Karunya INSTITUTE OF TECHNOLOGY AND SCIENCES

(Declared as Deemed to be University under Sec.3 of the UGC Act, 1956)

MoE, UGC & AICTE Approved

NAAC A++ Accredited

An internship report submitted by

SAM LEO S - Reg.No URK20CS2005

in partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

under the supervision of

Dr. M.Rajeswari **B.Tech. M.Tech. Ph.D.**

DIVISION OF COMPUTER SCIENCE AND ENGINEERING

KARUNYA INSTITUTE OF TECHNOLOGY AND SCIENCES

(Declared as Deemed to be University under Sec-3 of the UGC Act, 1956)

Karunya Nagar, Coimbatore - 641 114. INDIA

July 2023

Karunya INSTITUTE OF TECHNOLOGY AND SCIENCES
(Declared as Deemed to be University under Sec.3 of the UGC Act, 1956)
MoE, UGC & AICTE Approved
NAAC A++ Accredited

DIVISION OF COMPUTER SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that the report entitled, "SMART PHONE PRICE PREDICTION USING MACHINE LEARNING" is a bonafide record of Internship work done at INTEL UNNATI during the academic year 2023-2024 by

SAM LEO S (Reg. No: URK20CS2005)

in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering of Karunya Institute of Technology and Sciences.

Guide Signature

Dr. M.Rajeswari

B.Tech. M.Tech. Ph.D.

ACKNOWLEDGEMENT

First and foremost, I praise and thank ALMIGTHY GOD whose blessings have bestowed in me the will power and confidence to carry out my internship.

I am grateful to our beloved founders **Late. Dr. D.G.S. Dhinakaran, C.A.I.I.B, Ph.D** and **Dr. Paul Dhinakaran, M.B.A, Ph.D**, for their love and always remembering us in their prayers.

We entend Our tanks to **Dr. Prince Arulraj, M.E., Ph.D., Ph.D.,** our honorable vice chancellor, **Dr. E. J. James, Ph.D.,** and **Dr. Ridling Margaret Waller, Ph.D.,** our honorable Pro-Vice Chancellor(s) and **Dr. R. Elijah Blessing, Ph.D.,** our respected Registrar for giving me this opportunity to do the internship.

I would like to thank **Dr. Ciza Thomas, M.E., Ph.D.,** Dean, School of Engineering and Technology for his direction and invaluable support to complete the same.

I would like to place my heart-felt thanks and gratitude to **Dr. J. Immanuel John Raja, M.E., Ph.D.,** Head of the Division, Computer Science and Engineering for his encouragement and guidance. I feel it a pleasure to be indebted to, Mrs. **Dr. M.Rajeswari** **B.Tech. M.Tech. Ph.D.** Division of CSE & Mr. Srivatsa sinha M.tech for their invaluable support, advice and encouragement.

I also thank all the staff members of the School of CST for extending their helping hands to make this in Internship a successful one. I would also like to thank all my friends and my parents who have prayed and helped me during the Internship.

ABSTRACT

The advancement of technology, cell phones have become an essential component of daily life. Mobile has brand, internal memory, wifi, battery life, camera, and the accessibility of 4G is currently changing how people choose their mobile phone models. But people don't always connect those factors to the cost of mobile phones; in this case, this paper aims to solve the issue by training the mobile phone dataset with machine learning algorithms like Support Vector Machine, Decision Tree, K Nearest Neighbours, and Nave Bayes,Logical regression before making predictions about the price level. In order to forecast smartphone pricing based on accuracy score ,we applied the proper algorithms. This not only improves customer choice on mobile phones but also offers guidance firms who sell mobile phones how to balance their many feature offerings with fair pricing. This theory of price prediction will provide assist clients in making future smart phone selections.

The research aims to estimate smartphone pricing by applying these algorithms and taking into account their accuracy scores. This strategy not only expands the options available to customers when choosing mobile phones, it also offers advice to mobile phone sales companies on how to balance feature offerings with reasonable price. The price prediction hypothesis created in this research will help customers decide wisely about next smartphone purchases.

In conclusion, this study trains a dataset of mobile phone prices using machine learning methods. By doing this, it hopes to improve customers' ability to choose a mobile phone and help businesses set reasonable prices for their goods. Future smartphone purchase decisions by customers will benefit from the proposed price prediction hypothesis.

