iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** VII **Month of publication:** July 2023

DOI: https://doi.org/10.22214/ijraset.2023.54914

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VII Jul 2023- Available at www.ijraset.com

Ad Click Prediction: A Comparative Evaluation of Logistic Regression and Performance Metrics

Niharika Namdev[1], Nandini Tomar[2]
[1]*Asst. Professor, CSE Department, IIMT College of Engineering, Greater Noida*
[2]*Asst. Professor, CSE Department, Dronacharya Group of Institution, Greater Noida*

Abstract: ***This research paper investigates the effectiveness of logistic regression as a predictive modelling technique for ad click prediction. The study aims to explore the performance of logistic regression and evaluate its predictive power using various evaluation metrics. The primary objective is to assess the accuracy, precision, recall, F1 score, and area under the receiver operating characteristic curve (AUC-ROC) of logistic regression models in predicting ad clicks. The paper begins with a comprehensive review of the literature on ad click prediction and logistic regression. It highlights the importance of accurate click-through rate (CTR) estimation for advertisers and the potential benefits of logistic regression in this context. The theoretical background of logistic regression is also discussed, providing an understanding of its underlying principles and assumptions. Next, the methodology section describes the dataset used for the study, which includes historical ad impressions and click data. The data pre-processing steps, including feature selection and transformation, are explained. Logistic regression models are then trained on the pre-processed data, and the model performance is evaluated using various evaluation metrics.***
The results section presents the findings of the study. It includes a detailed analysis of the accuracy, precision, recall, F1 score, and AUC-ROC obtained from the logistic regression models. The performance of the models is compared against benchmark models or alternative algorithms commonly used in ad click prediction. The results highlight the strengths and limitations of logistic regression in predicting ad clicks. Furthermore, the discussion section provides insights into the implications of the results. It discusses the interpretability of logistic regression models and their potential for providing actionable insights to advertisers. The limitations and potential challenges of logistic regression in ad click prediction are also addressed.
Finally, the conclusion section summarizes the key findings of the research and provides recommendations for future studies. It emphasizes the significance of logistic regression as a reliable and interpretable method for ad click prediction, while also recognizing the need for further research to improve its performance.
Keywords: Ad clicks prediction, logistic regression, evaluation metrics, click-through rate (CTR), and performance analysis.

I. INTRODUCTION

Ad click prediction plays a vital role in online advertising, where businesses invest significant resources to reach their target audience and generate revenue. The ability to accurately predict ad clicks helps advertisers optimize their ad campaigns, allocate budgets effectively, and improve return on investment (ROI). By understanding the likelihood of users clicking on specific ads, advertisers can make informed decisions about ad placement, ad content, targeting strategies, and bidding strategies.
Logistic regression, as a widely used statistical modeling technique, plays a significant role in ad click prediction. It provides a framework for estimating the probability of ad clicks based on various features associated with ads, users, and contextual information. Logistic regression models are well-suited for binary classification problems, where the outcome of interest (in this case, ad click or no click) is represented as a binary variable.

The role of logistic regression in ad click prediction can be summarized as follows:

1) *Probability Estimation:* Logistic regression models estimate the probability of ad clicks based on the provided features. By fitting a logistic regression model to historical ad click data, advertisers can obtain probability estimates for new ad impressions. These probabilities can be used to rank ads or determine bidding strategies, allowing advertisers to allocate their resources effectively and maximize the likelihood of ad clicks.
2) *Feature Importance:* Logistic regression provides insights into the importance of different features in predicting ad clicks. By examining the estimated coefficients of the logistic regression model, advertisers can identify which features have the most significant impact on the likelihood of ad clicks. This information helps advertisers understand the factors that drive user engagement and tailor their ad campaigns accordingly.

3) *Interpretability:* Logistic regression models offer interpretability, making them valuable in the advertising domain. Advertisers can understand how each feature contributes to the probability of ad clicks based on the logistic regression coefficients. This interpretability enables advertisers to make informed decisions about ad targeting, ad design, and content optimization, as they can identify the specific features that are most influential in driving ad engagement.
4) *Model Performance Evaluation:* Logistic regression provides a reliable framework for evaluating the performance of ad click prediction models. Metrics such as accuracy, precision, recall, and area under the receiver operating characteristic curve (AUC-ROC) can be computed to assess the effectiveness of the logistic regression model in predicting ad clicks. This evaluation helps advertisers compare different models, select the most appropriate one, and make data-driven decisions to optimize their ad campaigns.
5) *Scalability and Efficiency:* Logistic regression is computationally efficient and scalable, making it suitable for handling large-scale ad click prediction tasks. With the vast amount of ad impressions and user interactions happening in real-time, logistic regression allows for quick model training and prediction, enabling advertisers to make timely decisions and respond to changing market dynamics effectively.

