A Comparison of Machine Learning Classifiers on Laptop Products Classification Task

Ayesha Ayub Syed, Yaya Heryadi, Lukas, Antoni Wibowo

***Abstract*— During the current pandemic situation, the laptop market is accelerating due to increased demands of 'work from home' and 'learn from home'. The expansion of the market and the increased volume of laptop products on e-commerce websites require effective and efficient products management and categorization. Better products categorization offers a smooth navigation and shopping experience to the customer. This research paper handles laptop products classification as a multiclass classification problem. It proposes a method to automatically classify laptop products into three categories, namely, 'Budget', 'Midrange', and 'Flagship' using machine learning classifiers. Various classifiers including Support Vector Machines, Multinomial Logistic Regression, Decision Trees, and Artificial Neural Network are used for the classification task. The classifiers are evaluated in terms of accuracy, recall, precision, and F1-score metrics. The results reveal an outstanding accuracy of 99% for SVM (Linear kernel), 98% for SVM (Gaussian kernel), Multinomial Logistic Regression, and Decision Tree classifier, 91% with the Artificial Neural Network, and 72% with SVM (Polynomial kernel) on our laptop products dataset.**

***Index Terms*—machine learning, multinomial logistic regression, multiclass classification, neural network, support vector machines**

I. Introduction

THE expansion of e-commerce businesses and the increased volume of products on e-commerce websites make product classification an intricate task. Product classification deals with the correct placement of the products in the relevant categories. It is a key feature for e-commerce websites that also facilitates marketing. Product categorization tends to increase conversion rates and return higher sales to the company. One of the major benefits of product classification is the improvement of website usability and navigation experience to the user. Users can look for the desired product quickly and easily. A high level of website usability influences user behavior in a positive way [1]. Users visit multiple websites before making the final purchase decision and are more likely to purchase from the website offering an optimal price and a better shopping experience [2].

For the automation of classification tasks, Machine Learning (ML) offers promising methods and algorithms. Classification in machine learning is a learning problem where a system learns to predict class labels on a set of data points. As a supervised learning problem, target class labels are also provided as an input to the classification algorithm. Classification can either be binary or multiclass. In binary classification, there are two classes to be predicted while multiclass classification problems involve predicting more than two classes. Some of the machine learning algorithms available for classification include Support Vector Machines (SVM), Decision Tree, Naïve Bayes, K-Nearest Neighbor, Multi-Layer Perceptron (MLP) [3]. The classification algorithms perform differently on different datasets. The performance of the classifier depends on the application, choice of features as well as nature of the dataset.

This research work focuses on the multiclass classification task for a laptop products dataset. Here, the term 'Laptop Products Classification' refers to the categorization of laptop products in three classes namely, 'Budget', 'Mid-range' and 'Flagship' products. The inputs to the classifier are features like laptop company, product type, size, weight, RAM, and price. The classifier predicts whether the given product belongs to 'Budget', 'Midrange' or, the 'Flagship' class. Various machine learning classifiers including Support Vector Machines (SVM), Multinomial Logistic Regression, Decision Tree, and Artificial Neural Network (ANN) are used to predict the class of laptop products. The performance of classifiers is compared in terms of classification accuracy, recall, precision, and F1-score.

The significance of this work is the better management of laptop products on an e-commerce website. From the user's point of view, it helps customers to find the required laptop product easily, efficiently, and according to their financial budget. It eliminates the need to scroll through hundreds of products to find the required one. For the business, it is beneficial because the smooth navigation and shopping experience are likely to bring the customer again to the website in the future, increasing the business sales. It also helps the business to manage the products more effectively resulting in increased productivity. The same research idea is also applicable to the categorization of smartphone products, tablets, and smartwatches, etc.

Section II of the paper covers the literature review and discusses some of the related work. Section III discusses the research methodology and experiment details. Section IV presents data visualization. Section V reveals the results of classification and discusses the results. Section VI presents the conclusion of the paper and provides directions for future work.

Ayesha Ayub Syed is a Ph.D. student at the Department of Doctor of Computer Science, Bina Nusantara University, Jakarta, Indonesia; (e-mail: ayeshaayubsyed@yahoo.com).

Yaya Heryadi, Faculty of the Department of Doctor of Computer Science, Bina Nusantara University, Jakarta, Indonesia; (email: yayaheryadi@binus.edu).

Lukas, Cognitive Engineering Research Group (CERG), Faculty of Engineering, Universitas Katolik Indonesia Atma Jaya, Indonesia; (email: lukas@atmajaya.ac.id).

Antoni Wibowo, Faculty of the Department of Computer Science, Bina Nusantara University, Jakarta, Indonesia; (email: anwibowo@binus.edu).