CHAPTER 1

1.INTRODUCTION

Science and technology have made the globe much more fantastic than it ever was in the contemporary era. Given that pricing has a significant impact on customer decisions to purchase smartphones, the public's attention is currently being drawn to the shifting costs of mobile phones. Brand loyalty is a significant factor, and [1] shows the market shares of recent mobile phone brands in China, from which we can infer that the market share is directly correlated with consumer preferences: Huawei ranks first with a market share of 28.1%, followed by Oppo with 13.3% and Apple with 11.3%. Despite this, a number of elements, such the camera, memory, battery, and screen, will impact the cost of mobile phones. It should be obvious that numerous The association between cellphone pricing and consumer behaviours was established by academics; however, the study gap is that individuals have little understanding of how different functions correspond to different pricing levels. As can be seen from the information above, the importance of this research goes beyond simply being able to predict the price of mobile phones based on various properties. It also allows consumers to have some understanding of the price range of mobile phones, which improves people's ability to make rational purchasing decisions in the future. The study's goal is to use machine learning techniques to estimate mobile pricing levels given smartphone feature information. This aids in raising people's awareness of both the pricing and features of mobile phones. The dataset was obtained from the Kaggle website. Four machine learning algorithms (DT, KNN, and LG) were used to fit the training dataset of mobile price and make a prediction on the price level after preprocessing the data and selecting the most important four features (in this case, they are ram, battery_power, px_width, and px_height).

1.2DECISION TREE

A supervised machine learning technique for classification is called decision trees. By learning direct decision rules, decision trees try to develop a prediction model and infer the values of target variables from data attributes. A tree can be thought of as an invariable piecewise estimate. Additionally, decision trees are employed in a variety of industries, including medical and data mining research, due to the straightforward algorithmic basis, the efficiency, and the visualisation of trees. describes how the decision tree is used to predict the kind of ECG beats, whereas Abraham and Mark determine how to gather the most accurate data. one that is indicative of the many decision-tree models.

The decision tree model is the second model that may match our mobile phone dataset as a result. To choose the optimal parameter, we utilise grid search CV once again. The result is a balanced class weight, with the values of y causing weights to change on their own in a manner that is inversely proportionate to class frequencies in the input data. We select "entropy" as the information gain since the parameter "criterion" refers to the capacity to evaluate the quality of a segmentation.

1.3 K-NEAREST NEIGHBOURS

Neighbors-based classification is a kind of non-generalizing learning technique based on actual cases; it only maintains examples from the training data without attempting to create a typical internal model. The most popular method, known as KNN, is based on each query point's closest neighbours. The KNN classification algorithm's steps are simple to understand: First, choose a value for k (which defines how many nearest neighbours we should select), then calculate the distance between each item in the collection of training objects and the test object, and finally choose the training object that is the closest to the test object. The next step is to choose the class that has the most objects that match it. Finally, there is only must repeat the procedure until the desired class is attained. Regarding the use of this methodology, researchers classify ECG signals It has a 97.5% accuracy rate and is useful for diagnosing cardiac problems.

1.4 LOGISTIC REGRESSION

The majority of medical areas, social sciences, and machine learning all employ logistic regression. For instance, Boyd et al. utilised logistic regression to create the

