iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 Issue: VI Month of publication: June 2020

DOI: http://doi.org/10.22214/ijraset.2020.6302

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue VI June 2020- Available at www.ijraset.com

Performance Evaluation of Different Supervised Learning Algorithms for Mobile Price Classification

Keval Pipalia[1], Rahul Bhadja[2]
[1, 2]Dept. of Computer Engineering, Marwadi Education Found.

Abstract: This research-paper aims at comparing the accuracy of different classification algorithms used in supervised machine learning. Classification Problem is about to find out in which class each example is related within a given dataset. It is used to classify the data instances into different groups according to some characteristics. We used several famous supervised learning algorithms - Logistic Regression, K-Nearest Neighbours (KNN), Decision Tree, Support Vector Machine (SVM), and Gradient Boosting to classify the range of Mobile price. We have created multiple classifiers for Mobile Price classification and compared their accuracy on the data taken from kaggle. Results are compared in terms of outcome accuracy score achieved from the research experiment. Conclusion is made for the best classifier for the mobile price classification problem.

Keywords: Machine Learning, Supervised Learning, Classification, Logistic Regression, Decision Tree, KNN, SVM, Gradient Boost Algorithm.

I. INTRODUCTION

Scientific Computing anticipates on executing computer algorithms. Given a particular available environment and hardware, algorithm's accuracy is a deciding factor. There are many supervised learning algorithms available for classification problem and many languages available to execute it like Python, R, MATLAB etc. But Python particularly has the best libraries and tools used in Machine Learning. Python provides scikit-learn library which contains simple and efficient tools for executing the supervised learning algorithms.

Millions of mobile are sold and purchased globally. So here the kaggle- mobile price classification is an example dataset for the given type of problem i.e. finding optimal class. The same work can be done to classify real price-range of all products like cars, bikes, Electronic items, medicine, Housing-price etc.

Here, result is taken as a single opinion score 'accuracy' – best used in comparison of algorithms. Price-class is calculated and decides whether the mobile is very economical, economical, and expensive or very expensive. In this paper, we compare the popular supervised learning- approaches (Classification, Logistic Regression, Decision Tree, K-Nearest Neighbour (KNN), Support Vector Machine (SVM), and Gradient Boost Algorithm) in the context of classification. It gives the opinion on the mobile price depending on the features used.

II. LITERATURE REVIEW

Several articles have studied and investigated machine learning algorithms. First Mark Schmidt focused on SVM and Structural SVM and compared it with logistic regression algorithm with accuracy testing. He found SVM to be more effective compared to Logistic Regression. The paper lack the descriptive learning of all different classification algorithms.

Sethi, Kapil & Gupta, Ankit & Gupta, Gaurav & Jaiswal, Varun. (2019). Comparative Analysis of Machine Learning Algorithms on Different Datasets. They have prepared a good research but have focused particular on effect of different size of datasets instead of Different Algorithms.

Finally, B. Omar, B. Zineb, A. Cortés Jofré and D. González Cortés, "A Comparative Study of Machine Learning Algorithms for Financial Data Prediction," 2018 International Symposium on Advanced Electrical and Communication Technologies (ISAECT), Rabat, Morocco, 2018, pp. 1-5, doi: 10.1109/ISAECT.2018.8618774. Have proposed good comparative study but they focused more upon artificial neural networks and other algorithms, they lack the comparison of mostly used algorithms for classification.