ISBN: 978-988-14049-1-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

II. Literature Review

Machine Learning (ML) is a field that is based on concepts and principles from multiple disciplines including Mathematics, Computer Science, Statistics, Cognitive Science, and Optimization Theory [4]. ML tasks are categorized into supervised learning, unsupervised learning, and reinforcement learning. From the supervised learning category, classification and regression are well-known tasks. In classification, the output is discrete e.g., class labels while output in regression takes on continuous values [3].

Researchers in [5] implemented the Multinomial Naïve Bayes algorithm for catalog classification. The products are categorized into classes like 'Electronics' and then subclasses like 'printer'. The number of products is about 40,000 collected from different databases including Amazon, Flipkart, etc. An overall number of 1000 classes for 40,000 products in the system. Multinomial Naïve Bayes is mostly applied for document classification. The foundation of the Naïve Bayes Classifier is the Bayes theorem based on probability. Naïve Bayes assumption about features is that the features are independent. With X as a feature vector of size n, $X = (x_1, x_2, x_3, \ldots . x_n)$ and y as the class label variable, Naïve Bayes predicts the class label as,

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)} \tag{1}$$

$$P(y|x_1, x_2, \ldots . x_n) = \frac{P(x_1|y)P(x_2|y)P(x_3|y), \ldots . P(x_n|y)}{P(x_1)P(x_2)P(x_3) \ldots . P(x_n)} \tag{2}$$

$$y = argmax_y P(y) \prod_{i=1}^{n} P(x_i|y) \tag{3}$$

The work of [6] is based on the use of the Naïve Bayes classifier and the Decision Tree classifier to predict the classes of mobile phones with given features as 'Economical' or 'Expensive'. The researchers collected the dataset from GSMArena.com. The features collected in the dataset include display size, weight, thickness, internal memory, camera, video quality, RAM, and battery. Two feature selection algorithms InfoGainEval and WrapperattributEval were applied to select the features that are most important in predicting the output class. The results are compared across the classifiers in terms of accuracy achieved with the selection of minimum features.

The Decision Tree algorithm is a popular supervised learning algorithm that works well with classification and regression tasks. The algorithm models a tree-like flowchart that has a root node, decision nodes, branches, and leaf nodes. The algorithm divides the data into small parts to identify the patterns that can be used for making a prediction. The learning strategy behind decision trees is the divide and conquer strategy. The entire dataset is at the root node. The algorithm chooses the feature that best predicts the target class. The entries are divided into groups of feature values. This decision creates the first set of branches. The divide and conquer process continues on the nodes until a stop criterion is reached [7]. The popular decision tree algorithms include ID3, C4.5, and CART algorithm [8].

[9] worked on product categorization on a dataset collected from Amazon distributers, using machine learning classifiers including Naïve Bayes, K-Nearest Neighbor, and Tree Classifier. Features for each item were determined using the bag-of-words model. The features set was processed using the standard pre-processing techniques like stop word removal, punctuation and number removal, lowercasing, and lemmatization. After feature processing, feature importance was determined using a modified MI formula and finally, the features were selected using the forward and backward search strategies. Naïve Bayes finished with 76.9% accuracy, KNN resulted in 69.4% accuracy, and the tree classifier performed the best with 86% accuracy but the execution took a long time (8 hours) to complete as compared to Naïve Bayes (3 seconds) and KNN (4 minutes).

Support Vector Machine (SVM) is another powerful machine learning algorithm for solving classification and regression problems. It has been reported to have outperformed other supervised machine learning algorithms and has become quite popular for classification in recent years due to its good generalization ability [10]. SVMs are focused on finding a hyperplane in an n-dimensional feature space that separates/classifies the data points. The algorithm chooses the hyperplane with the maximum margin so that the future data points can be classified more accurately. The points closest to the hyperplane are the support vectors and these vectors help in maximizing the margin of the hyperplane. To classify data that is not linearly separable, SVM has a technique known as the kernel trick. The kernel function takes a low dimensional input feature space and transforms it to a higher dimensional space. During this process, several complex data transformations take place to classify the data based on the output labels provided.

[11] implemented the KNN algorithm using several distance measures like Euclidean distance, Manhattan distance, and Chebyshev distance. The dataset utilized by the researchers is the KDD dataset. It is Knowledge Discovery and Data mining dataset with 41 features and class labels as 'normal' or 'attack'. The dataset is used in Intrusion Detection System (IDS). The results were evaluated using accuracy value, sensitivity and specificity measures, and FPR (False Positive Rate) and FNR (False Negative Rate). The results demonstrated the performance of Manhattan distance to be superior as compared to other distance metrics on the KDD dataset.

The researchers in [12] used Artificial Neural Networks (ANN) for automation of the classification of water quality. The dataset was obtained from the laboratory measurements and included environmental factors like pH, chemical oxygen demand, biological oxygen demand, dissolved oxygen, total suspended solids, and ammonia. The classification accuracy of 80% with an RMSE value of 0.468 was reported.

