

Generated using Undegraduate Thesis LaTeX Template, Version 1.4. Department of Computer Science and Engineering, Ahsanullah University of Science and Technology, Dhaka, Bangladesh.

This project report was generated on Wednesday 7th September, 2022 at 10:24pm.