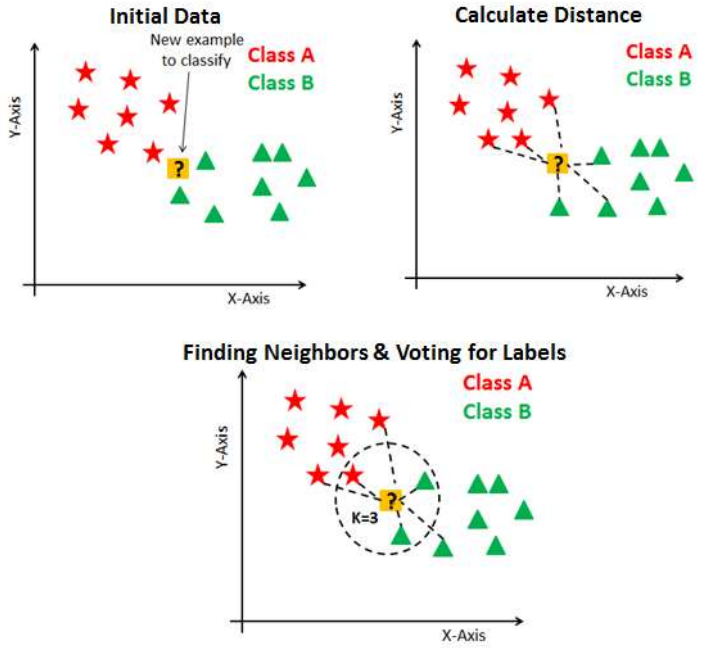


FIG.1

Trauma and Illness Severity Score (TRISS), which is frequently used to forecast death in wounded patients. Logistic regression has been used to generate several more medical

measures that are used to evaluate a patient's severity. Based on patient parameters that have been observed (such as age, sex, body mass index, outcomes of different blood tests, etc.), logistic regression may be used to forecast the likelihood of getting a certain illness (such as diabetes or coronary heart disease). Using a voter's age to forecast whether they would support the Nepali Congress, the Communist Party of Nepal, or any other party is another example income, gender, race, place of domicile, prior election votes, etc. The method is also applicable to engineering, particularly when determining the likelihood that a certain process, system, or product may fail. Additionally, it is utilised in marketing applications like forecasting a customer's probability to buy a product or cancel a subscription, among others.In economics, it may be used to forecast whether someone will join the labour market, and in business, it can be used to forecast if a homeowner would fail on a mortgage. Natural language processing employs conditional random fields, a logistic regression extension to sequential data.

CHAPTER 2

2.1 DATA PREPROCESSING

The dataset, titled "Mobile Price Classification," is acquired from Kaggle.com and used to assess methodologies. The data's historical context is that someone is attempting to launch his own mobile firm, and he is trying to figure out how a smartphone's pricing relates to its features (such wifi connectivity and battery life). You can view the specifics of the data for "train" and "test" It is clear that the train data has 2000 samples, 20 features, and 1 column of the target price range, whereas the test data contains 1000 samples and the same 20 characteristics, but without a price range. This cannot serve as a benchmark when we forecast the goal price range. Using the train-test split approach, we only utilise the training dataset as the whole set of data in this case, with the remaining 600 samples serving as testing data. Fortunately, no values are missing, and the descriptive statistics of the data show

TABLE 1-- INFORMATION OF THE TRAIN DATA

Column	Non-Null count	Dtype	Column	Non-Null count	Dtype
Battery power	2000	Int64	Px height	2000	Int64
blue	2000	Int64	Px width	2000	Int64
Clock speed	2000	Float64	Ram	2000	Int64
Dual sim	2000	Int64	Sc h	2000	Int64
Fc	2000	Int64	Sc w	2000	Int64
Four g	2000	Int64	Talk time	2000	Int64
Int memory	2000	Int64	Three g	2000	Int64
M dep	2000	Float64	Touch screen	2000	Int64
Mobile wt	2000	Int64	Wifi	2000	Int64
N cores	2000	Int64	Price range	2000	Int64
Pc	2000	Int64			

(Note: Battery power is expressed in milliampere-hours (mAh); Bluetooth is present; Dual SIM is supported; Front Camera is expressed in Megapixels; Internal Memory is expressed in Gigabytes; Mobile depth and weight are referred to as "m dep and mobile wt"; The processor's core count is indicated by the letter "n," while "pc" stands for "Primary Camera" and "px" stands for "Pixel Resolution Height and Width," "ram" stands for "Random Access Memory" in megabytes, and "sc h" and "sc w" stand for "Screen Height and Width of Mobile in cm" respectively. The amount of time you can chat on the phone on a single charge is called talk_time; The price range is separated into four sections, with the lowest price level represented by numbers 0, 1, and 3. correspondingly, the median price level, high price level, and extremely high price level.