II. RELATED WORK

Ad click prediction plays a vital role in online advertising, where businesses invest significant resources to reach their target audience and generate revenue.

The ability to accurately predict ad clicks helps advertisers optimize their ad campaigns, allocate budgets effectively, and improve return on investment (ROI). By understanding the likelihood of users clicking on specific ads, advertisers can make informed decisions about ad placement, ad content, targeting strategies, and bidding strategies.

Logistic regression, as a widely used statistical modeling technique, plays a significant role in ad click prediction. It provides a framework for estimating the probability of ad clicks based on various features associated with ads, users, and contextual information.

Logistic regression models are well-suited for binary classification problems, where the outcome of interest (in this case, ad click or no click) is represented as a binary variable.

The role of logistic regression in ad click prediction can be summarized as follows:

1) *Probability Estimation:* Logistic regression models estimate the probability of ad clicks based on the provided features. By fitting a logistic regression model to historical ad click data, advertisers can obtain probability estimates for new ad impressions. These probabilities can be used to rank ads or determine bidding strategies, allowing advertisers to allocate their resources effectively and maximize the likelihood of ad clicks.
2) *Feature Importance:* Logistic regression provides insights into the importance of different features in predicting ad clicks. By examining the estimated coefficients of the logistic regression model, advertisers can identify which features have the most significant impact on the likelihood of ad clicks. This information helps advertisers understand the factors that drive user engagement and tailor their ad campaigns accordingly.
3) *Interpretability:* Logistic regression models offer interpretability, making them valuable in the advertising domain. Advertisers can understand how each feature contributes to the probability of ad clicks based on the logistic regression coefficients. This interpretability enables advertisers to make informed decisions about ad targeting, ad design, and content optimization, as they can identify the specific features that are most influential in driving ad engagement.
4) *Model Performance Evaluation:* Logistic regression provides a reliable framework for evaluating the performance of ad click prediction models. Metrics such as accuracy, precision, recall, and area under the receiver operating characteristic curve (AUC-ROC) can be computed to assess the effectiveness of the logistic regression model in predicting ad clicks. This evaluation helps advertisers compare different models, select the most appropriate one, and make data-driven decisions to optimize their ad campaigns.
5) *Scalability and Efficiency:* Logistic regression is computationally efficient and scalable, making it suitable for handling large-scale ad click prediction tasks. With the vast amount of ad impressions and user interactions happening in real-time, logistic regression allows for quick model training and prediction, enabling advertisers to make timely decisions and respond to changing market dynamics effectively.

III.PROPOSED ALGORITHM METHODOLOGY

In this section, we present some classification algorithms in machine learning and the methodology of this research.

A. Logistic Regression (LR)

Logistic regression (LR) is a method that is often used for classification, which is a statistical analysis technique applied for predictive models. This classification is one of the most popular machine learning algorithms that come under supervised learning techniques. Moreover, this classification model usually achieves high algorithm performance, so it is often applied in the industrial world. There are several types of logistic regression, namely binary and multinomial logistic regression. Binary logistic regression is used when the response variable is dichotomous. That is, there are only two categories. Meanwhile, multinomial linear regression is used when the response variable has more than two categories. This research uses binary linear regression. Another advantage of the logistic regression model is the ability to process large volumes of data at high speed because it requires less computational capacity, such as memory and processing power. This makes the model very suitable for data scientists to get multiple solutions with fast results. Logistic regression is also used extensively in the fields of medicine and social sciences, as well as in marketing, such as predicting a customer's propensity to buy a product or unsubscribe.

This logistic regression is a predictive model similar to linear regression based on the logistic function or the sigmoid function. The difference between the results of linear regression and logistic regression is that the range of values in linear regression is a real number, while the range of values in logistic regression is between 0 and 1. Then it also does not require a linear relationship between input and output variables since it uses a nonlinear log transformation approach to predict the odds ratio. In general, the assumptions of LR include:

1) There is no need for linearity between the independent and response variables.
2) There is no need to assume multivariate normality or equal variance between independent variables.
3) There is no need for the assumption of homoscedasticity.
4) The dependent variable must be dichotomous.
5) Do not need to transform into metric form.
6) The categories must be separate or exclusive to the independent variables.
7) Requires a relatively large sample for predictor variables, for example, a minimum of 50 data samples.
8) The odds ratio is a probability value.