III.METHODOLOGY

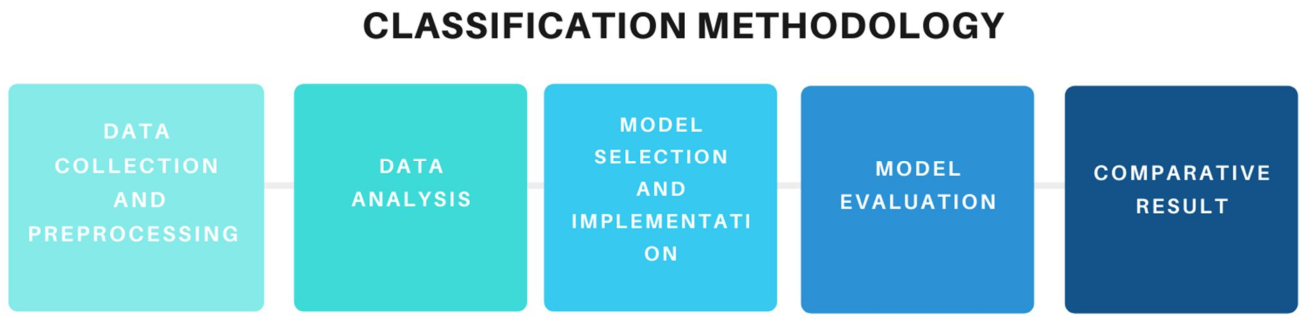

CLASSIFICATION METHODOLOGY
DATA COLLECTION AND PREPROCESSING
DATA ANALYSIS
MODEL SELECTION AND IMPLEMENTATI ON
MODEL EVALUATION
COMPARATIVE RESULT

Fig. 1: Machine learning pipeline for classification problem.

A. Data Collection and Pre-processing

Data collection is the process of gathering and measuring information on targeted variables in an established system, which then enables one to answer relevant questions and evaluate outcomes. [1]

The data of Mobile Price is available on Kaggle. (https://www.kaggle.com/iabhishekofficial/mobile-price-classification)

Data pre-processing is a data mining technique that involves transforming raw data into an understandable format.

Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviours or trends, and is likely to contain many errors. Data pre-processing is a proven method of resolving such issues. Data pre-processing prepares raw data for further processing. [2]

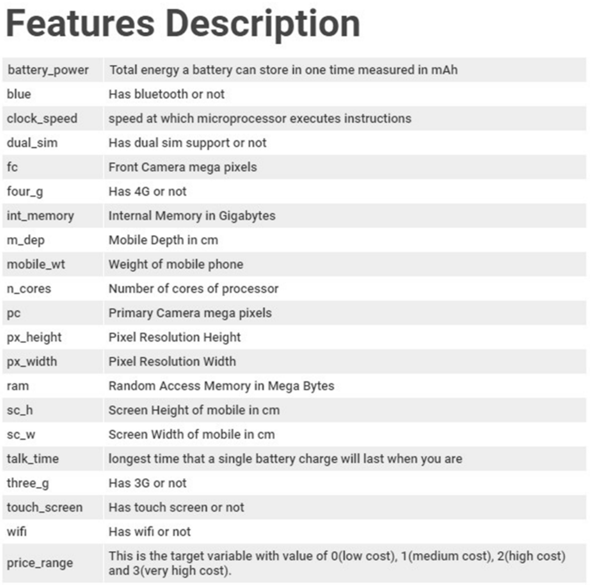
Given data has 21 feature columns of different 2000 instances. The given data has been cleaned and null values has been removed. Prior to the performing of the machine learning techniques, the data pre-processing was performed on the data set. Incomplete data such as data that which is lacking attribute values, missing values within the records were delete from the data set. Outlier analysis was performed. In WEKA a filter called Interquartile-Range was used to perform outlier analysis.

B. Data Analysis

Data Analysis is the process of systematically applying statistical and/or logical techniques to describe and illustrate, condense and recap, and evaluate data. [3]

Python's Scikit-Learn Machine Learning Toolbox has been used for the Exploratory Data Analysis, Data Processing and Model Development. Python's Plotting Libraries like Matplotlib and Seaborn have been used for the data Visualizations.

Dataset as 21 features and 2000 entries. The meanings of the features are given below.



Features Description

battery_power	Total energy a battery can store in one time measured in mAh
blue	Has bluetooth or not
clock_speed	speed at which microprocessor executes instructions
dual_sim	Has dual sim support or not
fc	Front Camera mega pixels
four_g	Has 4G or not
int_memory	Internal Memory in Gigabytes
m_dep	Mobile Depth in cm
mobile_wt	Weight of mobile phone
n_cores	Number of cores of processor
pc	Primary Camera mega pixels
px_height	Pixel Resolution Height
px_width	Pixel Resolution Width
ram	Random Access Memory in Mega Bytes
sc_h	Screen Height of mobile in cm
sc_w	Screen Width of mobile in cm
talk_time	longest time that a single battery charge will last when you are
three_g	Has 3G or not
touch_screen	Has touch screen or not
wifi	Has wifi or not
price_range	This is the target variable with value of 0(low cost), 1(medium cost), 2(high cost) and 3(very high cost).

Fig. 2: Feature Description of all the columns used.

C. Correlation Matrix Heat Map

A correlation matrix is a table showing correlation coefficients between variables. Each cell in the table shows the correlation between two variables. A correlation matrix is used to summarize data, as an input into a more advanced analysis, and as a diagnostic for advanced analyses. [4]

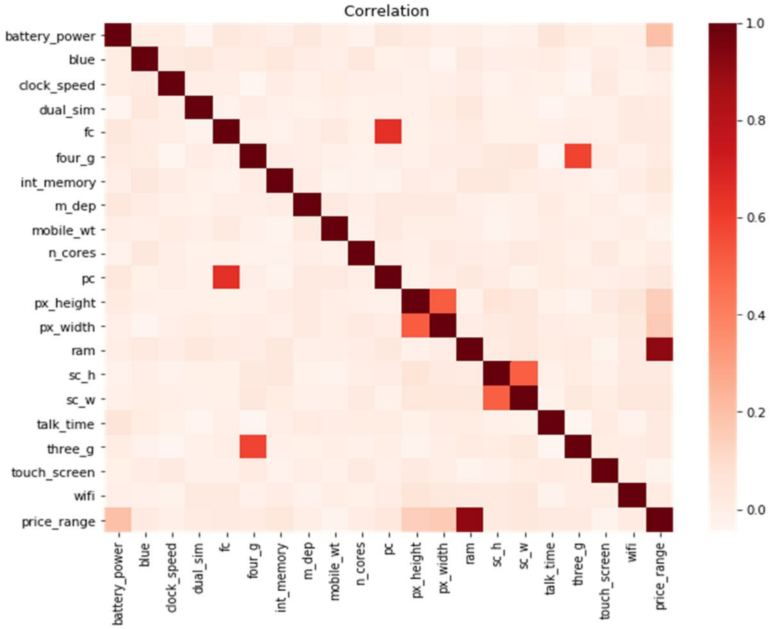


Fig. 3: Result of correlation matrix heat map plotting.

D. Model Selection and Implementation

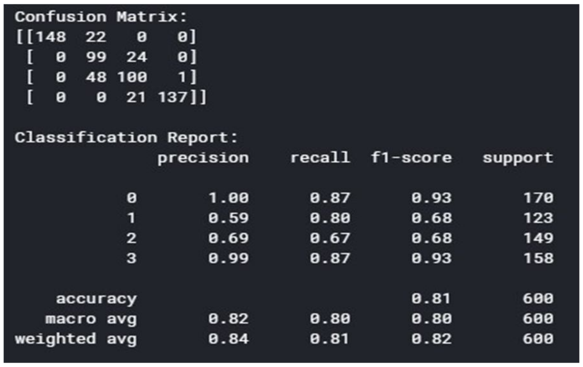
1) *Machine Learning:* Machine learning (ML) is categorized under artificial intelligence of (AI) which facilitates the computer with efficiency to perform and learn even after not being particularly programmed. ML is a strategy for information examination that robotizes logical model building [5-7]