[13] used ANN as well as a hybrid ANN-Bat Optimization Algorithm for the classification of medical diagnosis. The results indicated that the proposed ANN combined with the metaheuristic Bat algorithm performed better in terms of accuracy.

ANN is suitable for predictive modeling tasks when the number of output classes is large and there is a large amount of data supporting the model. ANN uses neurons as a computational unit. An ANN consists of an input layer, hidden layer/layers, and an output layer. The neurons in each layer are connected to all other neurons in the following layer. A weight is associated with each of the connections. With $x_1, x_2, x_3, \ldots . x_i$ as input, w_{ij} as the weight associated with each input, s_j as the weighted sum at node j, and h_j as the output, the network can be mathematically expressed as,

$$s_j = \sum_{i=0}^{n} w_{ij} x_i \tag{4}$$

$$h_j = \tanh(s_j) = \frac{e^{2s_j} - 1}{e^{2s_j} + 1} \tag{5}$$

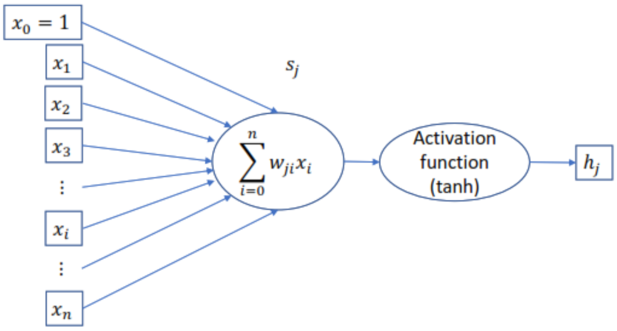

Fig. 1. A single node of a neural network

III. Methodology And Experiment Details

The methodology for the experiment conducted is presented in Fig. 2. The dataset for this work is downloaded from AtapData (https://atapdata.ai/). The dataset has 1,304 laptop products with detailed specs contained in 12 columns. The names of the columns are shown in Table I.

A. Data Preparation

To achieve the objective of classifying laptop products into 'Budget', 'Midrange', and 'Flagship' categories, we have added one more column to the dataset named 'Class'. Depending on the price and specs provided in the dataset and some research on laptops prices in general, we assigned a class to each entry in the dataset. The class assignment rule is if the price of the laptop product is in the range of €100 - €599, we assigned it the class 'Budget', if the price is between €600 - €999, we assigned it the class 'Midrange' and, if the price is between €1000 - €6000, the 'Flagship' class is assigned.

TABLE I
Columns in the Original Dataset

No.	Column
1	Company
2	Product
3	TypeName
4	Inches
5	ScreenResolution
6	CPU
7	RAM
8	Memory
9	GPU
10	OpSys
11	Weight
12	Price_euros

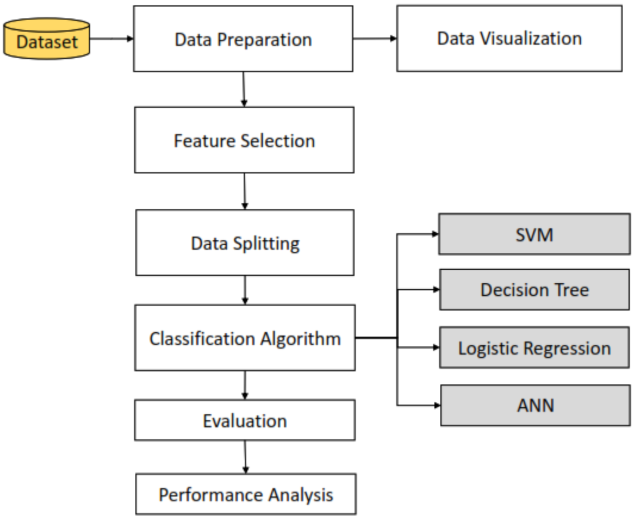

Fig. 2. Research Methodology

B. Feature Selection

The next and most important step is selecting the features or attributes from the dataset for the classification task. From the 12 features in the dataset, we have chosen 6 features. The selected features are numerical and categorical variables. These are 'Company', 'Product', 'Inches', 'Weight', 'RAM', and 'Price'. The features having the categorical data e.g., 'Company' and 'Product' are assigned a numerical value for each category.

C. Data Splitting

In this phase, the data is split into two sets namely the training set and testing set. 80% of the data is used for training or fitting the model while the rest 20% of the data is spared for testing and evaluating the classification performance.

D. Classification Algorithm

We fit our model on four machine learning classifiers e.g., SVM, ANN, Decision Tree, and Multinomial Logistic Regression. For the SVM classifier, we experimented using the Linear Kernel, the Gaussian Kernel, as well as, the Polynomial Kernel.