This paper mainly addresses the usage of an algorithmic technique name Multiple Linear Regression. We achieve a greater accuracy on sales revenue using multiple linear regressions.

B. Data-Set Description

Dataset In this research, we use a dataset taken from Kaggle's website about the ad-click prediction. Then, we process this dataset to predict about customer and whether that customer clicked the ad and made the purchase. Therefore, we apply several classification algorithms to predict it. This data set contains the following features:

This data set contains the following features:

1) 'User ID': unique identification for consumer
2) 'Age': customer age in years
3) 'Estimated Salary': Avg. Income of consumer
4) 'Gender': Whether consumer was male or female
5) 'Purchased': 0 or 1 indicated clicking on Ad

Age and estimated salary, they have different ranges. We convert the values of age and estimated salary within the range of 0 to 1.once these values are converted in same range; it is easy to plot them

The response feature is purchased. This feature has two possible outcomes that are 0 and 1 where 0 refers to the case where a user didn't click the advertisement (class 0), while class 1 refers to the scenario where a user clicks the advertisement (class 1). This research divides data into 67% in training data and 33% in testing data.

Data-set description is very essential for understanding our data. In order to get insights from our data, we mainly use two commands. The first one is .info() command which gives us information about the number of rows and columns in the data-set and the other one is .describe()which explains various parameters like count(),min(),standard deviation(), max() etc.

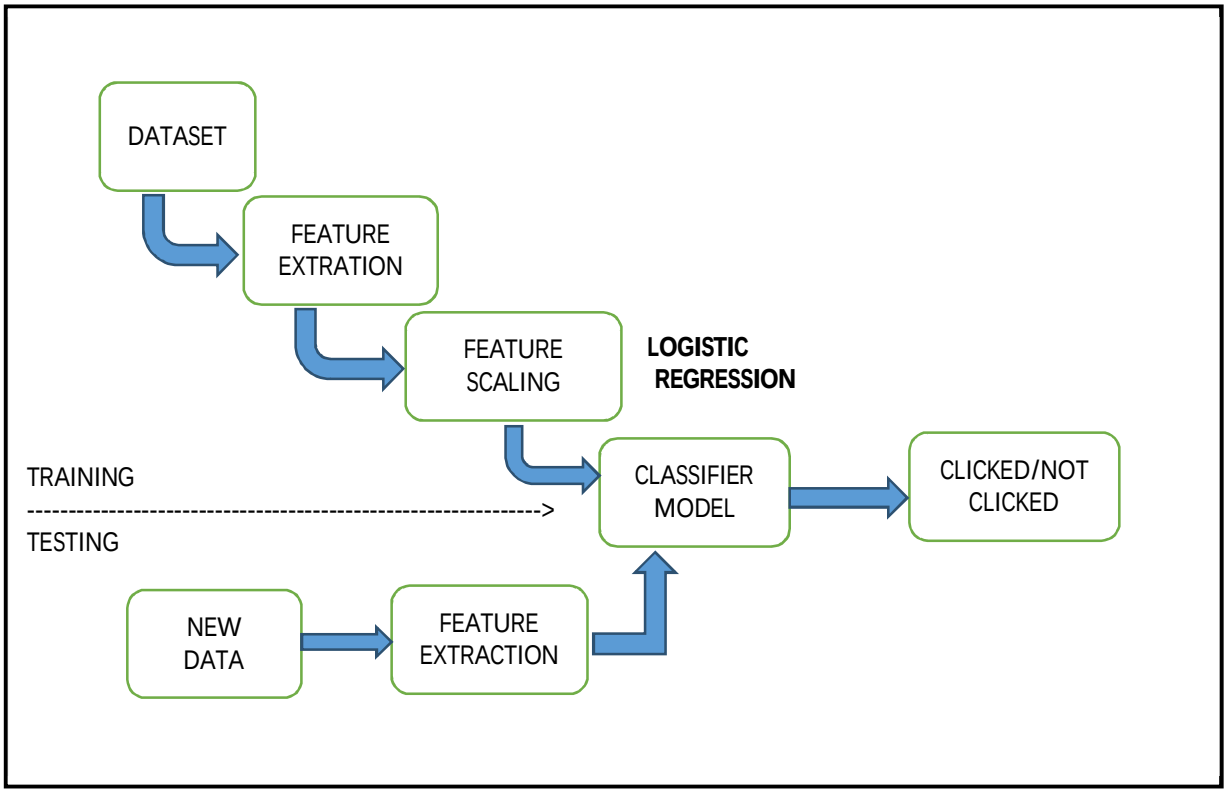
	User ID	Gender	Age	EstimatedSalary	Purchased
0	15624510	Male	19	19000	0
1	15810944	Male	35	20000	0
2	15668575	Female	26	43000	0
3	15603246	Female	27	57000	0
4	15804002	Male	19	76000	0
...	...	...	...	...	...
395	15691863	Female	46	41000	1
396	15706071	Male	51	23000	1
397	15654296	Female	50	20000	1
398	15755018	Male	36	33000	0
399	15594041	Female	49	36000	1