TABLE 2-- INFORMATION OF THE TEST DATA

Column	**Non-Null count**	**Dtype**	**Column**	**Non-Null count**	**Dtype**
Battery power	1000	Int64	Px height	1000	Int64
blue	1000	Int64	Px width	1000	Int64
Clock speed	1000	Float64	Ram	1000	Int64
Dual sim	1000	Int64	Sc h	1000	Int64
Fc	1000	Int64	Sc w	1000	Int64
Four g	1000	Int64	Talk time	1000	Int64
Int memory	1000	Int64	Three g	1000	Int64
M dep	1000	Float64	Touch screen	1000	Int64
Mobile wt	1000	Int64	Wifi	1000	Int64
N cores	1000	Int64	Pc	1000	Int64

TABLE 3-- DESCRIPTIVE STATISTICS OF MOBILE PHONE SAMPLE

	count	mean	std	min	25%	50%	75%	max
battery_power	2000	1238.519	439.4182	501	851.75	1226	1615.25	1998
blue	2000	0.495	0.5001	0	0	0	1	1
clock_speed	2000	1.52225	0.816004	0.5	0.7	1.5	2.2	3
dual_sim	2000	0.5095	0.500035	0	0	1	1	1
fc	2000	4.3095	4.341444	0	1	3	7	19
four_g	2000	0.5215	0.499662	0	0	1	1	1
int_memory	2000	32.0465	18.14571	2	16	32	48	64
m_dep	2000	0.50175	0.288416	0.1	0.2	0.5	0.8	1
mobile_wt	2000	140.249	35.39965	80	109	141	170	200
n_cores	2000	4.5205	2.287837	1	3	4	7	8
pc	2000	9.9165	6.064315	0	5	10	15	20
px_height	2000	645.108	443.7808	0	282.75	564	947.25	1960
px_width	2000	1251.516	432.1994	500	874.75	1247	1633	1998
ram	2000	2124.213	1084.732	256	1207.5	2146.5	3064.5	3998
sc_h	2000	12.3065	4.213245	5	9	12	16	19
sc_w	2000	5.767	4.356398	0	2	5	9	18
talk_time	2000	11.011	5.463955	2	6	11	16	20
three_g	2000	0.7615	0.426273	0	1	1	1	1
touch_screen	2000	0.503	0.500116	0	0	1	1	1
wifi	2000	0.507	0.500076	0	0	1	1	1
price_range	2000	1.5	1.118314	0	0.75	1.5	2.25	3

2.2 CORRELATION MATRIX

The correlation matrix, which displays the pairwise correlation coefficients between variables in a dataset, is visualised as a correlation heatmap. Finding the strength and direction of the correlations between the variables is helpful. A complete positive correlation is represented by a correlation value of 1, whereas a perfect negative correlation is represented by a correlation coefficient of -1