E. Evaluation

We evaluated the classification model on the test dataset in terms of accuracy, precision, recall, and f-score. We plotted confusion matrices for clear visualization of True Positives, False Positives, True Negatives, and False Negatives.

F. Performance Analysis

We finally analyzed the performance of all four classifiers to determine which classifier performed best on our laptop products dataset.

IV. Data Visualization

This Section presents a visualization of some of the data from the dataset to know some interesting trends in the laptops market, using bar charts, histograms, word clouds, and graphs.

Fig. 3 presents a bar chart of the laptop market players in the dataset. We find Dell, Lenovo, and HP to be the dominant market players having the largest number of products. Fig. 4 presents the horizontal bar graph showing the laptop products in the dataset. Among the 6 categories, ‘Notebooks’ are found to be the common trend in laptop products.

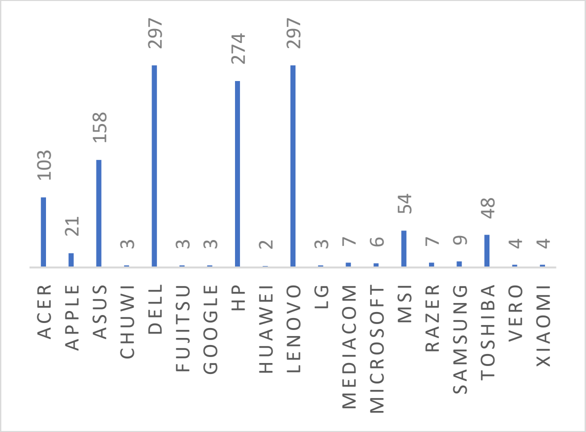

Fig. 3. Laptop Market Players in the dataset

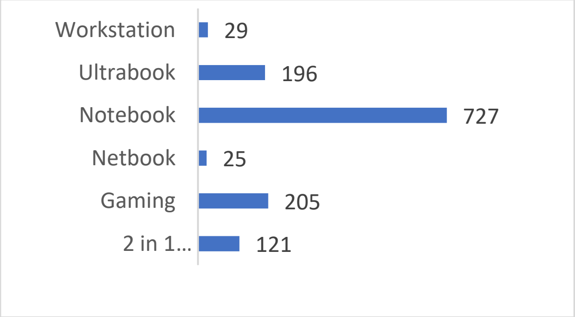

Fig. 4. Laptop Products in the dataset

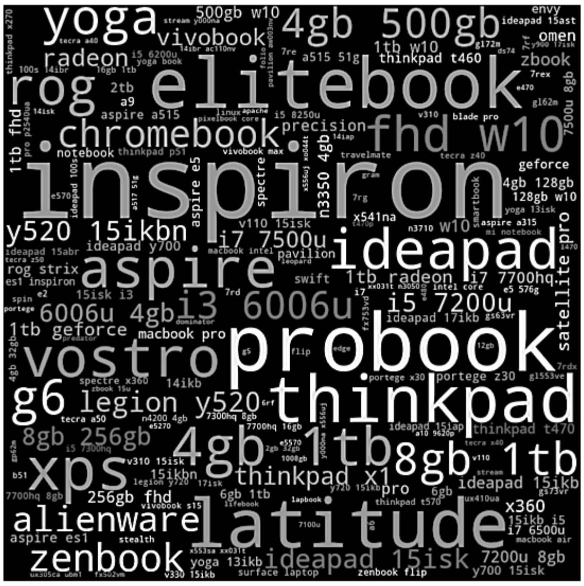

Fig. 5. Word Cloud Laptop Product types in the dataset

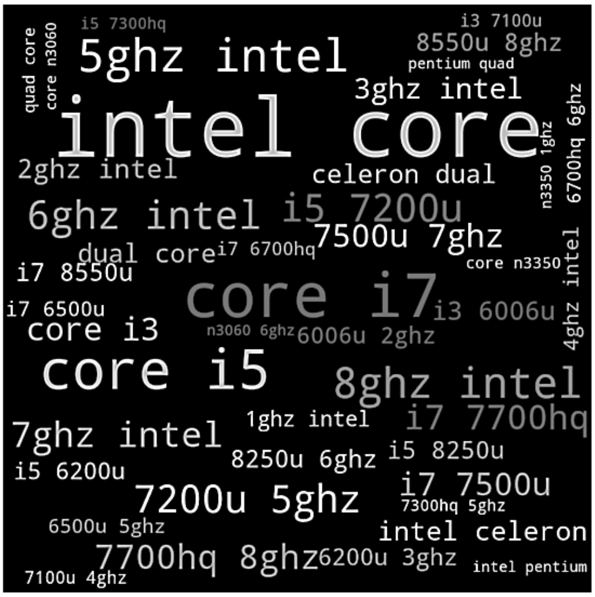

Fig. 6. Word Cloud of CPU in the dataset

Fig. 5 presents the word cloud generated from the ‘Product type’ column in the dataset. It highlights ‘Inspiron’, ‘Probook’, ‘EliteBook’, ‘ThinkPad’, and ‘Latitude’ as some of the best laptop products. The word cloud in Fig. 6 highlights some of the top processors as ‘Intel core’, ‘Core i5’, and ‘Core i7’. Fig. 7 presents a pair plot of the dataset features used for building the classification model. It reflects the pairwise relationships between the attributes ‘Company’, ‘Product’, ‘Inches’, ‘RAM’, ‘Weight’, ‘Price’ on all three classes ‘flagship’, ‘midrange’, and ‘entry-level/budget’. It can be observed from Fig. 7, that ‘Price’ is the most important feature that determines the class of a laptop as ‘budget’, ‘midrange’, or ‘flagship’. Some of the plots in the pair plot reflect that the classes are linearly separable (last row and last column of the pair plot) while in some of the other plots there is quite a lot of overlap.

V. Classification Results, Evaluation & Discussion

This Section presents the testing results of our classification model using four different classifiers. We have used four metrics for performance evaluation: Accuracy, Precision, Recall, and F-score. The confusion matrices are plotted for clear visualization of the results. The metrics are defined as,

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \tag{6}$$

$$Precision = \frac{True\ positive\ predictions}{Total\ positive\ predictions} \tag{7}$$

$$Recall = \frac{True\ positive\ predictions}{Total\ actual\ positive\ predictions} \tag{8}$$

$$F1\ score = 2\left(\frac{precision\ x\ recall}{precision + recall}\right) \tag{9}$$