400 rows × 5 columns

Fig. 1 dataset

C. Method

An architecture overview is shown in Fig. 1. At the beginning, we input the dataset we need to classify. In classifying this data set, we used machine learning classifiers: LR. this classifier was applied to predict if a particular user would click on an online Advertisement. Furthermore, to obtain the best classifier method, we evaluate the performance of classification methods with confusion matrix and several metrics: sensitivity, specificity, precision, accuracy, F1-score, and AUC-ROC.



ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VII Jul 2023- Available at www.ijraset.com

D. Performance Evaluation

1) Confusion Matix

In machine learning, a confusion matrix, also known as an error matrix, is a table that allows visualization and evaluation of the performance of a classification model. It is commonly used to assess the accuracy of a classification algorithm by comparing the predicted and actual classes of a set of samples.

A confusion matrix is typically organized in a square matrix format, where each row represents the actual class labels, and each column represents the predicted class labels. The elements of the matrix indicate the counts or percentages of samples that fall into different combinations of predicted and actual classes.

The representation of the confusion matrix is a matrix table with four combinations of predicted values, and the actual value where the table can be seen in Table I. Suppose there are two classification results, namely positive (labeled 1) and negative (labeled 0), then the four combinations include 1). True Positive (TP) is the amount of positive data that is predicted to be true as positive, 2). False Negative (FN), which is the amount of data that is positive but is predicted to be negative 3). True Negative (TN) where the number of data that is negative and is predicted to be true as negative, and 4). False Negative (FN) is the amount of data that is positive but is predicted to be negative. Next, when a prediction result is a real number, a threshold value of t is needed to distinguish positive and negative classes, after which the confusion

Matrix can be made.

Confusion Matrix format as dispayed in the output of python

		Predicted	
		-	+
Actual	-	TN	FP
	+	FN	TP

Table I

2) Accuracy Of The Model

The accuracy of a classification model is a measure of how often the model correctly predicts the class or label of a given data point. It is calculated by dividing the number of correctly classified data points by the total number of data points.

$$\text{Accuracy} = \frac{TP + TN}{\text{Total number of predictions}}$$

TP: TRUE POSITIVE
TN: TRUE NEGATIVE

3) Misclassification

It is defined as the ratio of incorrect predictions to total predictions.

$$\text{Misclassification} = \frac{FP + FN}{\text{Total number of predictions}}$$

FP: FALSE POSITIVE
FN:FALSE NEGATIVE

4) Recall

It calculates the proportion of correctly predicted positive instances out of all actual positive instances. It indicates the model's ability to capture true positives and avoid false negatives.

$$TPR = \frac{TP}{\text{Total number of Actual Positives}}$$

5) Precision

It measures the proportion of correctly predicted positive instances out of all instances predicted as positive. It helps identify the model's ability to minimize false positives.

$$\text{Precision} = \frac{\text{TP}}{\text{Total number of Predicted Positives}}$$

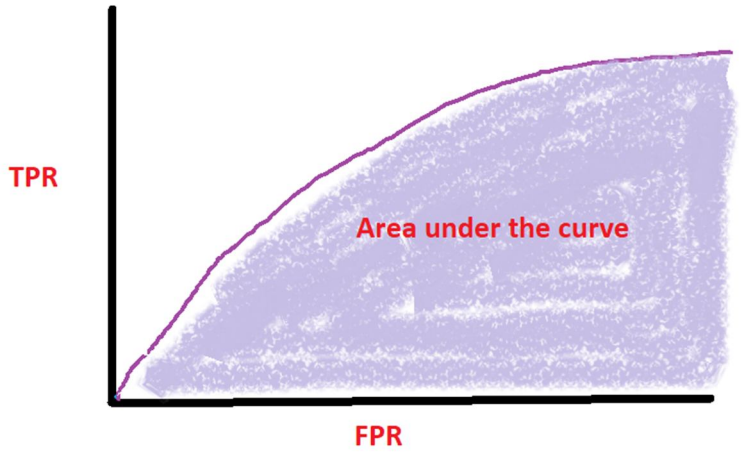
6) F1 Score

It combines precision and recall into a single metric, providing a balanced measure of a model's performance.

