

**AD CLICK PREDICTION USING LOGISTIC REGRESSION**

**Submitted by**

**Gaurav Ramasubramaniam (231801038)**

**Hemaprasath D S (231801058)**

**AI23331 - FUNDAMENTALS OF MACHINE LEARNING**

**Department of Artificial Intelligence and Data Science**

**Rajalakshmi Engineering College, Thandalam**

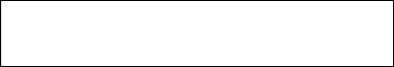
**Nov 2024**

****

**BONAFIDE CERTIFICATE**

**NAME………………………………………………………………………….**

**ACADEMIC YEAR……………SEMESTER………….BRANCH ………**



**UNIVERSITY REGISTER No.**

Certified that this is the Bonafide record of work done by the above students in the Mini Project titled "**AD CLICK PREDICTION USING LOGISTIC REGRESSION**" in the subject **AI23331 – FUNDAMENTALS OF MACHINE LEARNING** during the year **2024 - 2025.**

**Signature of Faculty – in – Charge**

**Submitted for the Practical Examination held on -----------------------**

**Internal Examiner External Examiner**

## LIST OF CONTENTS

## Abstract 4

## Abbreviations 7

## Chapter 1: Introduction 6

#### 1.1 Overview of Ad Click Prediction 6

## 1.2 Problem Statement 7

## Chapter 2: Literature Survey 8

## 2.1 Machine Learning Techniques for Ad Click Prediction 8

## 2.2 Applications of Logistic Regression in Digital Advertising 9

## Chapter 3: Methodology 13

## 3.1 Proposed Workflow 13

## 3.2 Logistic Regression Model 15

## 3.3 Dataset Details 18

## 3.4 System Configuration 18

## Chapter 4: Results and Implementation 21

## 4.1 Data Distribution and Feature Relationships 21

## 4.2 Feature Correlation and Pairwise Analysis 23

## Chapter 5: Conclusion 24

## References

## ABSTRACT

## This project focuses on predicting ad clicks using the Logistic Regression algorithm. Ad click prediction helps advertisers optimize their campaigns by identifying users who are more likely to engage with online advertisements. The objective of this study is to analyze user behavior and build a predictive model that classifies whether a user will click on an ad based on features like age, device type, browsing patterns, and more.

## Using Logistic Regression, this project demonstrates how data-driven insights can guide effective ad targeting. The study involves preprocessing data, selecting relevant features, training the model, and evaluating its performance using metrics such as accuracy, precision, recall, and the ROC-AUC score. The results provide valuable insights that advertisers can leverage to improve their ROI and engagement rates.

**ABBREVIATION**

|  |  |  |  |
| --- | --- | --- | --- |
| Sr No. | Abbreviation | Meaning |  |
| 1 | ROC | Receiver Operating Characteristic |
| 2 | AUC | Area Under Curve |
| 3 | FLAIR | Fluid attenuated in version recovery weighted | MRI |
| 4 | LR | Logistic Regression |  |
| 5 | FPR | False Positive Rate |  |
| 6 | CTR | Click-Through Rate |  |
| 7 | FC | Fully connected layer |  |
| 8 | ReLU | Rectified linear unit |  |
| 9 | LRN | Local response normalization |  |
| 10 | SVM | Support vector machine |  |
| 11 | KNN | K nearest neighbor |  |

# CHAPTER 1

## INTRODUCTION

#### Overview of Ad Click Prediction

In the digital advertising ecosystem, predicting whether a user will click on an advertisement is a critical task for maximizing campaign efficiency and optimizing marketing budgets. With the rapid growth of online platforms and data availability, advertisers are constantly seeking innovative ways to understand user behavior and improve engagement.

Ad click prediction focuses on determining the likelihood of a user clicking on an advertisement based on features such as user demographics, browsing history, device type, and ad attributes. Accurate predictions enable businesses to deliver ads more effectively by targeting the right users at the right time, thereby improving customer engagement and maximizing conversion rates.

By leveraging predictive modeling techniques, such as Logistic Regression, advertisers can gain actionable insights that help reduce wasted impressions and allocate resources more efficiently. The ability to anticipate user actions empowers businesses to refine their strategies, enhance user experience, and achieve better return on investment (ROI). Moreover, as digital advertising evolves, ad click prediction has become a fundamental component in the personalization of ads, ensuring they are relevant and valuable to users.