2) *Classification Algorithms*

a) *Logistic Regression:* Logistic Regression is a supervised learning algorithm used for classification. The model created by this algorithm is based on logistic function [8] [9]. A logistic function, also referred to as a logistic curve is a sigmoid curve with the below equation is the Euler's number, x0 is the x-value of the sigmoid midpoint, M is the curve's maximum value and s is the steepness of the curve.

$$f\ (x)\ =\ \frac{M}{1\ +\ e^{-k(x-x0)}}$$

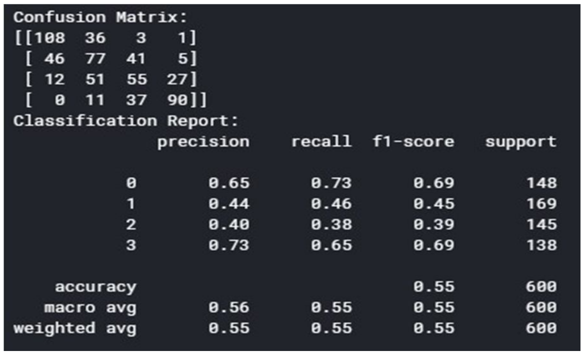
It is a logistic function to convert the output of a linear regression into classes. Higher linearity between the feature and the target variable contributes to better performance of the Logistic Regression model. In the multiclass case as ours, the training algorithm uses the one-vs-rest (OvR) scheme. The logistic regression class of scikit-learn implements regularized logistic regression. It can handle both dense and sparse input [10].



```
Confusion Matrix:
[[148  22   0   0]
 [  0  99  24   0]
 [  0  48 100   1]
 [  0   0  21 137]]

Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.87      0.93       170
           1       0.59      0.80      0.68       123
           2       0.69      0.67      0.68       149
           3       0.99      0.87      0.93       158

    accuracy                           0.81       600
   macro avg       0.82      0.80      0.80       600
weighted avg       0.84      0.81      0.82       600
```

Fig. 4: Confusion Matrix and Classification Report of Logistic Regression Algorithm.

b) *K – Nearest Neighbour (KNN):* KNN is a classification algorithm as given in [11] where objects are classified by voting several labelled training examples with their smallest distance from each object. This method performs well even in handling the classification tasks with multi-categorized classification. Its disadvantage is that KNN requires more time for classifying objects when a large number of training examples are given. KNN should select some of them by computing the distance of each test objects with all of the training examples. KNN is a modest algorithm that stores all accessible suitcases and classifies new suitcases based on a similarity measure. KNN has symmetrical names (a) Memory-Based reasoning [12] (b) Example-Based Reasoning (c) Instance-Based Learning (d) Case-Based Reasoning and (e) Lazy Learning. KNN utilized for relapse and grouping for prescient issues [13]. Be that as it may, it is broadly utilized as a part of grouping troubles in the business.



```
Confusion Matrix:
[[108  36   3   1]
 [ 46  77  41   5]
 [ 12  51  55  27]
 [  0  11  37  90]]
Classification Report:
              precision    recall  f1-score   support

           0       0.65      0.73      0.69       148
           1       0.44      0.46      0.45       169
           2       0.40      0.38      0.39       145
           3       0.73      0.65      0.69       138

    accuracy                           0.55       600
   macro avg       0.56      0.55      0.55       600
weighted avg       0.55      0.55      0.55       600
```

Fig. 5: Confusion Matrix and Classification Report of KNN Algorithm.

c) *Decision Tree:* A Decision Tree text classifier in [14] is a tree in which internal nodes are labelled by terms, branches departing from them are labelled by the weight that the term has in the text document and leafs are labelled by categories. Decision Tree constructs using ‘divide and conquer’ strategy. Each node in a tree is associated with set of cases. This strategy checks whether all the training examples have the same label and if not then select a term partitioning from the pooled classes of documents that have same values for term and place each such class in a separate subtree.

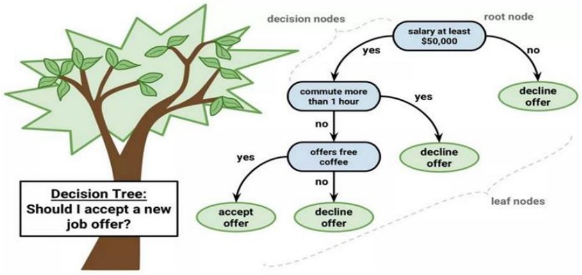
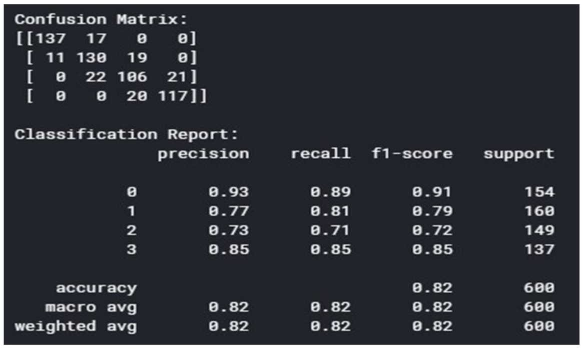


Fig. 6: Image to explain Decision Tree Algorithm.

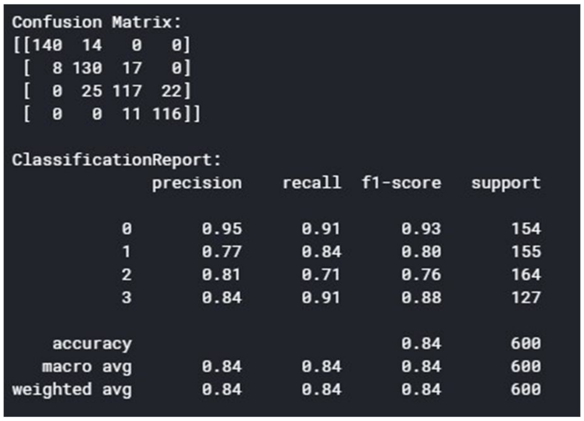
In contrast to other supervised learning algorithms, a decision tree algorithm can be utilized for taking care of regression and classification issues as well. The general thought process of utilizing Decision Tree is to make a training model that can use to predict class or estimation of objective factors by taking in choice standards derived from earlier data (training data). In Fig. 4 we have shown a sample picture of decision trees. [15]



```
Confusion Matrix:
[[137  17   0   0]
 [ 11 130  19   0]
 [  0  22 106  21]
 [  0   0  20 117]]

Classification Report:
              precision    recall  f1-score   support

           0       0.93      0.89      0.91       154
           1       0.77      0.81      0.79       160
           2       0.73      0.71      0.72       149
           3       0.85      0.85      0.85       137

    accuracy                           0.82       600
   macro avg       0.82      0.82      0.82       600
weighted avg       0.82      0.82      0.82       600
```

Fig. 7: Confusion Matrix and Classification Report of Decision Tree Algorithm.

d) *Support Vector Machine(SVM):* In ML, SVM are supervised learning models associated with learning algorithms that inspect data used for classification and regression analysis [16]. 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 [17]. When facts are not categorized, supervised learning is not possible, and an unsupervised learning approach is compulsory [18], which efforts to invention normal clustering of the data to groups, and then map new data to these formed groups. The clustering algorithm which delivers an enhancement to the SVM is called support vector clustering (SVC) and is often used in trade applications either when facts are not categorized or when only some facts are categorized as a pre-processing for a classification pass [19]. The mechanism of classifying the data into different classes by definition a line which splits the training files into classes. There are a few straight hyperplanes, SVM calculation tries to augment the separation in the focal of the few classes that are mind boggling and this is said as edge augmentation [20]. If the line makes the most of the space among the classes is recognized, the probability to simplify well to unobserved data is increased.



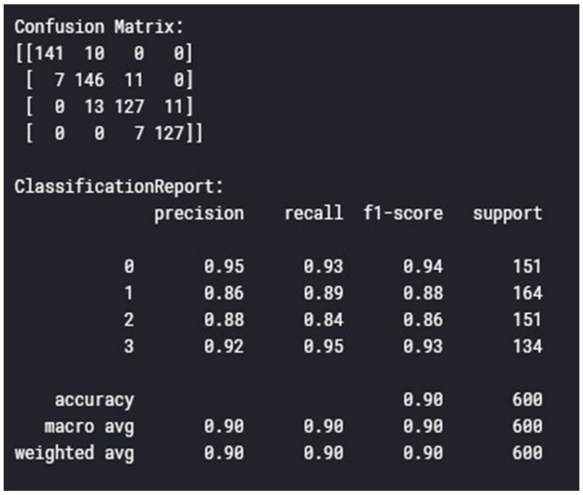
```
Confusion Matrix:
[[140  14   0   0]
 [  8 130  17   0]
 [  0  25 117  22]
 [  0   0  11 116]]

ClassificationReport:
              precision    recall  f1-score   support

           0       0.95      0.91      0.93       154
           1       0.77      0.84      0.80       155
           2       0.81      0.71      0.76       164
           3       0.84      0.91      0.88       127

    accuracy                           0.84       600
   macro avg       0.84      0.84      0.84       600
weighted avg       0.84      0.84      0.84       600
```

Fig. 8: Confusion Matrix and Classification Report of Support Vector Machine Algorithm.

e) *Gradient Boosting*

Gradient Boosting = Gradient Descent + Boosting

- Fit an additive model (ensemble) $\Sigma_t\ \rho_t h_t(x)$ in a forward stage-wise manner.
- In each stage, introduce a weak learner to compensate the shortcomings of existing weak learners.
- In Gradient Boosting, “shortcomings” are identified by gradients.
- Recall that, in Adaboost, “shortcomings” are identified by high-weight data points.
- Both high-weight data points and gradients tell us how to improve our model. [21]

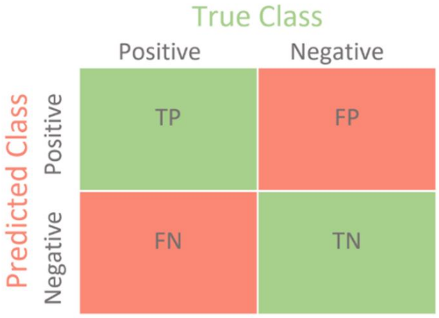


```
Confusion Matrix:
[[141  10   0   0]
 [  7 146  11   0]
 [  0  13 127  11]
 [  0   0   7 127]]

ClassificationReport:
              precision    recall  f1-score   support

           0       0.95      0.93      0.94       151
           1       0.86      0.89      0.88       164
           2       0.88      0.84      0.86       151
           3       0.92      0.95      0.93       134

    accuracy                           0.90       600
   macro avg       0.90      0.90      0.90       600
weighted avg       0.90      0.90      0.90       600
```

Fig. 9: Confusion Matrix and Classification Report of Gradient Boosting Algorithm.

E. Model Evaluation

A binary classification problem has only two classes to classify, preferably a positive and a negative class. Now let’s look at the metrics of the Confusion Matrix.



		True Class	
		Positive	Negative
Predicted Class	Positive	TP	FP
	Negative	FN	TN

Fig. 10: Confusion Matrix for Binary Classification.

1) *True Positive (TP):* It refers to the number of predictions where the classifier correctly predicts the positive class as positive.
2) *True Negative (TN):* It refers to the number of predictions where the classifier correctly predicts the negative class as negative.
3) *False Positive (FP):* It refers to the number of predictions where the classifier incorrectly predicts the negative class as positive.
4) *False Negative (FN):* It refers to the number of predictions where the classifier incorrectly predicts the positive class as negative.

a) *Accuracy:* It gives you the overall accuracy of the model, meaning the fraction of the total samples that were correctly classified by the classifier. To calculate accuracy, use the following formula:

$$\frac{TP + TN}{(TP + TN + FP + FN)}$$

b) *Misclassification Rate:* It tells you what fraction of predictions were incorrect. It is also known as Classification Error. You can calculate it using

$$\frac{FP + FN}{(TP + TN + FP + FN)}$$

c) *Precision:* It tells you what fraction of predictions as a positive class were actually positive. To calculate precision, use the following formula:

$$\frac{TP}{TP + FP}$$

d) *Recall:* It tells you what fraction of all positive samples were correctly predicted as positive by the classifier. It is also known as True Positive Rate (TPR), Sensitivity, and Probability of Detection. To calculate Recall, use the following formula:

$$\frac{TP}{TP + FN}$$

e) *Specificity:* It tells you what fraction of all negative samples are correctly predicted as negative by the classifier. It is also known as True Negative Rate (TNR). To calculate specificity, use the following formula:

$$\frac{TN}{(TN + FP)}$$

f) *F1-score:* It combines precision and recall into a single measure. Mathematically it’s the harmonic mean of precision and recall. It can be calculated as follows: [22]

$$F_1\text{−score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}$$

F. Classification Report

A Classification report is used to measure the quality of predictions from a classification algorithm. How many predictions are True and how many are False. More specifically, True Positives, False Positives, True negatives and False Negatives are used to predict the metrics of a classification report. [23]

IV. CONCLUSIONS

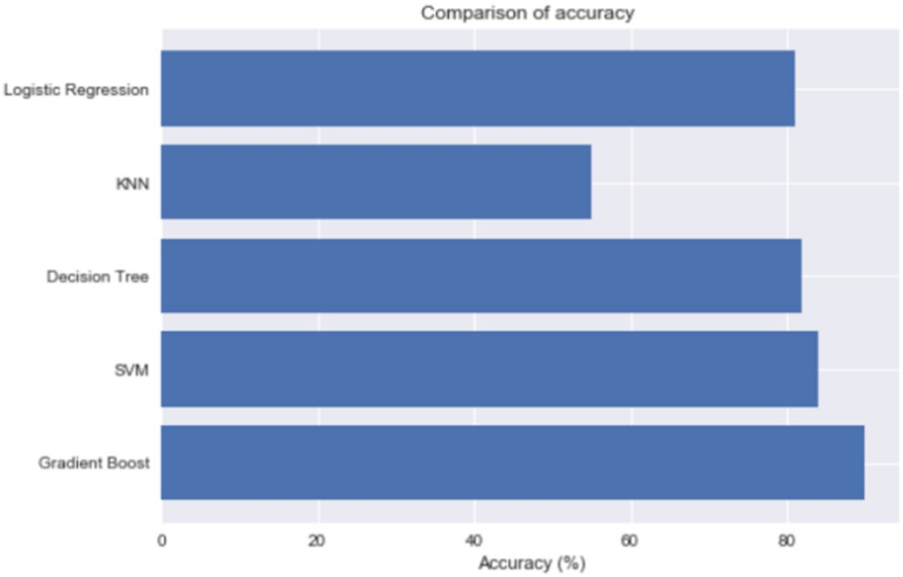


Fig. 11: Comparison of accuracy (F1-score) of different classifiers.

The principal part of this work is to compare five distinctive supervised machine learning classifiers and find the best accurate algorithm. We researched all classifiers execution on Mobile Price classification data and the Gradient Boost classifier gives the most elevated order exactness 90% dependent on F1 score and K-Nearest Neighbours (KNN) gives the least precision 55%. So we can conclude that even on the less training data, Gradient Boosting and SVM algorithms classifies very well and the accuracy can be increased by using big datasets. The main reason of low accuracy rate for some algorithms is low number of instances in the data set. In our study, there are a few bearings for future work in this field. We just explored some popular supervised machine learning algorithms, more algorithms can be picked to assemble an increasingly precise model.