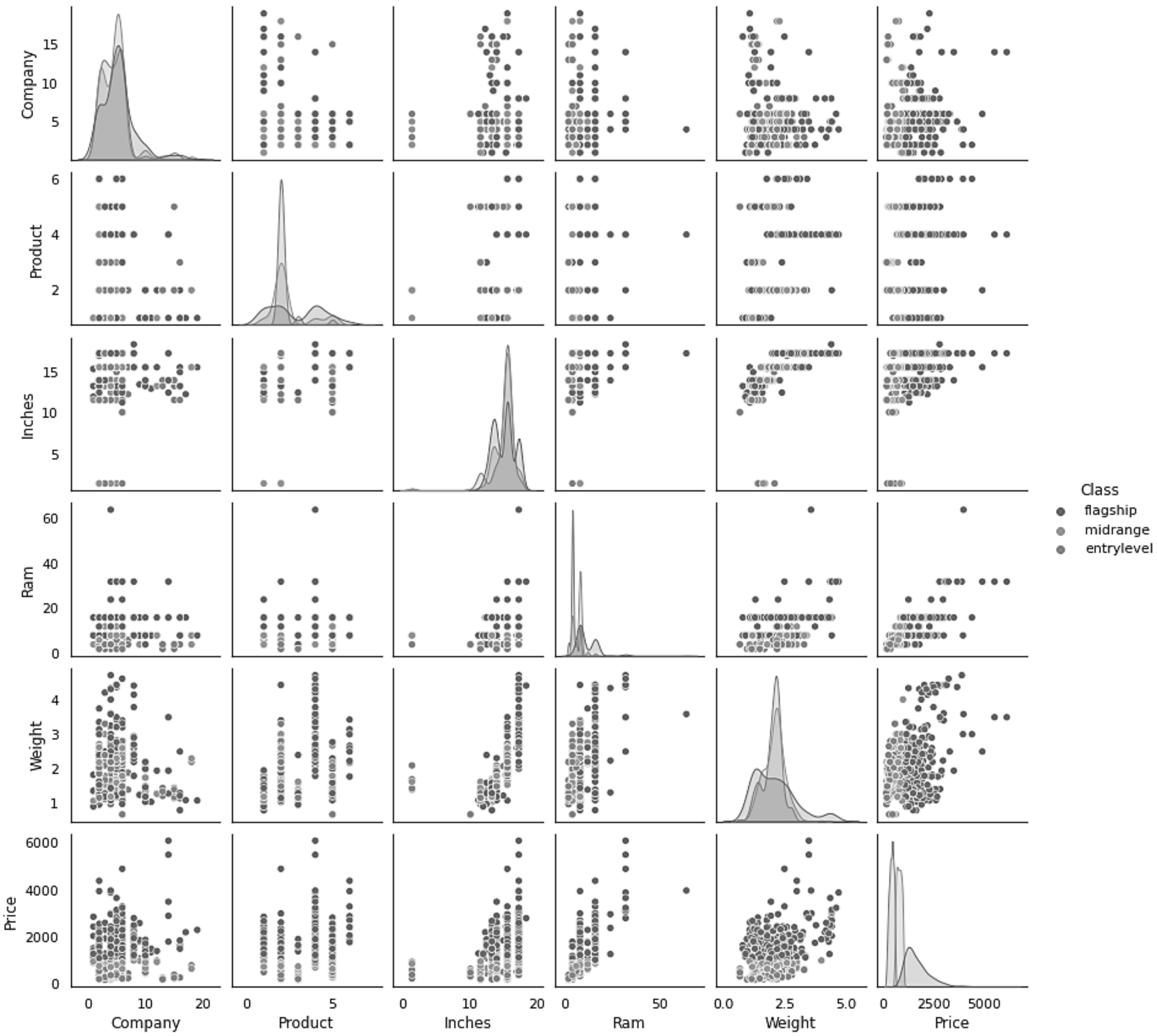

Fig. 7. Pair plot of the selected features from the dataset

Table II presents classification results using the SVM (Linear Kernel). The model has achieved 99% accuracy using the SVM linear classifier. Precision and recall scores for each class are also presented. Fig. 8 shows the corresponding confusion matrix. It is observed that overall there are two misclassifications, one for the 'Midrange' class (Class 1) and the second for the 'Flagship' class (Class 2).

TABLE II
CLASSIFICATION REPORT FOR SVM LINEAR KERNEL

	Precision	Recall	F1-score
Class 0 - Budget	0.98	1.00	0.99
Class 1 - Midrange	0.99	0.99	0.99
Class 2 - Flagship	1.00	0.98	0.99
Accuracy			0.99
Macro Average	0.99	0.99	0.99
Weighted Average	0.99	0.99	0.99

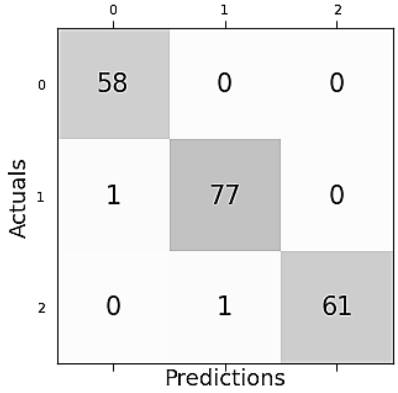

Fig. 8. Confusion Matrix for SVM Linear Kernel Classifier

Table III presents the classification results using SVM (Gaussian Kernel). The model achieved 98% accuracy and 0.98 as the average recall, precision, and F1-score. Fig. 9 presents the corresponding confusion matrix. It is observed that in this case, the total number of misclassifications is 4.

TABLE III
CLASSIFICATION REPORT FOR SVM GAUSSIAN KERNEL

	Precision	Recall	F1-score
Class 0 - Budget	0.97	1.00	0.98
Class 1 - Midrange	0.97	0.97	0.97
Class 2 - Flagship	1.00	0.97	0.98
Accuracy			0.98
Macro Average	0.98	0.98	0.98
Weighted Average	0.98	0.98	0.98

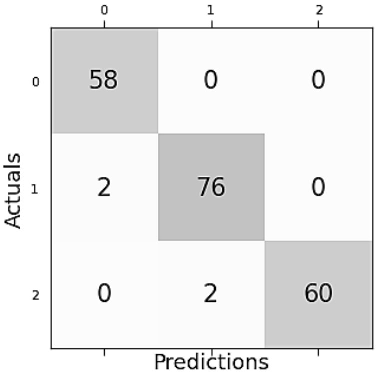