$$\text{F1 Score} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}}$$

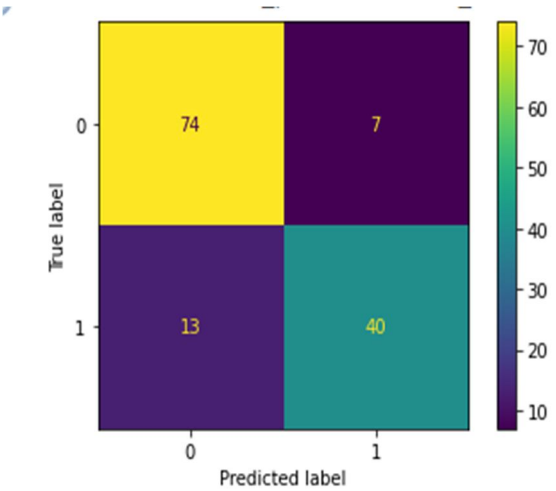
7) *AUC-ROC*

It evaluates the model's ability to distinguish between classes by plotting the true positive rate against the false positive rate. It provides an overall measure of the model's discriminatory power



Higher the Area under the curve, higher is TPR and that indicates that the model is doing a good job!

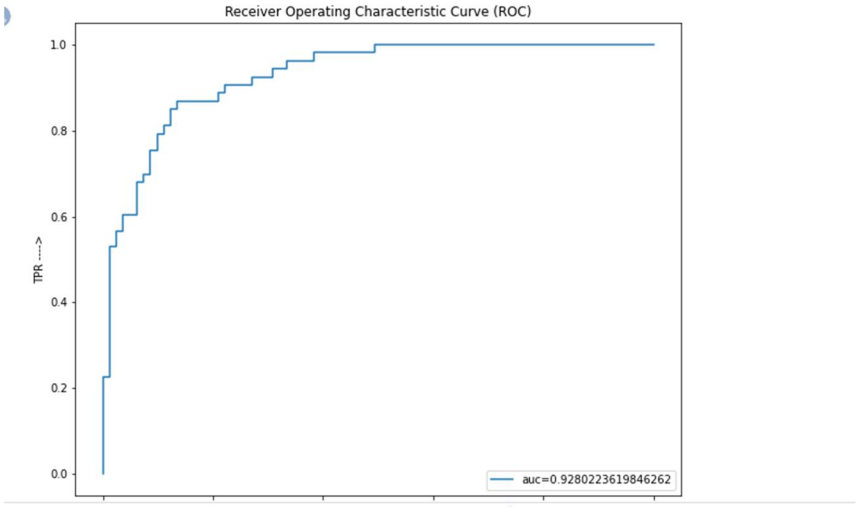
IV. OBSERVATIONS AND RESULTS



```
print(classification_report(y_test,y_pred))
```