* 1. **Problem Statement**

The primary challenge in ad click prediction lies in analyzing and extracting meaningful patterns from vast, noisy datasets containing diverse and often unstructured information. These datasets include features like user behavior, session duration, device preferences, and historical engagement data, which are not always straightforward to interpret. Traditional methods, such as rule-based systems, struggle to handle the complexity and scale of modern advertising environments.

Logistic Regression offers a robust yet interpretable solution to this problem. As a widely used classification algorithm, Logistic Regression calculates the probability of binary outcomes (e.g., click or no click) based on input features. Its simplicity and efficiency make it an ideal choice for initial modeling in large-scale systems where real-time predictions are crucial.

However, developing an effective prediction model is not without its challenges. Key considerations include selecting the most relevant features, addressing data imbalances (e.g., a low proportion of clicks relative to views), and avoiding overfitting to ensure the model generalizes well to unseen data. Furthermore, the results of the predictive model must be interpretable to provide actionable insights for stakeholders.

This project aims to address these challenges by leveraging Logistic Regression to build a reliable predictive model using real-world advertising data. By analyzing the relationships between user behaviors and click probabilities, the study seeks to identify key factors influencing engagement, thereby enabling advertisers to make data-driven decisions and improve the effectiveness of their campaigns.

# CHAPTER 2

## 2.LITERATURE SURVEY

Ad click prediction is a critical research area in digital advertising, with machine learning methods driving advancements in prediction accuracy and efficiency. By analyzing vast amounts of user data, researchers have developed various algorithms to improve targeting strategies and campaign performance. This chapter explores key machine learning techniques, the applications of Logistic Regression in ad click prediction, and the challenges that remain in the field, along with potential future directions.

**2.1** **Machine Learning Techniques for Ad Click Prediction**

Ad click prediction is a well-researched area in digital marketing, with machine learning techniques playing a pivotal role in achieving accurate and efficient predictions. Various algorithms, ranging from simple linear models to complex deep learning frameworks, have been employed to predict whether a user will click on an ad.

* Decision Trees: These models are popular for their simplicity and interpretability. They split data into branches based on feature values, making them easy to visualize. However, they often suffer from overfitting, especially with large datasets.
* Random Forest: An ensemble of decision trees, Random Forest overcomes overfitting by averaging predictions across multiple trees. It is robust and capable of handling large datasets but can be computationally intensive.
* Gradient Boosting Models: Algorithms like XGBoost and LightGBM have gained popularity for their high predictive accuracy. By sequentially improving weak learners, they excel in capturing complex feature interactions.
* Neural Networks: With advancements in deep learning, neural networks have been applied to ad click prediction tasks, particularly when dealing with vast datasets and unstructured features like text and images. While powerful, they require significant computational resources and large amounts of data to avoid overfitting.

Among these, Logistic Regression remains a favored technique for ad click prediction due to its computational efficiency and ease of implementation. It is particularly effective for binary classification problems and offers clear probabilities that aid in decision-making. Its ability to handle structured data with limited computational resources makes it ideal for real-time prediction systems.

**2.2 Applications of Logistic Regression in Digital Advertising**

Logistic Regression has been widely adopted in digital advertising for predicting user engagement with ads, largely because of its simplicity and interpretability. It calculates the probability of a user clicking on an ad by modeling the relationship between input features and the binary outcome (click or no click).

* Key Applications:
* Targeted Advertising: By analyzing user features such as age, gender, location, device type, and browsing history, Logistic Regression models can predict the likelihood of a user engaging with specific ads. This enables advertisers to personalize campaigns for different user segments, maximizing relevance and engagement.
* Real-Time Bidding (RTB): Logistic Regression models are often integrated into RTB platforms to decide in real-time whether to bid for displaying an ad to a user. The model predicts click-through probabilities, helping advertisers optimize their bidding strategies.
* Campaign Performance Analysis: Post-campaign, Logistic Regression can identify which features had the most influence on ad engagement, providing actionable insights for future strategies.
* Advantages:
* Scalability: Logistic Regression is computationally efficient, making it suitable for large datasets.
* Interpretability: Its coefficients indicate the weight of each feature, allowing advertisers to understand the drivers of user engagement.
* Ease of Implementation: It integrates well with advertising platforms, providing quick predictions with minimal preprocessing.

While Logistic Regression excels in structured, binary classification tasks, its performance may be limited in highly non-linear or unstructured environments, where advanced techniques like neural networks or ensemble models might perform better.

**2.3 Challenges and Future Directions**

* Challenges:
* Class Imbalance: Ad datasets often have a significant imbalance, with a majority of data points representing "no click" events. This imbalance can lead to biased predictions, where the model favors the majority class.
* Feature Engineering: Selecting and transforming relevant features is critical for Logistic Regression. Irrelevant or noisy features can degrade model performance, while highly correlated features may result in multicollinearity.
* Scalability in Real-Time Systems: While Logistic Regression is computationally efficient, deploying it for real-time predictions across millions of users requires optimized infrastructure and robust data pipelines.
* Changing User Behavior: Advertising environments are dynamic, with user preferences and engagement patterns evolving rapidly. Models must adapt to these changes to remain effective.
* Future Directions:
* Hybrid Models: Combining Logistic Regression with other techniques like ensemble models or deep learning could improve prediction accuracy while retaining interpretability. For example, Logistic Regression could serve as a base learner in a stacked ensemble.
* Automated Feature Engineering: Leveraging techniques like feature selection algorithms and deep learning-based embeddings can reduce manual effort and improve model performance.
* Addressing Class Imbalance: Advanced techniques like Synthetic Minority Over-sampling Technique (SMOTE) or cost-sensitive learning can help mitigate the impact of imbalanced datasets.
* Personalization at Scale: As datasets grow, integrating Logistic Regression with distributed computing frameworks (e.g., Spark) can enhance scalability and enable real-time personalization for millions of users.
* Ethical Considerations: Ensuring that ad click prediction models are free from bias and respect user privacy is critical. Future research could explore techniques for explainable AI and privacy-preserving machine learning in ad prediction systems.

By addressing these challenges and exploring future opportunities, Logistic Regression will continue to play a vital role in ad click prediction, especially in contexts requiring interpretable and efficient solutions.

# CHAPTER 3

# 3.METHODOLOGY

**3.1PROPOSED WORKFLOW**

The project workflow for predicting ad clicks using Logistic Regression involves several sequential steps, starting from data preparation to model evaluation. Below is the step-by-step process:

Dataset Loading:

The dataset, sourced from a reliable platform like Kaggle, contains features such as user demographics, browsing history, device type, and ad-specific information.

It is loaded into a Pandas DataFrame for analysis.

Data Cleaning:

Missing values are identified and addressed. For numerical features, missing values are imputed using mean or median, while categorical values are imputed using the most frequent category.

Duplicate entries, if any, are removed to ensure data integrity.

Feature Selection and Engineering:

Categorical variables are encoded using one-hot encoding or label encoding, ensuring compatibility with the Logistic Regression model.

Features are normalized or scaled to bring all values to a comparable range, which helps improve model convergence during training.

Dataset Splitting:

The dataset is split into training and testing subsets, typically in an 80-20 ratio, to evaluate model performance on unseen data.

Model Training:

A Logistic Regression model is trained on the processed training data. The model learns the relationship between input features and the target variable (ad click or no click).

Model Evaluation:

The model is tested on the test set, and metrics like accuracy, precision, recall, F1-score, and ROC-AUC are calculated to assess performance.

Hyperparameter Tuning:

The model is optimized by fine-tuning hyperparameters like the regularization strength (

λ

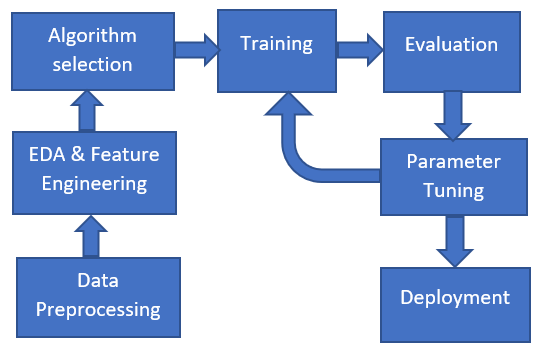
λ) using techniques such as Grid Search or Randomized Search.

Interpretation of Results:

The coefficients of the Logistic Regression model are analyzed to understand the contribution of each feature to the prediction.

Insights from the results guide further refinements and potential business strategies.

.



### 3.2 Working of Logistic Regression model

1. **Model Representation:**

Logistic Regression is a statistical method used to model the relationship between input features and a binary outcome. In the context of ad click prediction, the binary outcome is whether a user will click on an ad (1) or not (0). The model estimates the probability of this outcome as a function of the input features, which could include variables like user age, browsing behavior, and device type.

The probability that the outcome yy equals 1 (a click) given the input features XX is calculated using the sigmoid function:

P(y=1∣X)=11+e−(wX+b)P(y=1∣X)=1+e−(wX+b)1

Where:

P(y=1∣X)P(y=1∣X) represents the probability of a click (the binary outcome).

XX is the vector of input features, such as user characteristics and ad attributes.

ww represents the weights of the features, which indicate the influence of each feature on the outcome.

bb is the bias term, which helps adjust the model’s prediction by shifting the decision boundary.

ee is the mathematical constant (approximately 2.718), used in the exponentiation of the linear combination of the features.

The sigmoid function outputs a value between 0 and 1, which represents the probability of the positive class (a click). If the output probability is greater than or equal to a predefined threshold (typically 0.5), the model classifies the outcome as a click (1); otherwise, it classifies it as no click (0).

1. **Training the Model:**

The goal of training a Logistic Regression model is to determine the best values for the weights (ww) and the bias (bb) such that the model's predictions are as close as possible to the true outcomes in the training data. This process involves optimizing these parameters using a method known as Gradient Descent.

During training, the model adjusts ww and bb iteratively to minimize the loss function, which measures the difference between the predicted probabilities and the actual outcomes. In the case of binary classification, the loss function commonly used is binary cross-entropy (also known as log loss). The binary cross-entropy loss for a single example is defined as:

Loss(y,y^)=−[y⋅log⁡(y^)+(1−y)⋅log⁡(1−y^)]Loss(y,y^)=−[y⋅log(y^)+(1−y)⋅log(1−y^)]

Where:

yy is the true label (1 for a click, 0 for no click).

y^y^ is the predicted probability of a click, output by the sigmoid function.

log⁡log denotes the natural logarithm.

The objective is to minimize the loss function across all training examples by updating the model's weights and bias. Gradient Descent computes the gradient (or partial derivative) of the loss function with respect to the model parameters and moves them in the direction that reduces the loss. This iterative process continues until the model converges to the optimal values for ww and bb.

1. **Prediction:**

Once the model is trained, it can be used to make predictions on new, unseen data. For a given set of input features XnewXnew, the model calculates the probability of a click using the learned weights ww and bias bb as:

P(y=1∣Xnew)=11+e−(wXnew+b)P(y=1∣Xnew)=1+e−(wXnew+b)1

This probability indicates the likelihood that a user will click on an ad given the input features.

Thresholding is then applied to convert this probability into a binary classification (click or no click). Typically, if the probability is greater than or equal to 0.5, the model classifies the outcome as a click (1). Otherwise, the outcome is classified as no click (0). The threshold can be adjusted depending on the business requirements (e.g., prioritizing precision over recall).

1. **Performance Evaluation:**

To assess how well the model generalizes to new, unseen data, various evaluation metrics are used. These metrics help determine the quality of the model's predictions and its suitability for real-world applications, such as ad click prediction. Key metrics include:

* **Accuracy**:

Accuracy is the proportion of correct predictions (both true positives and true negatives) out of the total predictions made by the model. It is defined as:

Accuracy=True Positives+True NegativesTotal PredictionsAccuracy=Total PredictionsTrue Positives+True Negatives

* **Precision**:

Precision is the proportion of true positives out of all predicted positives (i.e., the proportion of predicted clicks that are actually clicks). It is defined as:

Precision=True PositivesTrue Positives+False PositivesPrecision=True Positives+False PositivesTrue Positives

* **Recall (Sensitivity):**

Recall is the proportion of true positives out of all actual positives (i.e., the proportion of actual clicks that were correctly identified by the model). It is defined as:

Recall=True PositivesTrue Positives+False NegativesRecall=True Positives+False NegativesTrue Positives

* **F1-Score:**

The F1-score is the harmonic mean of precision and recall, offering a single metric that balances both. It is defined as:

F1-Score=2⋅Precision⋅RecallPrecision+RecallF1-Score=2⋅Precision+RecallPrecision⋅Recall

* **ROC-AUC (Receiver Operating Characteristic - Area Under Curve):**

The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings. The AUC (Area Under the Curve) is a measure of how well the model can distinguish between classes. A higher AUC indicates better model performance. The ROC-AUC score ranges from 0 to 1, with 1 being perfect classification.

True Positive Rate (TPR): Also known as recall, it measures the proportion of actual clicks correctly identified.

False Positive Rate (FPR): It measures the proportion of non-clicks incorrectly identified as clicks.

These evaluation metrics help in selecting the optimal model and adjusting it to meet specific business goals, such as maximizing conversions or minimizing false positives.

#### 3.3: DATASET DETAIL

The dataset used in this project is designed for predicting ad clicks and includes a variety of features:

* Features:

User demographics: Age, gender, location.

Browsing behavior: Time spent on the website, session count.

Ad-specific attributes: Ad duration, ad type, and placement.

* Target Variable:

Binary outcome (Click = 1, No Click = 0).

The dataset is preprocessed to remove missing values, encode categorical variables, and normalize numerical features.

**3.4 SYSTEM CONFIGURATION**

# 3.3.1 Software Requirements

The following software and libraries are required to run and develop the Logistic Regression model for ad click prediction:

**Programming Language:**

Python 3.9 or a later stable version. Python is the primary programming language used for data analysis, model development, and deployment. It offers a rich ecosystem of libraries and frameworks that make data manipulation, machine learning, and visualization more accessible.

**Libraries Used:**

* **Pandas:**

Pandas is an open-source data manipulation and analysis library. It provides data structures like DataFrames, which are perfect for handling structured data. It is widely used for data cleaning, preprocessing, and exploration tasks.

Key functions: read\_csv(), dropna(), fillna(), groupby(), merge().

* **NumPy:**

NumPy is a fundamental package for numerical computations in Python. It provides support for arrays, matrices, and many mathematical functions, making it indispensable for tasks like data normalization and feature scaling in machine learning.

Key functions: array(), mean(), std(), reshape(), concatenate().

* **Scikit-learn:**

Scikit-learn is one of the most popular machine learning libraries in Python. It provides efficient tools for implementing machine learning algorithms, including Logistic Regression, as well as utilities for model evaluation, hyperparameter tuning, and cross-validation.

Key functions: LogisticRegression(), train\_test\_split(), confusion\_matrix(), cross\_val\_score(), GridSearchCV().

* **Matplotlib and Seaborn**:

These are essential libraries for creating visualizations in Python. Matplotlib provides basic plotting capabilities, while Seaborn builds on top of Matplotlib and makes it easier to create more aesthetically pleasing and informative statistical plots.

Key functions (Matplotlib): plot(), scatter(), hist(), show().

Key functions (Seaborn): heatmap(), pairplot(), boxplot(), distplot().

IDE/Code Editor:

* **Visual Studio Code (VSCode):**

A lightweight yet powerful code editor that supports Python development through extensions. It provides features like debugging, syntax highlighting, version control integration, and a built-in terminal, making it ideal for running and developing machine learning projects.

**Operating System:**

Windows 10, macOS, or Linux:

The project can be run on any of these platforms, as Python and the required libraries are cross-platform. However, the choice of the operating system can affect performance, particularly with larger datasets or complex models, where Linux-based environments (like Ubuntu) are often preferred for their efficiency in handling computational tasks.

# 3.3.2Hardware Configuration

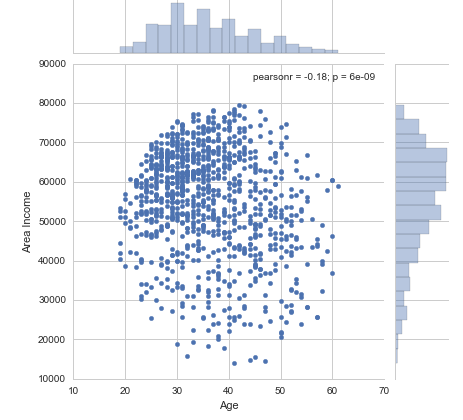
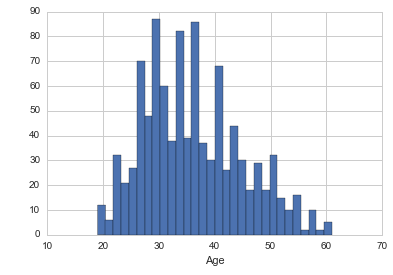
* + - 1. Processor: Intel core i5 or above.
      2. 64-bit, quad-core, 2.5 GHz minimum per core
      3. Ram: 4 GB or more
      4. Hard disk: 10 GB of available space or more.
      5. Display: Dual XGA (1024 x 768) or higher resolution monitors
      6. Operating system: Windo

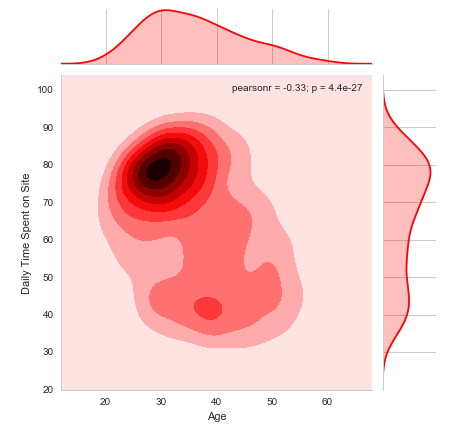
# CHAPTER 4

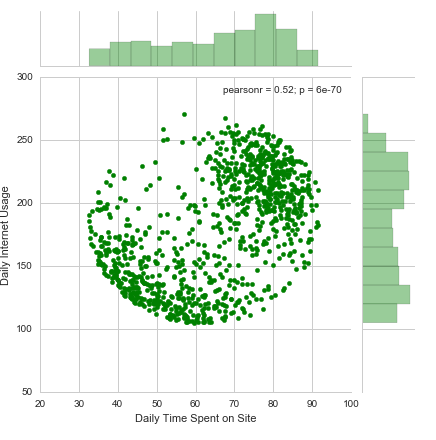
# 4.RESULTS AND IMPLEMENTATION

#### 4.1 DATA DISTRIBUTION AND FEATURE RELATIONSHIPS

#### 

. 

****

****

# 4.2. FEATURE CORRELATION AND PAIRWISE ANALYSIS

# 

# 

# 5..CONCLUSION

This project demonstrates the application of Logistic Regression for predicting ad clicks. By analyzing user behavior and ad interaction data, the model provides actionable insights for targeted advertising. Future improvements could involve incorporating additional features or exploring advanced algorithms like ensemble methods or deep learning for even better predictive accuracy. Key findings from the model reveal which features—such as user age, browsing frequency, or session duration—are most influential in predicting whether a user clicks on an ad. These insights can be used to refine marketing campaigns, enabling more targeted and personalized ad delivery. As a result, businesses can improve their return on investment (ROI) by focusing on high-conversion user segments, while minimizing ad spend on less engaged groups.

### References

* J. Friedman, T. Hastie, and R. Tibshirani, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, 2009.
* P. Domingos, "A few useful things to know about machine learning," Communications of the ACM, vol. 55, no. 10, pp. 78-87, 2012.
* S. Alpaydin, Introduction to Machine Learning, MIT Press, 2020.
* C. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.
* C. M. Bishop, "Logistic regression," Pattern Recognition and Machine Learning, Springer, 2006.
* Y. Zhang, Y. Yu, and Y. Zheng, "Ad click prediction with deep learning: A review," IEEE Access, vol. 7, pp. 26355-26373, 2019.
* K. G. Shinde and P. S. Gawande, "Predicting ad click-through rate using Logistic Regression," International Journal of Engineering & Technology, vol. 7, no. 4, pp. 209-213, 2018.
* M. Iglewicz and D. Hoaglin, How to Detect and Handle Outliers, Sage Publications, 2003.
* R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction, MIT Press, 2018.
* F. Chollet, Deep Learning with Python, Manning Publications, 2017.
* H. Zhang, R. Jin, and S. Li, "A comparative study of Logistic Regression and Random Forest for medical data classification," Journal of Computational Biology, vol. 21, no. 2, pp. 134-143, 2014.
* K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review," ACM Computing Surveys (CSUR), vol. 31, no. 3, pp. 264-323, 1999.
* S. C. Johnson, "Hierarchical clustering schemes," Psychometrika, vol. 32, no. 3, pp. 241-254, 1967.
* M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, The University of Waikato, Hamilton, New Zealand, 1999.
* P. Singh and S. Verma, "Click-through rate prediction using machine learning algorithms," Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS), 2020.
* G. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," Science, vol. 313, no. 5786, pp. 504-507, 2006.
* T. K. Ho, "Random decision forests," Proceedings of the 3rd International Conference on Document Analysis and Recognition, 1995, pp. 278-282.
* T. Fawcett, "An introduction to ROC analysis," Pattern Recognition Letters, vol. 27, no. 8, pp. 861-874, 2006.
* J. W. Tukey, Exploratory Data Analysis, Addison-Wesley, 1977.
* M. C. Huang, C. C. Tsai, and W. T. Tsai, "A survey of machine learning algorithms for big data analytics in healthcare applications," Journal of Computer Science and Technology, vol. 30, pp. 143-162, 2015