V. ACKNOWLEDGMENT

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

REFERENCES

[1] Data is available on Kaggle uploaded by Abhishek Sharma : https://www.kaggle.com/iabhishekofficial/mobile-price-classification
[2] https://en.wikipedia.org/wiki/Data_collection
[3] https://www.techopedia.com/definition/14650/data-preprocessing
[4] https://ori.hhs.gov/education/products/n_illinois_u/datamanagement/datopic.html#:~:text=Data%20Analysis%20is%20the%20process,and%20recap%2C%20and%20evaluate%20data.&text=An%20essential%20component%20of%20ensuring,appropriate%20analysis%20of%20research%20findings.
[5] https://www.displayr.com/what-is-a-correlation-matrix/#:~:text=A%20correlation%20matrix%20is%20a,Create%20your%20own%20correlation%20matrix
[6] Sharma, L., Gupta, G. and Jaiswal, V., 2016, December. Classification and development of tool for heart diseases (MRI images) using machine learning. In Parallel, Distributed and Grid Computing (PDGC), 2016 Fourth International Conference on (pp. 219-224). IEEE.
[7] Chauhan, D. and Jaiswal, V., 2016, October. An efficient data mining classification approach for detecting lung cancer disease. In Communication and Electronics Systems (ICCES), International Conference on (pp. 1-8). IEEE.
[8] Negi, A. and Jaiswal, V., 2016, December. A first attempt to develop a diabetes prediction method based on different global datasets. In Parallel, Distributed and Grid Computing (PDGC), 2016 Fourth International Conference on (pp. 237-241). IEEE.
[9] https://hal.inria.fr/hal-00860051/document
[10] Aaron Defazio, Francis Bach, Simon Lacoste-Julien. SAGA: A Fast Incremental Gradient Method with Support for Non-Strongly Convex Composite Objectives. Advances In Neural Information Processing Systems, Nov 2014, Montreal, Canada
[11] http://scikit-learn.org/stable/documentation.html

[12] Tam, Santoso A and Setiono R., “A comparative study of centroid-based, neighborhood-based and statistical approaches for effective document categorization”, ICPR '02 Proceedings of the 16th International Conference on Pattern Recognition (ICPR'02) ,vol.4 , no. 4 , 2002, pp.235–238.

[13] Domingos, P., 2012. A few useful things to know about machine learning. Communications of the ACM, 55(10), pp.78-87.

[14] Mitchell, T.M., 2006. The discipline of machine learning (Vol. 3). Carnegie Mellon University, School of Computer Science, Machine Learning Department.

[15] Russell Greiner and Jonathan Schaffer, “Exploratorium – Decision Trees”, Canada. 2001. URL: http://www.cs.ualberta.ca/~aixplore/learning/ Decision Trees

[16] Decision Trees, Retrieve from: https://dataaspirant.com/2017/01/30/how-decision-treealgorithm-works/, Last Accessed: 5 Octobor,2019

[17] Wagstaff, K., 2012. Machine learning that matters. arXiv preprint arXiv:1206.4656.

[18] Bennett, K.P. and Parrado-Hernández, E., 2006. The interplay of optimization and machine learning research. Journal of Machine Learning Research, 7(Jul), pp.1265-1281.

[19] Caruana, R. and Niculescu-Mizil, A., 2006, June. An empirical comparison of supervised learning algorithms. In Proceedings of the 23rd international conference on Machine learning (pp. 161-168). ACM.

[20] Javidi, B., 2002. Image recognition and classification: algorithms, systems, and applications. CRC Press.

[21] College of Computer and Information Science Northeastern University[A Gentle Introduction to Gradient Boosting by Cheng Li]

[22] https://towardsdatascience.com/confusion-matrix-for-your-multi-class-machine-learning-model-ff9aa3bf7826

[23] https://muthu.co/understanding-the-classification-report-in-sklearn/#:~:text=A%20Classification%20report%20is%20used,classification%20report%20as%20shown%20below