CORRELATION MATRIX

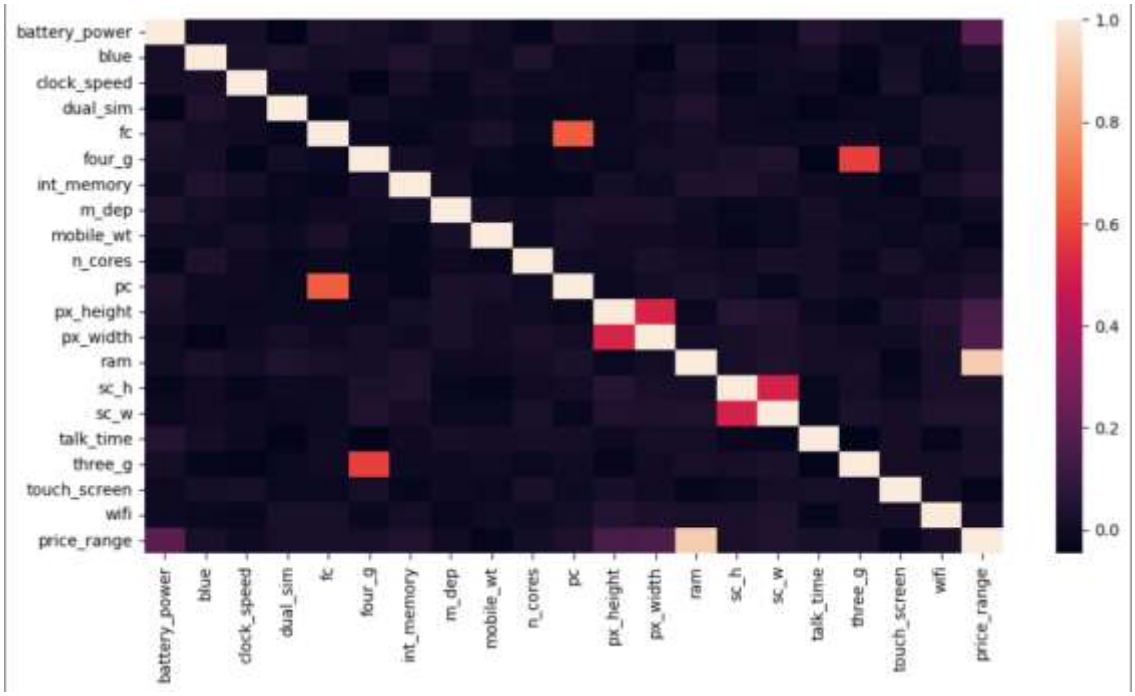


FIG 2.1

2.3 STANDARD SCALAR

With the use of the preprocessing method known as StandardScaler, numerical characteristics may be transformed by being scaled to have a zero mean and unit variance. It is a component of the Python scikit-learn package. The training and testing datasets are both transformed by StandardScaler using the mean and standard deviation of each feature in the training dataset. This makes sure that both datasets receive the same scaling treatment consistently.

StandardScaler is used to scale all features to the same value, which is useful for some machine learning techniques. When features are standardised, several algorithms, including logistic regression, support vector machines, and k-nearest neighbours, perform better or converge more quickly.


#StandardScaler
from sklearn.preprocessing import StandardScaler
std=StandardScaler()
X_std=std.fit_transform(X)
data_test_std=std.transform(data_test)
X_std
array([[-0.90259226, -0.9900495 , 0.83077943, ..., -1.70686097,
-1.00601811, 0.98609664],
[-0.40515057, 1.0100505 , -1.2530642 , ..., 0.55964063,
0.99401789, -1.01409939],
[-1.5376065 , 1.0100505 , -1.2530642 , ..., 0.55964063,
0.99401789, -1.01409939],
...,
[1.53077136, -0.9900495 , -0.76274085, ..., 0.55964063,
0.99401789, -1.01409939],
[0.62252745, -0.9900495 , -0.76274085, ..., 0.55964063,
0.99401789, 0.98609664],
[-1.6583306 , 1.0100505 , 0.58562134, ..., 0.55964063,
0.99401789, 0.98609664]])

FIG.2.2

CHAPTER 3
3.1 TRAINING THE MODEL

3.1.1 DECISION TREE

A common supervised machine learning technique for classification and regression applications is the decision tree. Each internal node represents a feature or characteristic, each branch represents a decision rule based on that attribute, and each leaf node represents the result or the class label in this tree-like flowchart structure.

Decision trees are favoured by data scientists and analysts because they are simple to understand and use. They emulate human decision-making logic by posing a sequence of questions depending on the characteristics of the data, which makes the decision-making process clear and easy to grasp.

Decision trees have a number of benefits, including their usability, interpretability, and capacity to handle both category and numerical data. However, they can be overfitting-prone and sensitive to even little changes in the training data, especially if the tree is excessively deep or complicated. These problems may be solved utilising regularisation approaches including pruning, hyper parameter tweaking, and ensemble methods.

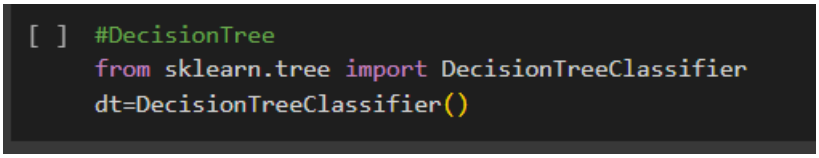


FIG 3.1

3.1.2 KNN

KNN offers a number of benefits, including its simplicity, adaptability to changes in the training data, and versatility in addressing both classification and regression issues. The prediction time rises as the size of the training set increases, although it can be sensitive to the amount of the input features.

It's important to remember that KNN is a lazy learning algorithm, which means that during the training stage it doesn't create an explicit model. Instead, it memorises the whole training dataset and uses the stored instances to directly make predictions at runtime.

KNN is an effective technique for many applications, especially when the decision boundary or decision surface is poorly defined or when the underlying data distribution is complicated.

```
[ ]  #KNN
     from sklearn.neighbors import KNeighborsClassifier
     knn=KNeighborsClassifier()
```

FIG 3.2

3.1.3 LOGISTIC REGRESSION

Numerous benefits of logistic regression include its ease of use, interpretability, and effectiveness in training and making predictions. It performs well if the underlying data distribution is not significantly skewed or if the relationship between the input characteristics and the binary outcome can be represented by a linear decision boundary.

For multi-class classification issues with more than two classes, extensions of logistic regression, such as multinomial logistic regression, can be employed.

It's crucial to understand that logistic regression makes the assumption that the correlation between the input characteristics and the binary outcome's log-odds is linear. The use of feature engineering techniques or other machine learning algorithms may be more acceptable if the connection is nonlinear.

In general, logistic regression is a popular and adaptable method for binary classification. activities, giving insights into how input factors affect the likelihood of a particular outcome.

```
[ ]  #LogisticRegression
     from sklearn.linear_model import LogisticRegression
     lr=LogisticRegression()
```

FIG 3.3

CHAPTER 4

4.1 ACCURACY SCORE

The percentage of accurate predictions generated by a model out of all predictions is measured by the accuracy score, a popular assessment metric in classification tasks. It has a range of 0 to 1, with 0 denoting no valid predictions and 1 denoting all of them being correct. A higher accuracy number denotes a greater percentage of accurate forecasts, whereas a lower value denotes a greater percentage of inaccurate predictions.

It's crucial to evaluate the accuracy score in light of the issue, though. A high accuracy rating might be deceiving in cases when classes are uneven since the model may be biassed in favour of the dominant class. Additionally, taking into account additional evaluation metrics like accuracy, recall, F1-score, or the area under the ROC curve allows for a more thorough review of the efficiency of the model.

In order to comprehend true positive, true negative, false positive, and false negative predictions and develop a more complex understanding of the model's efficacy beyond only accuracy, it is essential to analyse the confusion matrix.

4.1.1 Decision Tree after split with accuracy score

You may evaluate the model's effectiveness using an accuracy score after segmenting the dataset and creating a decision tree. The percentage of accurate predictions the decision tree made on the testing set is represented by the accuracy score. It is determined by dividing the total number of forecasts by the number of correct guesses.

A higher accuracy number denotes a greater percentage of accurate forecasts, whereas a lower value denotes a greater percentage of inaccurate predictions. You may learn more about how effectively the decision tree predicts the class labels on the testing set by taking into account the accuracy score.

It's crucial to remember that the accuracy score may not give a whole view of the model's performance, hence it is advised to take additional assessment measures into account.

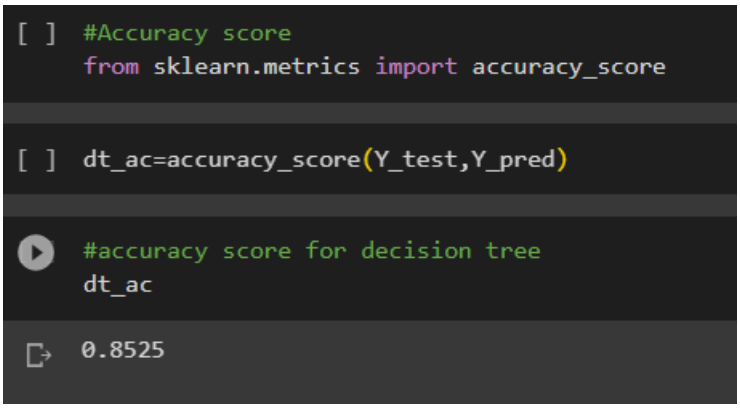


FIG 4.1

4.1.2 KNN after split with accuracy score

An accuracy score may be used to assess the model's performance following the partitioning of the dataset and training of the K-Nearest Neighbours (KNN) method. The percentage of accurate predictions the KNN model produced on the testing set is represented by this score. The amount of accurate predictions is calculated by dividing the total number of forecasts by the number of correct predictions, which are obtained by comparing the predicted class labels to the actual class labels in the testing set. A greater accuracy score suggests that the KNN model does a good job predicting the class labels, as indicated by a larger percentage of right predictions. A lower accuracy score, on the other hand, denotes more inaccurate predictions. The accuracy score should be read in conjunction with other assessment measures, it is crucial to mention. To gain a thorough grasp of the model's performance, look at metrics like accuracy, recall, F1-score, or the area under the ROC curve. The performance of the KNN model in producing correct predictions on the testing set may be evaluated using an accuracy score.