Fig. 9. Confusion Matrix for SVM Gaussian Kernel Classifier

Table IV indicates the classification results of the SVM (Polynomial Kernel). An accuracy of 72% and an average precision, recall, and F1-score of 0.83, 0.72, and 0.70 are observed. The confusion matrix in Fig. 10 reflects 55 incorrect predictions. Using the SVM classifier, the best classification performance is achieved with SVM 'linear kernel'.

TABLE IV
CLASSIFICATION REPORT FOR SVM POLYNOMIAL KERNEL

	Precision	Recall	F1-score
Class 0 - Budget	0.52	1.00	0.69
Class 1 - Midrange	0.93	0.32	0.48
Class 2 - Flagship	1.00	0.97	0.98
Accuracy			0.72
Macro Average	0.82	0.76	0.72
Weighted Average	0.83	0.72	0.70

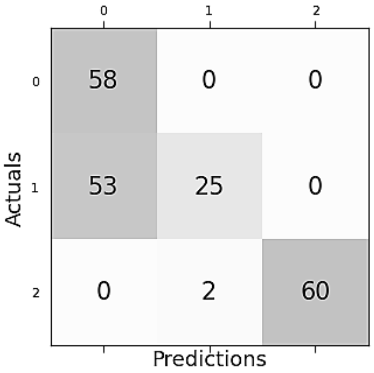

Fig. 10. Confusion Matrix for SVM Polynomial Kernel Classifier

Laptop products classification is also implemented using the simple neural network in Keras. Since we have six features, the number of inputs to the neural network is 6, there are three classes 'Budget', 'Midrange', and 'Flagship' to be predicted, so the number of outputs is specified to be 3. 'RELU' activations are used for the input and hidden layer. 'Softmax' activation is used as the output. The model compiles with the Adam optimizer and the cross-entropy loss function. With 192 parameters and 200 epochs, the model achieved a final prediction accuracy of 91%. Fig. 11 and Fig. 12 show the history plots of the classification model accuracy and loss. The training and testing accuracy and loss values are presented in Table V.

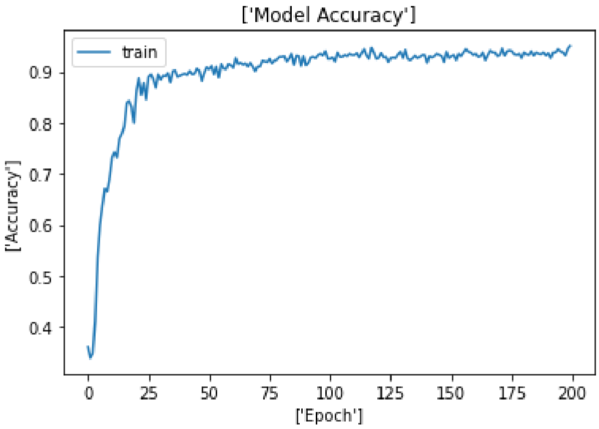

Fig. 11. Neural Network Model Accuracy

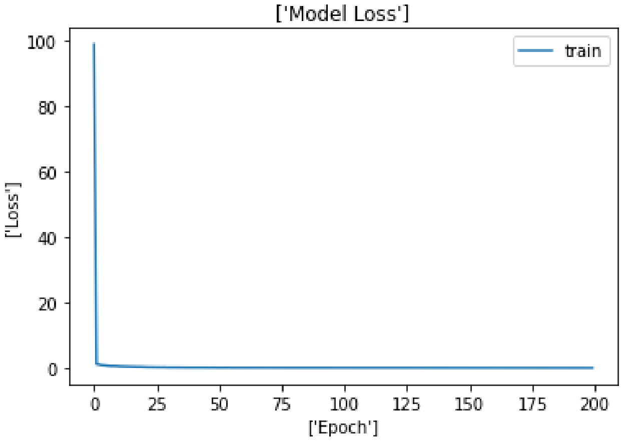

Fig. 12. Neural Network Model Loss

TABLE V
CLASSIFICATION REPORT FOR NEURAL NETWORK

	Training	Testing
Accuracy	0.9506	0.9091
Loss	0.1513	0.1623

Table VI and Table VII show the classification reports for the Decision Tree and the Multinomial Logistic Regression classifier. Both the classifiers predicted with an accuracy of 98% and overall, 2 to 3 incorrect predictions. Confusion matrices for both the classifiers are presented in Fig. 13 and Fig. 14.

TABLE VI
CLASSIFICATION REPORT FOR DECISION TREE CLASSIFIER

	Precision	Recall	F1-score
Class 0 - Budget	0.98	0.97	0.98
Class 1 - Midrange	0.97	0.99	0.98
Class 2 - Flagship	1.00	1.00	1.00
Accuracy			0.98
Macro Average	0.99	0.98	0.99
Weighted Average	0.98	0.98	0.98

On comparing the performance of all the classifiers in Table VIII, we can observe that SVM 'linear kernel' gave the best accuracy of 99% on our Laptop Products dataset. SVM Gaussian kernel, Decision tree, and multinomial logistic regression also performed well with an accuracy of 98%. From these observations, it is not recommended to use SVM 'polynomial kernel' on this type of dataset. With the ANN, the prediction accuracy is 91% that is also encouraging. But the dataset should be large to get an overall improvement in prediction accuracy.

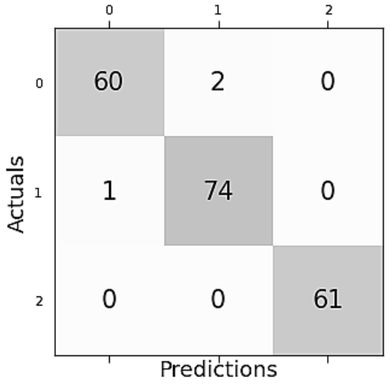

Fig. 13. Confusion Matrix for Decision Tree Classifier

TABLE VII
CLASSIFICATION REPORT FOR LOGISTIC REGRESSION CLASSIFIER

	Precision	Recall	F1-score
Class 0 - Budget	0.98	1.00	0.99
Class 1 - Midrange	1.00	0.97	0.99
Class 2 - Flagship	0.98	1.00	0.99
Accuracy			0.98
Macro Average	0.99	0.99	0.99
Weighted Average	0.99	0.99	0.99

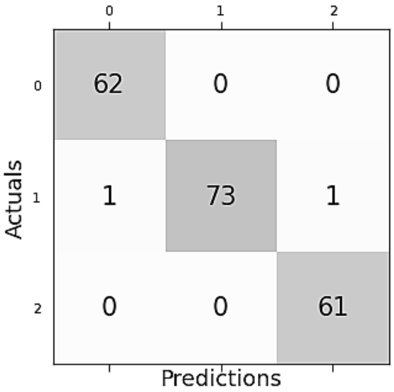

Fig. 14. Confusion Matrix for Logistic Regression Classifier

TABLE VIII
OVERALL CLASSIFICATION PERFORMANCE

Classifier	Accuracy
SVM Linear Kernel	**0.99**
SVM Gaussian Kernel	0.98
Decision Tree	0.98
Logistic Regression	0.98
Neural Network	0.91
SVM Polynomial Kernel	0.72

VI. CONCLUSION

The paper proposes a multiclass categorization of laptop products in three classes 'Budget', 'Midrange', and 'Flagship'. It compares the performance of SVM, MLR, DT, and ANN algorithms on a laptop products dataset. The features selected from the dataset for classification are Company, Product, Size, RAM, Weight, and Price. SVM (Linear kernel) performs best on the classification task. SVM (Gaussian kernel), MLR, and DT also performed well.

The work can be extended to develop a system for predicting the price of laptop products. The input to the system might be laptop specs and class from the dataset and the output would be the predicted price.

REFERENCES

[1] M. Sinha, L. N. Fukey, and S. Likitha, "Web user Experience and Consumer behavior: The Influence of Color, Usability, and Aesthetics on the Consumer Buying Behavior," Test Engineering and Management, vol. 82, pp. 16592-16600, 2020

[2] A. A. Syed and J. Suroso, "Factors Affecting Consumers' Decision for E-Hotel Booking," CommIT (Communication and Information Technology Journal), vol. 12, no. 2, pp. 111-123, 2018, DOI: 10.21512/commit.v12i2.4917

[3] A. Soofi and A. Awan, "Classification Techniques in Machine Learning: Applications and Issues," Journal of Basic and Applied Sciences, vol. 13, pp. 459–465, 2017, DOI: 10.6000/1927-5129.2017.13.76

[4] E. Alpaydın, "Introduction to Machine Learning, 2nd ed., The MIT Press, ISBN: 9780262012430, 2009

[5] S. Pandey, M. Supriya, and A. Shrivastava, "Data Classification Using Machine Learning Approach," Intelligent Systems Technologies and Applications, vol. 683, 2018, DOI: https://doi.org/10.1007/978-3-319-68385-0_10

[6] M. Asim and Z. Khan, "Mobile Price Class prediction using Machine Learning Techniques," International Journal of Computer Applications, vol. 179, no. 29, pp. 6–11, 2018, DOI: 10.5120/ijca2018916555

[7] B. Lantz, "Divide and Conquer – Classification Using Decision Trees and Rules," Machine Learning with R, 2nd ed., 2015

[8] H. Sharma and S. Kumar, "A Survey on Decision Tree Algorithms of Classification in Data Mining," International Journal of Science and Research, vol. 5, no. 4, pp. 2094–2097, 2016, doi: 10.21275/v5i4.nov162954

[9] S. Shankar and I. Lin, "Applying Machine Learning to Product Categorization," Computer Science Project, Stanford University, 2011.

[10] J. Cervantes, F. Garcia-Lamont, L. Rodrı́guez-Mazahua, and A. Lopez, "A comprehensive survey on support vector machine classification: Applications, challenges, and trends," Neurocomputing, vol. 408, pp. 189-215, 2020, DOI: 10.1016/j.neucom.2019.10.118

[11] P. Mulak and N. Talhar, "Analysis of Distance Measures Using K-Nearest Neighbor Algorithm on KDD Dataset," International Journal of Science and Research., vol. 4, pp. 2319–7064, 2013, Available: www.ijsr.net

[12] K. Sulaiman, L. Hakim Ismail, M. Adib Mohammad Razi, M. Shalahuddin Adnan, and R. Ghazali, "Water Quality Classification Using an Artificial Neural Network (ANN)," IOP Conference Series: Materials Science and Engineering, vol. 601, no. 1, 2019, doi: 10.1088/1757-899X/601/1/012005

[13] N. Kumar and D. Kumar, "Classification using Artificial Neural Network Optimized with Bat Algorithm," International Journal of Innovative Technology and Exploring Engineering, vol. 9, no. 3, pp. 696–700, 2020, doi: 10.35940/ijitee.c8378.019320