MINI PROJECT REPORT

**ADVERTISEMENT CLICK-THROUGH RATE PREDICTION**

**Abstract**

This program performs a comprehensive analysis of a dataset of online advertisements to predict the click-through rate of advertisements. The key steps include data cleaning, visualization, correlation analysis, and machine learning classification.

The program starts by importing necessary libraries and reading a dataset containing information about online ad interactions. Duplicate rows are removed, and some irrelevant columns are dropped. Categorical features are encoded, and a boxplot is used to visualize the distribution of numerical features. The dataset is then further analyzed by creating a correlation matrix. Box plots are generated to explore the relationship between different features and the target variable. Outliers are handled using the IQR method. Finally, the program applies various machine learning models (KNN, LDA, Random Forest, SVM) to predict ad clicks. Cross-validation is used to assess model performance, and the results are visualized with a boxplot. The Random Forest model is selected, trained, and evaluated on a test set, with accuracy, classification report, and confusion matrix being displayed. The program concludes with a scatter plot comparing the predicted and actual values.

**Introduction**

In the era of digital marketing, understanding user behavior and predicting ad click-through rates (CTR) is crucial for advertisers to optimize their campaigns and allocate resources effectively. The "Advertisement Click Through Rate Prediction" project aims to analyze a dataset containing information about users interacting with online ads and develop a predictive model to forecast whether a user will click on an ad. By the end of this project, advertisers will gain actionable insights into user behavior, allowing them to make data-driven decisions to enhance the effectiveness of their online advertising campaigns.

**Objective**

The primary objective of the "Advertisement Click Through Rate Prediction" project is to develop a predictive model that accurately forecasts whether a user will click on an online advertisement. This involves leveraging data analysis, machine learning techniques, and statistical methods. The project aims to equip advertisers with a reliable predictive tool that enhances their ability to make informed decisions and maximize the effectiveness of online advertising campaigns.

**Scope of the Project**

The scope of the "Advertisement Click Through Rate Prediction" project encompasses various stages, from data preprocessing to model development and evaluation. This helps in deriving actionable insights from the analysis and model predictions and provide recommendations for advertisers to optimize ad targeting and improve click-through rates based on the project findings.

**Significance of the Study**

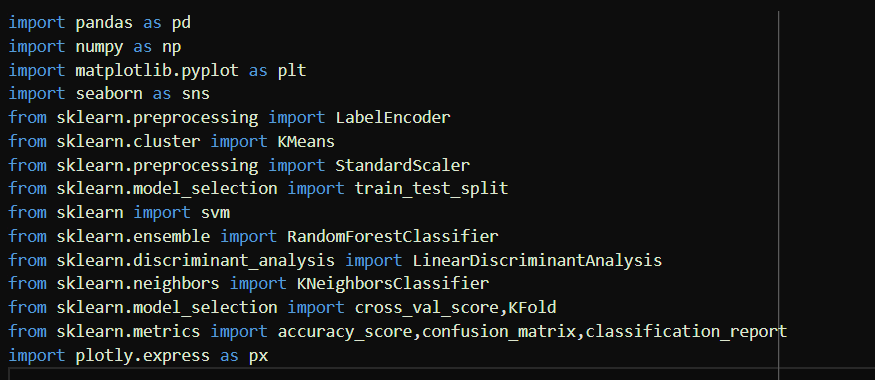
The "Advertisement Click Through Rate Prediction" project holds significant importance in the realm of digital marketing and online advertising. By predicting click-through rates accurately, advertisers can optimize their ad campaigns. This ensures that promotional content reaches a more receptive audience, leading to increased engagement and potentially higher conversion rates. Advertisers can allocate their resources more effectively by focusing on target audiences more likely to click on ads. This strategic approach helps maximize the impact of advertising budgets. The project's insights and predictions contribute to cost-efficient advertising strategies. Advertisers can avoid spending resources on audiences less likely to engage, reducing overall advertising costs.

In summary, the "Advertisement Click Through Rate Prediction" project is significant as it empowers advertisers with a predictive tool to make informed decisions, improve ad targeting, and ultimately enhance the efficiency and effectiveness of online advertising campaigns in a dynamic digital landscape.

**Procedure**

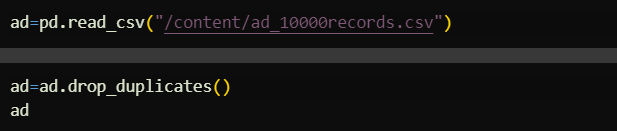
This Python program employs various classification models, including K-neighbors classifier and Random forest classifier, to predict click-through rates based on features such as details of the audience and type of the advertisement. The dataset is first loaded using Pandas and visualized using Seaborn and Plotly. The data is then cleaned, including grouping countries based on user counts. Encoded categorical features using LabelEncoder. Explored and handled outliers in specific numerical columns. Model performance is compared using cross-validation, and the models are trained and evaluated using metrics like Accuracy. The program also includes visualizations of predicted vs. actual points and real data vs. predictions for Randomforest classifier. Finally, the models are applied to new data for predicting delivery times in real-world scenarios.

1. **Importing required libraries**

****

Importing necessary libraries for data manipulation (pandas), numerical operations (numpy), plotting (matplotlib and seaborn), preprocessing (LabelEncoder, StandardScaler), machine learning models (svm, RandomForestClassifier, LinearDiscriminantAnalysis, KNeighborsClassifier), cross-validation (cross\_val\_score, KFold), and visualization (plotly.express).

1. **Reading the dataset**

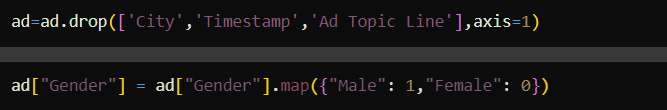


We load our dataset(ad\_10000records.csv) using pandas and drop any duplicate records.

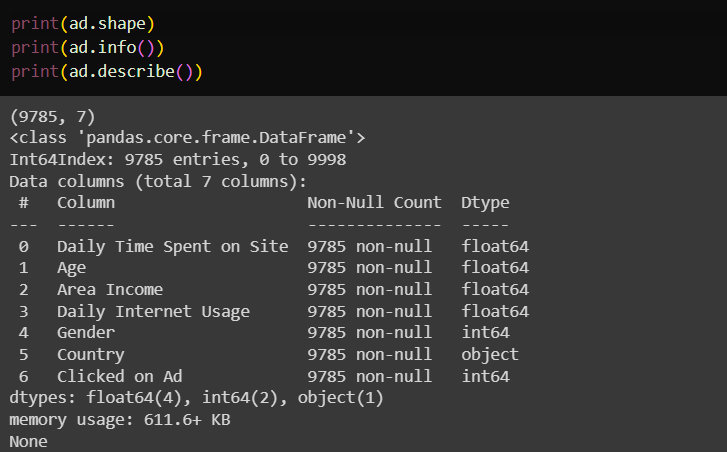
1. **Data cleaning**

Data cleaning steps contribute to preparing the data for further analysis and modeling. They address issues such as removing unnecessary columns, deriving new features, and encoding categorical variables.

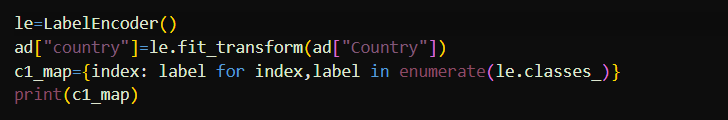
* 1. **Removing columns and mapping variables**



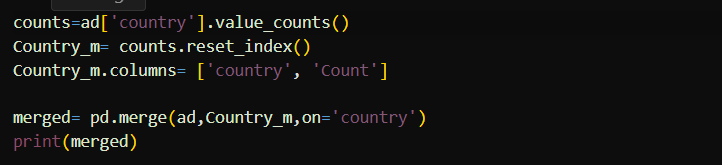
Mapping the "Gender" column to numerical values (1 for "Male" and 0 for "Female") and dropping unnecessary columns ('City', 'Timestamp', 'Ad Topic Line').

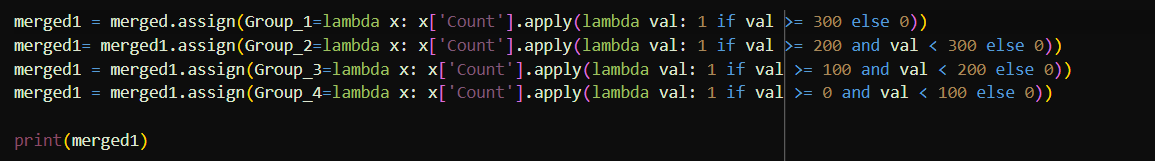


* 1. **Categorical variable encoding**



Using LabelEncoder to convert the "Country" column into numerical values and creating a mapping dictionary for the encoded values.

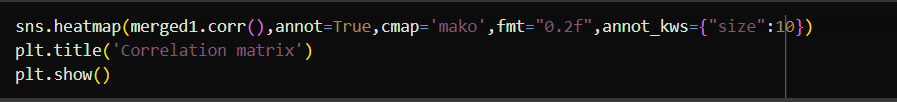




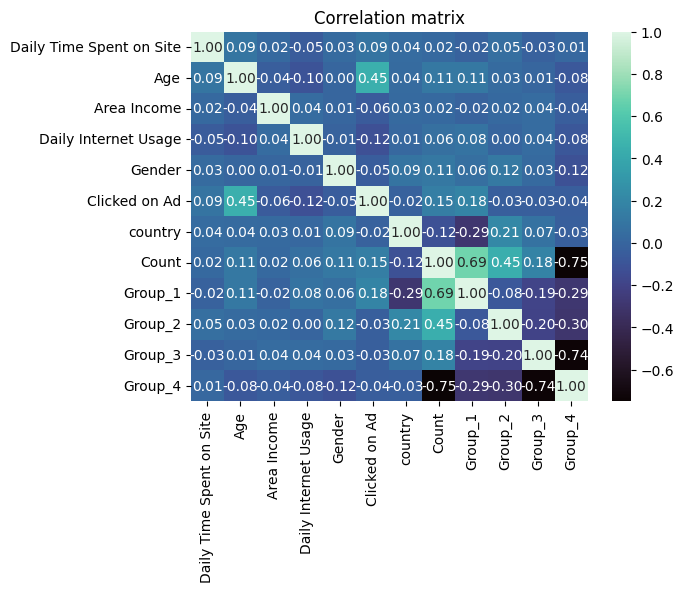
Counting the occurrences of each country and merging the counts with the main dataset. Creating groups based on the count of each country.

1. **Data visualization**

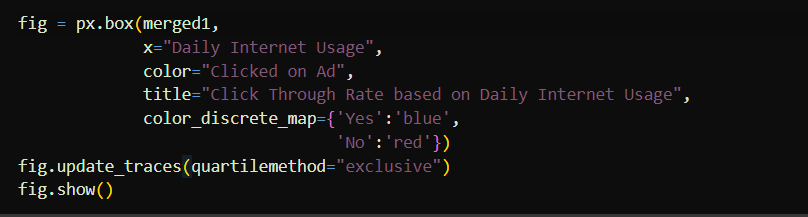
Data visualization offer insights into the relationships between different features and the target variable **Clicked on Ad** in the dataset. They can aid in understanding patterns, correlations, and potential influences on click through rate.

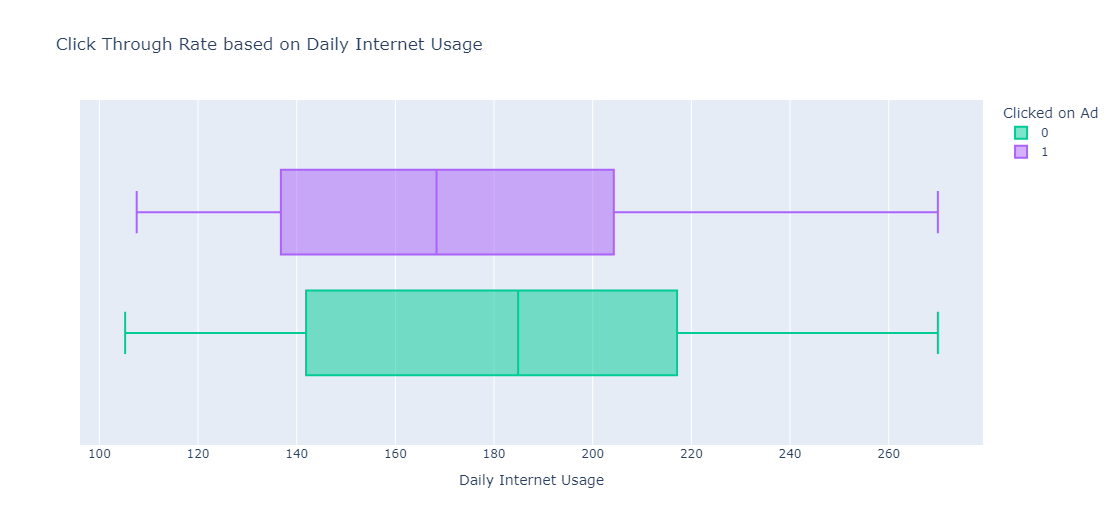
****

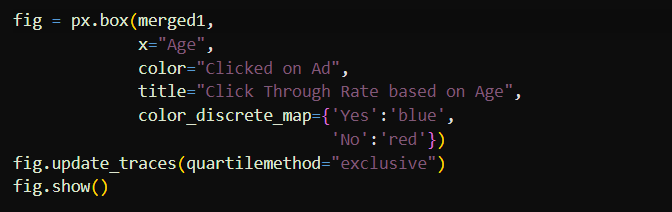
The heatmap provides a visual representation of the correlations between the selected numerical variables. It helps identify patterns, relationships, and potential multicollinearity between features. The color intensity and annotation values indicate the strength and direction of the correlations.

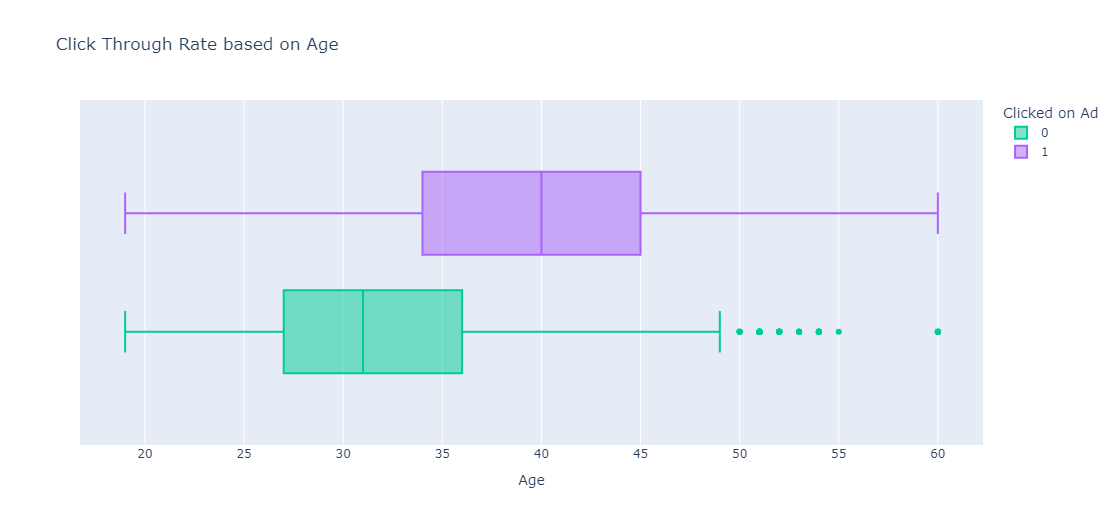


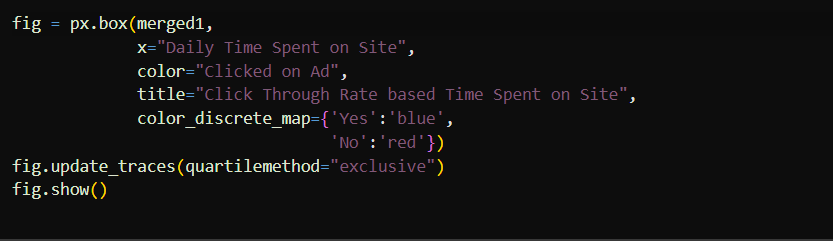
Create box plots for different features to analyze their impact on the "Clicked on Ad" variable.

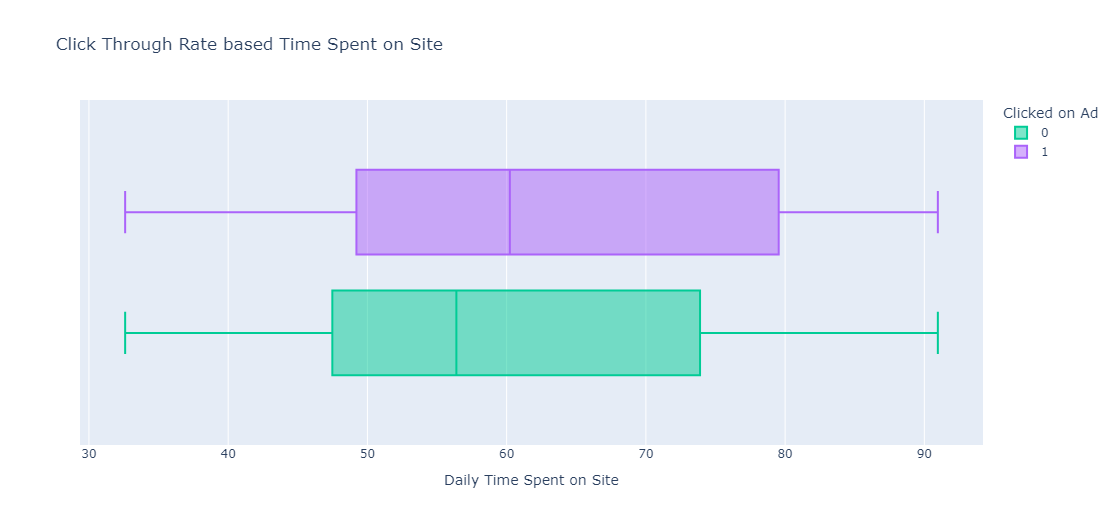


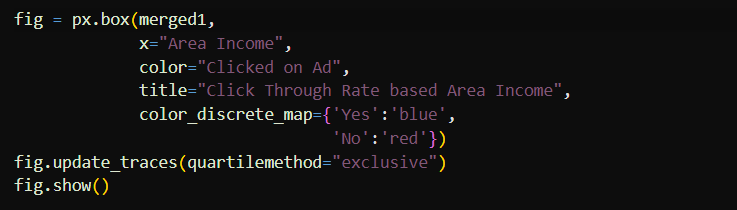




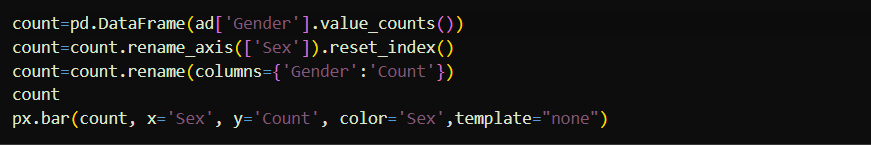


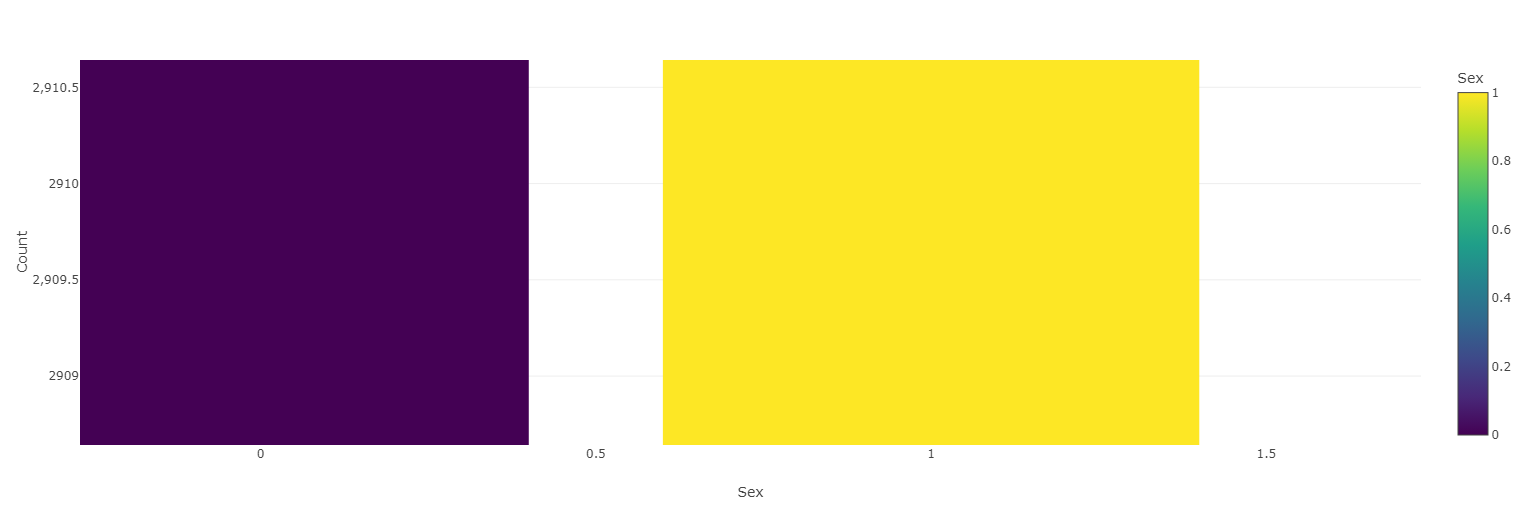




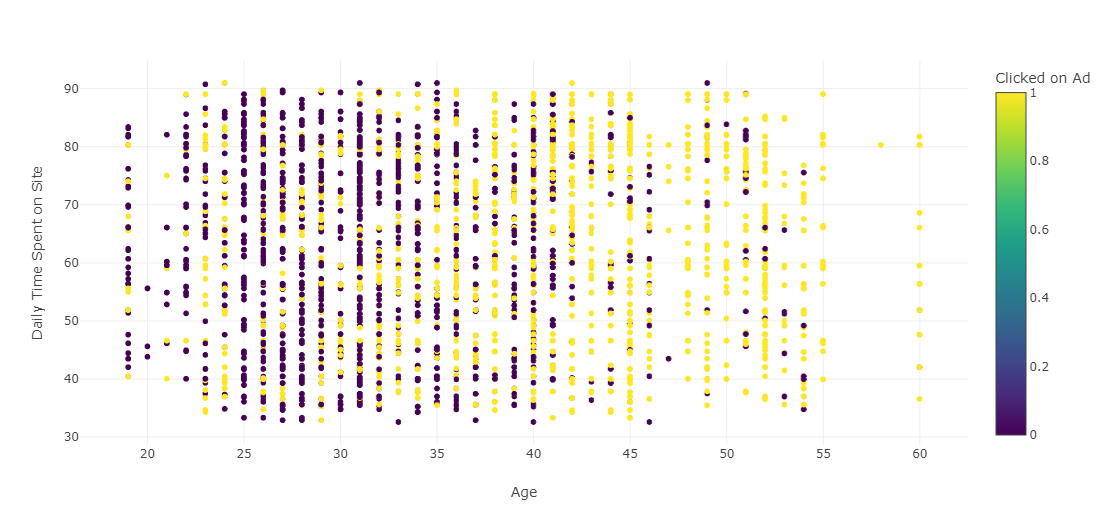




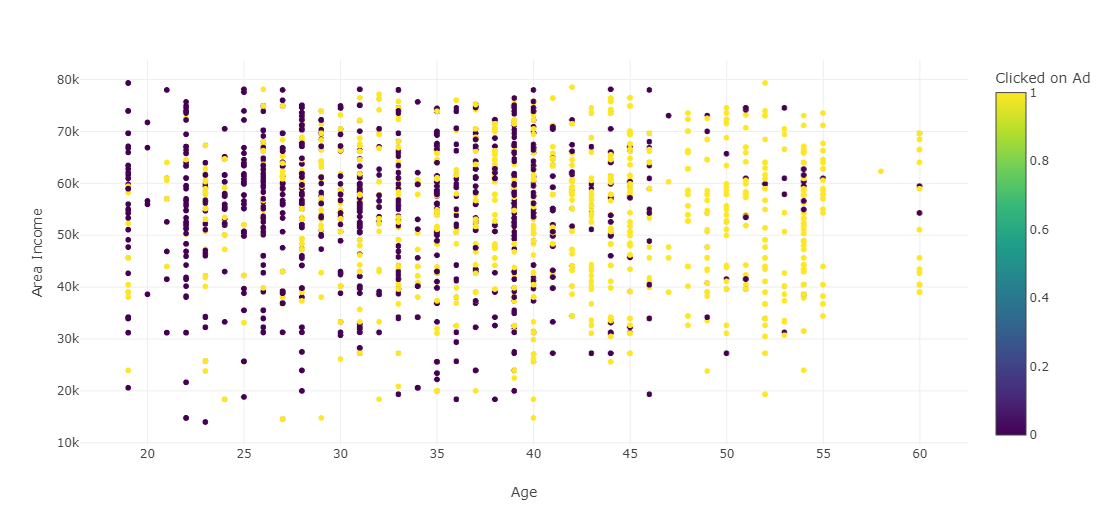




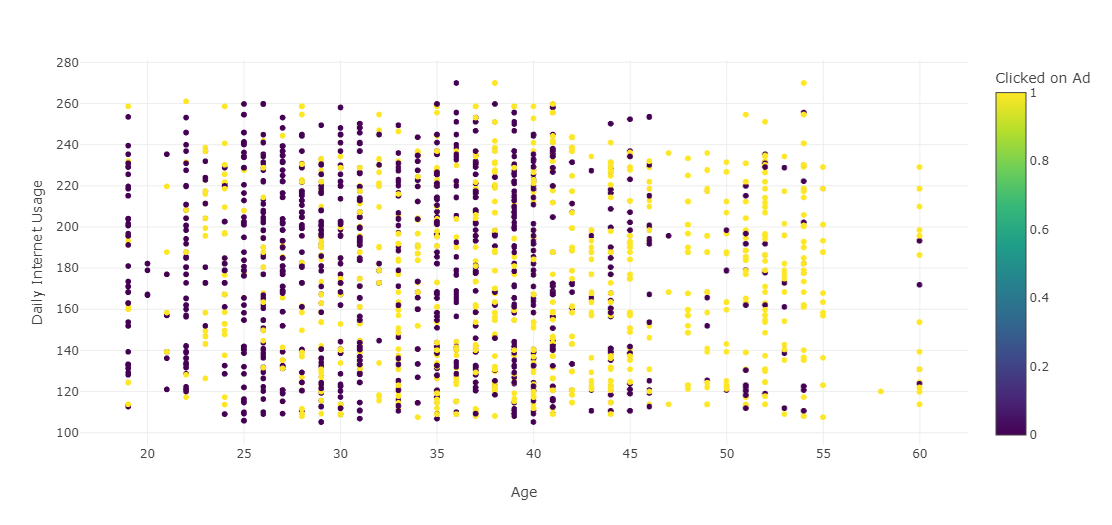
sctr



sctr1

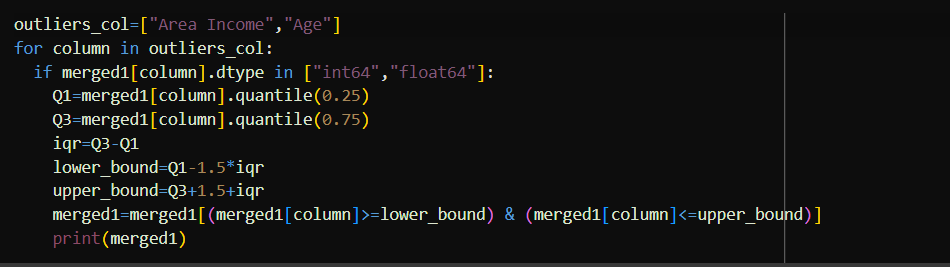


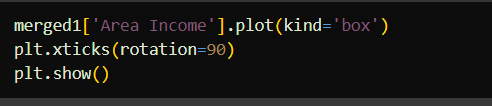
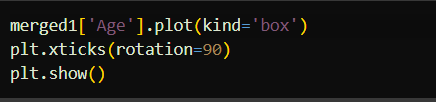
sctr2

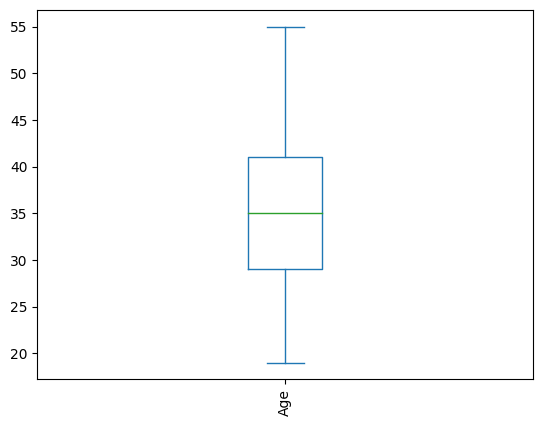
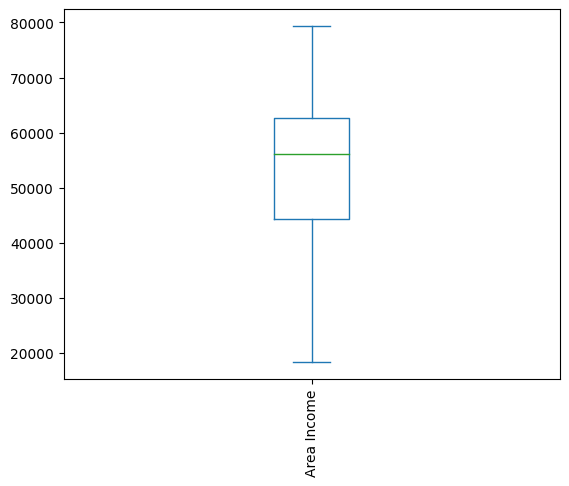


1. **Handling Outliers**

Identify and remove outliers using the Interquartile Range (IQR) method.

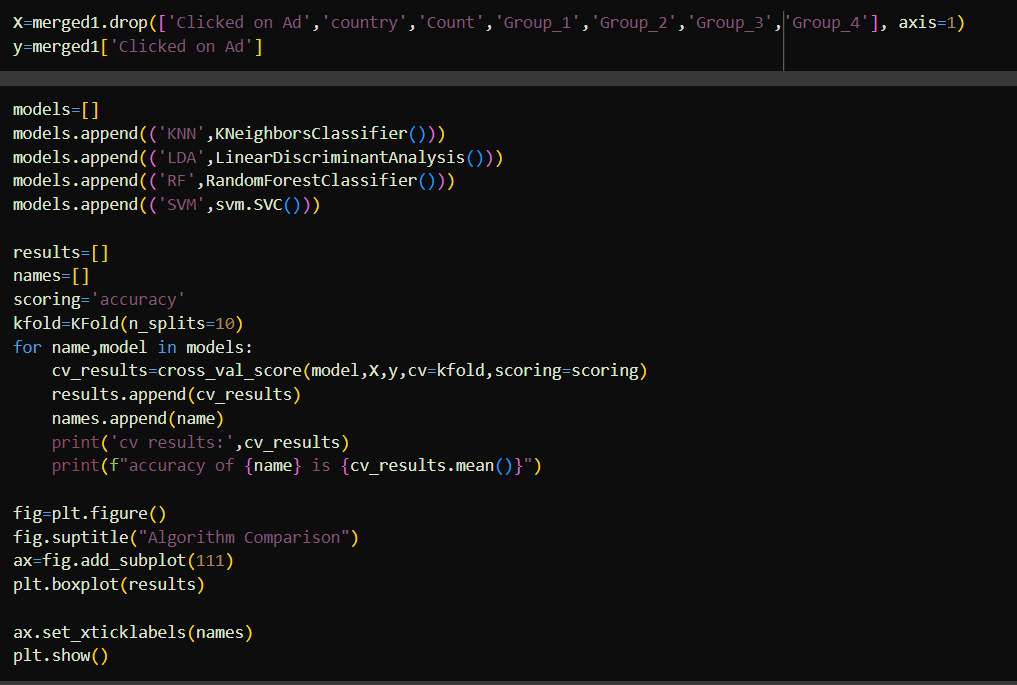


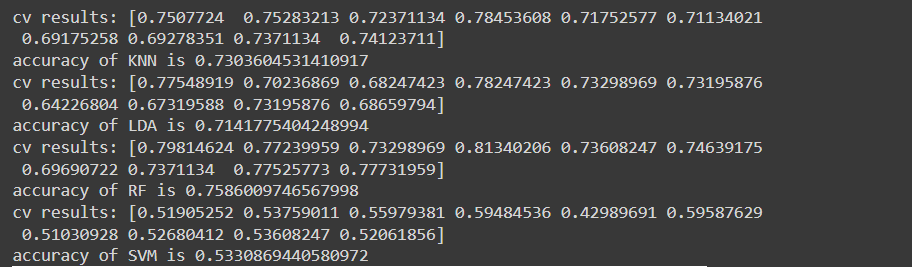
 

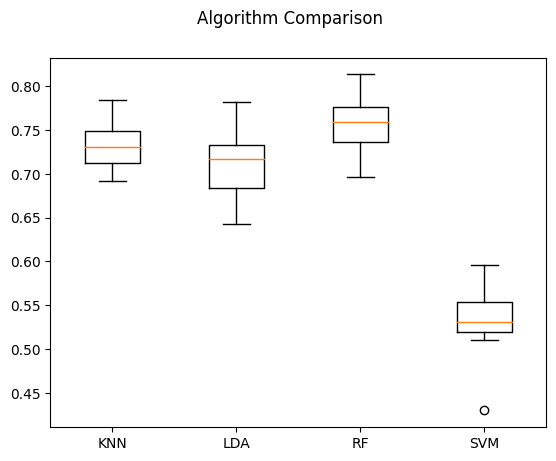
1. **Model comparison**

Here we are preparing the data for modeling and comparing the performance of different classification algorithms using cross-validation. This helps us compare the performance of different classification models on our dataset. The accuracy is used, and higher values indicate better performance. The results are printed for each model, allowing us to assess which algorithm performs best on our data.



Here we create a list of classification models **KNN**, **Linear Discriminant Analysis**, **Random forest classifier, SVM** anduse k-fold cross-validation to evaluate the models. After that we calculate the accuracy as the evaluation metric.



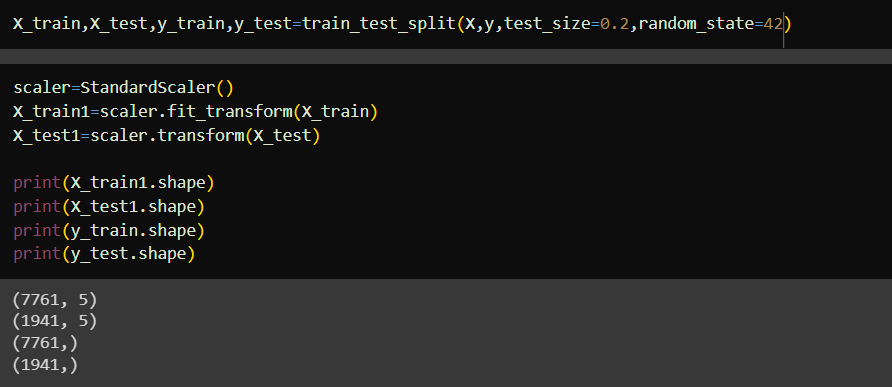


Here **Random forest classifier** has the highest accuracy value which indicates that they are the best models for our prediction.

1. **Model training and evaluation**

In this part we are preparing the data for training and testing a regression model. The steps include splitting the data into training and testing sets, and then scaling the numerical features using **StandardScaler**.

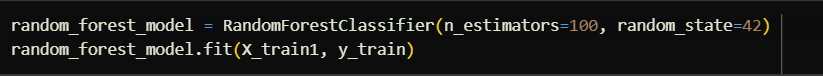
Split the data into training and testing sets. 80% of the data is used for training **X\_train**, **y\_train**, and 20% for testing **X\_test**, **y\_test**. The **random\_state** parameter ensures reproducibility.

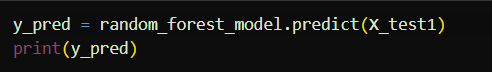
****

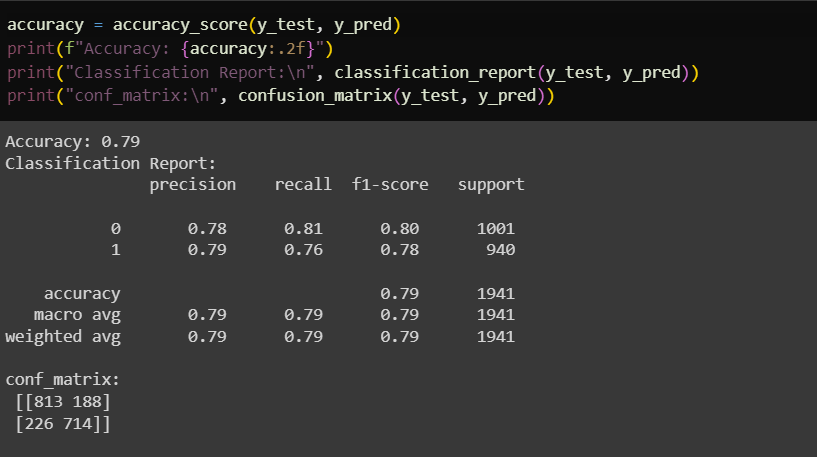
This section is crucial for preparing the data in a standardized form before feeding it into machine learning models.

1. **Prediction using Random forest classifier**

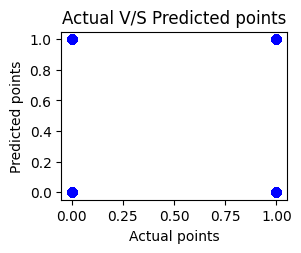
In this part, we train a Random forest classifier model on the scaled training data **X\_train1**, **y\_train** and then make predictions on the scaled testing data **X\_test1**. Subsequently, we evaluate the performance of the model using various classification metrics such as accuracy, confusion matrix and classification report.





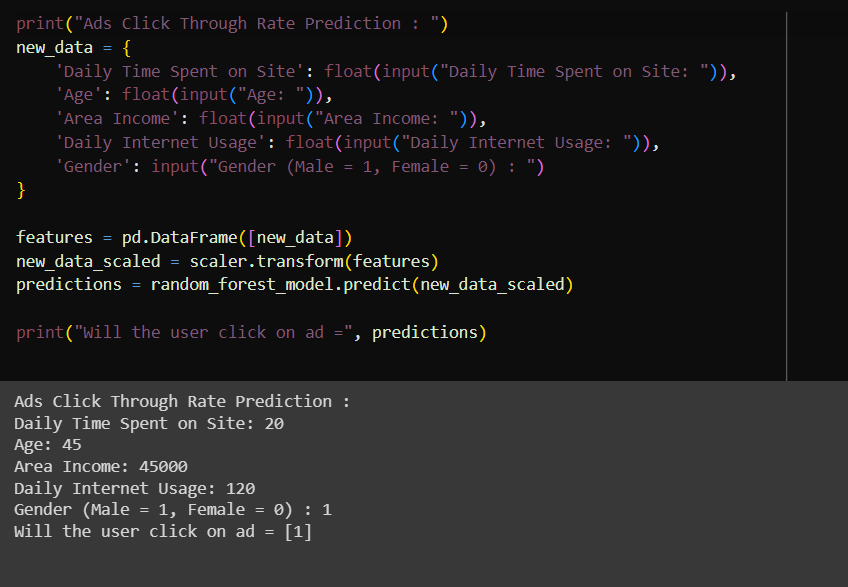


The comparison of true vs predicted values helps in visualizing how well the model predictions align with the actual outcomes.



1. **Model deployment**

Now that we have the best-performing model, it can be deployed to make the prediction. The model takes inputs such as internet usage, income, age, gender,etc and returns the prediction.



**Conclusion**

The Ads Click-Through Rate (CTR) Prediction project is a comprehensive data analysis and machine learning initiative aimed at forecasting whether a user will click on an advertisement based on various input features. The project successfully establishes a comprehensive machine learning pipeline for Ads CTR prediction, covering data preprocessing, exploratory analysis, model training, and evaluation.

Visualizations enhance the understanding of data patterns and assist in model interpretation. The Random Forest model emerges as a robust choice for predicting ad click-through rates. The ability to make predictions on new data showcases the practicality and potential deployment of the model in real-time scenarios.

In summary, the Ads CTR Prediction project provides a solid foundation for leveraging machine learning techniques to forecast user engagement with advertisements. The insights gained from the project contribute to informed decision-making in digital advertising strategies.