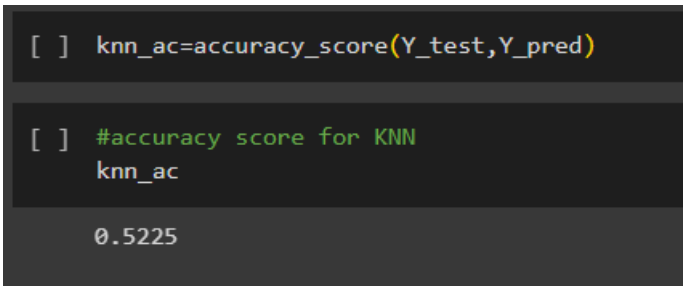


FIG 4.2

4.1.3 Logistic Regression after split with accuracy score

An accuracy score may be used to evaluate the model's performance after dividing the dataset and training the logistic regression method. The percentage of accurate predictions produced on the testing set by the logistic regression model is represented by the accuracy score.

The amount of accurate predictions is calculated by dividing the total number of forecasts by the number of correct predictions, which are obtained by comparing the predicted class labels to the actual class labels in the testing set.

A greater accuracy score represents a larger percentage of true predictions, demonstrating how effectively the Logistic Regression model predicts the binary class labels. A lower accuracy score, on the other hand, denotes more inaccurate predictions. It's critical to remember that the accuracy score should be weighed together To gain a thorough picture of the model's performance, use other assessment measures like accuracy, recall, F1-score, or the area under the ROC curve. The success of the Logistic Regression model in producing correct predictions on the testing set may be evaluated using an accuracy score.

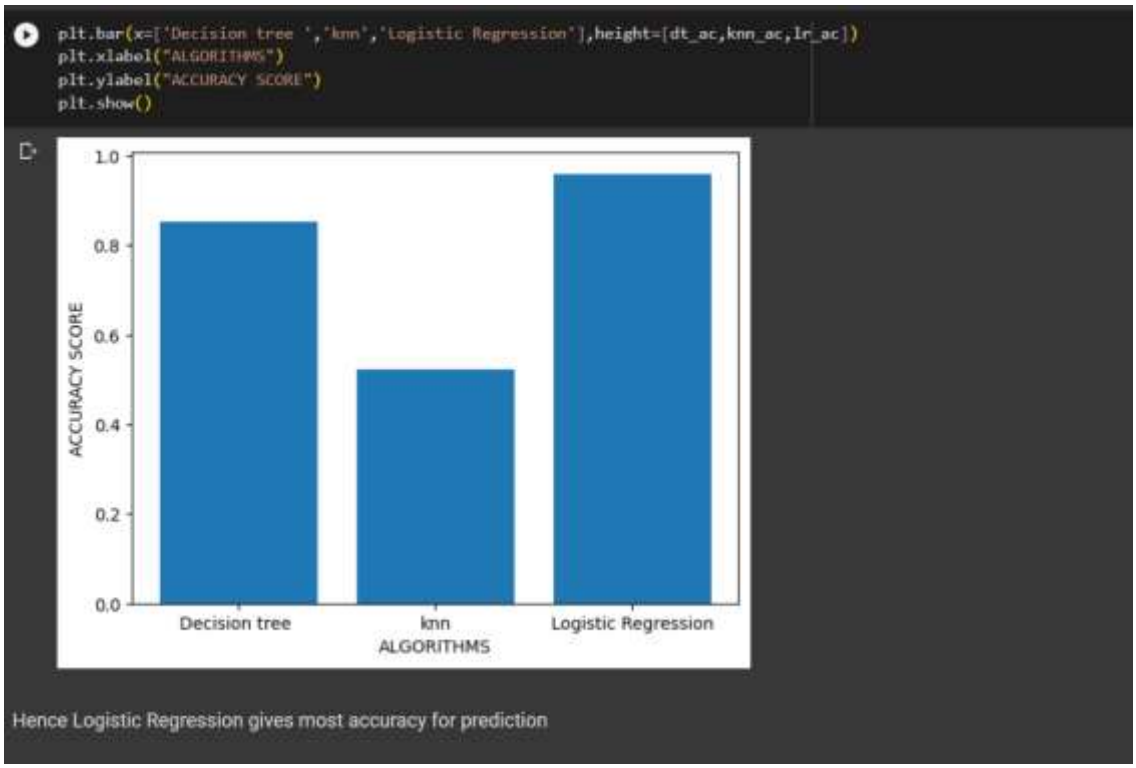
```
[ ]  lr_ac=accuracy_score(Y_test,Y_pred)

[ ]  #accuracy score for Logistic Regression
     lr_ac

     0.96
```

FIG 4.3

4.2 Model Accuracy and Selection



CHAPTER 5

5.CONCLUSION

The major purpose of the study is to estimate the cost of mobile devices based on attributes like Bluetooth compatibility and battery life. The study's conclusion suggests that the main goal has been accomplished. For data gathered from the Kaggle website published by Abhishek Shaema, the four machine learning techniques were successful in predicting the mobile price level using the features of ram, battery_power, px_width, and px_height (random access memory, battery power, pixel resolution height and width).To forecast the price level, we utilised Support Vector Machine, Decision Tree, K Nearest Neighbours, and Naive Bayes classifiers. The performance was tracked and evaluated using accuracy, precision, recall, and F1 score. Depending on the results As seen above, the SVM classifier produces the best results with and without feature selection, whereas the Nave Bayes classifier produces the worst results in terms of accuracy, precision, recall, and F1 score. The other two classifiers' differences are apparent since, during the investigation, the K-Nearest Neighbour classifier provided a result of around 95%, and the Decision Tree classifier produced an outcome of about 90%.including and excluding feature choices. Additionally, there has been a general improvement since the introduction of the feature selection process, which creates a strong basis for customers to use to minimise costs when selecting mobile phones at various price points. To make predictions utilising other sorts of feature selection using other machine learning algorithms, more investigation may be required. From a business standpoint, the forecast would be helpful to comprehend how to establish a price according to different functionalities of smartphones; on the other side, the outcome may be helpful to customers in choosing mobile phones at various price points.People now have a better knowledge of how functions relate to one another as a result and price level of mobile phones.

REFERENCE

[1] Li H., Wei Y., “Analysis the Impacting of “User Experience” for Chinese Mobile Phone’s Brands Market Changing”, Design, User Experience, and Usability. Practice and Case Studies, pp277-287, 2019.

[2] Cristina Butnariu, Catalin Lisa, Florin Leon and Silvia Curteanu, “Prediction of liquid-crystalline property using support vector machine classification”, Journal of Chemometrics, June 2013.

[3] Kassio M.G. Lima, Laurinda F.S. Siqueira, Camilo L.M. Morais, “SVM for FT-MIR prostate cancer classification: An alternative to the traditional methods”’ Journal of Chemometrics, July 2018.

[4] Parisa Pouladzadeh, Shervin Shirmohammadi, Aslan Bakirov, Ahmet Bulut & Abdulsalam Yassine, “Cloud-based SVM for food categorization”, Multimedia Tools and Applications, June 2014.

[5] Mohebbanaaz, L. V. Rajani Kumari & Y. Padma Sai, “Classification of ECG beats using optimized decision tree and adaptive boosted optimized decision tree”, Signal, Image and Video Processing, Oct 2021.

[6] Abraham Itzhak Weinberg & Mark Last, “Selecting a representative decision tree from an ensemble of decision-tree models for fast big data classification”, Journal of Big Data, Feb 2019.

[7] C. Venkatesan, P. Karthigaikumar & R. Varatharajan, “A novel LMS algorithm for ECG signal preprocessing and KNN classifier-based abnormality detection”, Multimedia Tools and Applications, Mar 2018.

[8] Shi H., Liu Y., “Naïve Bayes vs. Support Vector Machine: Resilience to Missing Data”, Artificial Intelligence and Computational Intelligence, pp680-687, 2011.

[9] Thomas Rincy N, Roopam Gupta, “An efficient feature subset selection approach for machine learning”, Multimedia Tools and Applications, Jan 2021.