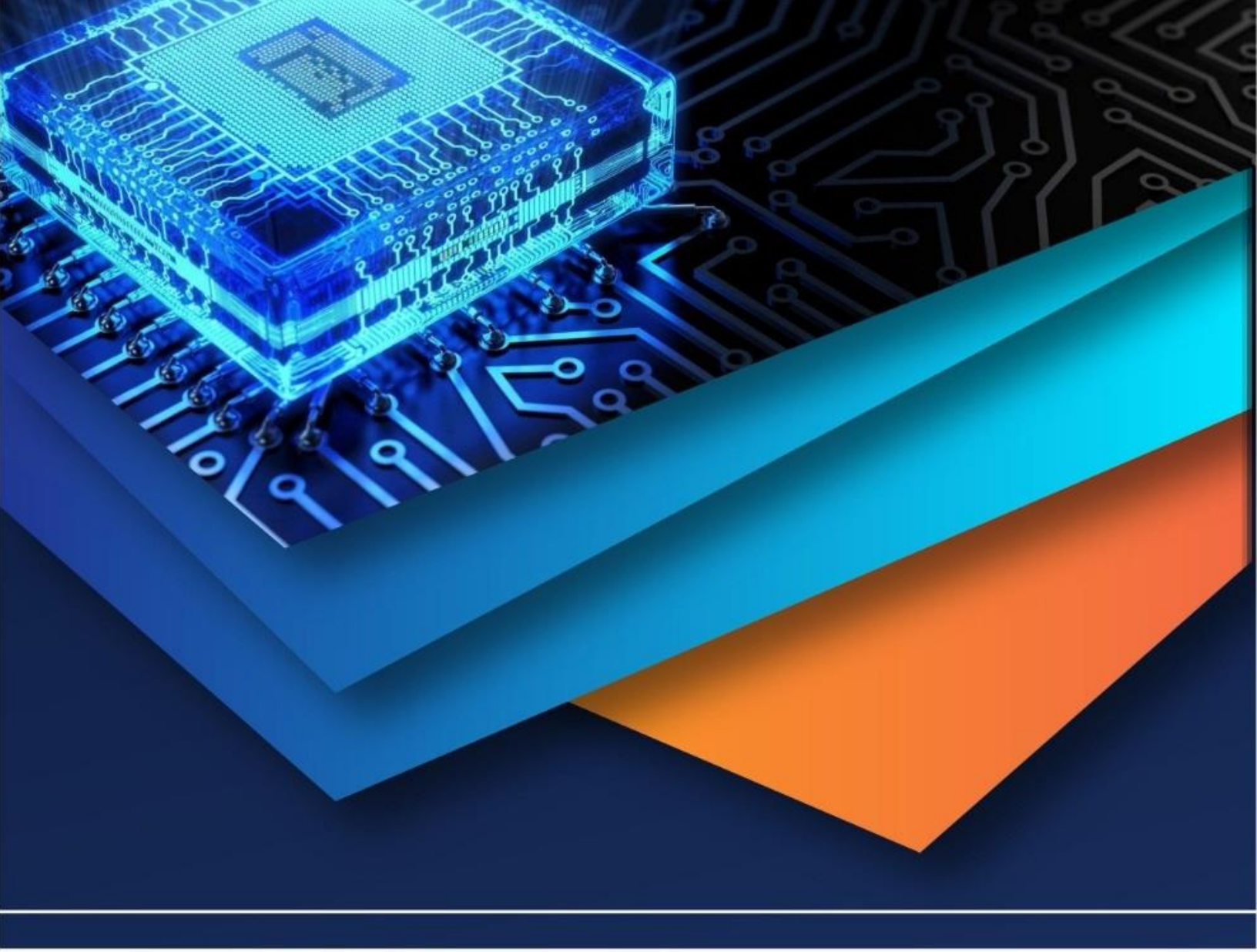
	precision	recall	f1-score	support
0	0.85	0.91	0.88	81
1	0.85	0.75	0.80	53
accuracy			0.85	134
macro avg	0.85	0.83	0.84	134
weighted avg	0.85	0.85	0.85	134


```
plt.show()
```



REFERENCES

[1] Https://www.kaggle.com/datasets
[2] 2)yasi dani ,maria ginting(march 2023)"classification of predicting customer ad clicks using logistic
[3] Regression and k-nearest neighbors "joiv : int. J. Inform. Visualization, 7(1) - march 2023 98-104 kodamagulla kausthub"commercials sales prediction using multiple linear regression"international research journal of engineering and technology (irjet) volume: 08 issue: 03 | mar 2021 .
[4] G. Shrivastava, v. Nagar, and s. K. Gill, "the effects of advertising on consumer buying behavior with special reference to fmcg industry," au-hiu int. Multidiscip. J., vol. 2, no. 1, pp. 1–8, 2022.
[5] A. Goldfarb, "what is different about online advertising?," rev. Ind. Organ., vol. 44, no. 2, 2014, doi: 10.1007/s11151-013-9399-3.
[6] D. Chakrabarti, d. Agarwal, and v. Josifovski, "contextual advertising by combining relevance with click feedback," 2008. Doi: 10.1145/1367497.1367554.
[7] y. Yang and p. Zhai, "click-through rate prediction in online advertising: a literature review," inf. Process. Manag., vol. 59, no. 2, 2022, doi: 10.1016/j.ipm.2021.102853.
[8] H. Cheng and e. Cantú-paz, "personalized click prediction in sponsored search," 2010. Doi: 10.1145/1718487.1718531.
[9] M. R. Farooqi and m. F. Ahmad, "the effectiveness of online advertising on consumers' mind - an empirical study," int. J. Eng. Technol., vol. 7, no. 2, 2018, doi: 10.14419/ijet.v7i2.11.11006.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY