Decoding Consumer Behavior: Predictive Analytics of Social Media Advertising Effectiveness

Pradeep Kandula, Kankanala L S V S Nagamani Charan, Kandala Abhigna, Rasagna T, Vinaya Kamal D N

ABSTRACT:

In the digital age, social media advertising has become an indispensable tool for marketing. This study investigates the effectiveness of social media advertisements in influencing consumer purchasing decisions. Using a dataset obtained from Kaggle, this analysis incorporates demographic and socio-economic factors—specifically, age, gender, and estimated salary—to predict consumer behavior towards product purchases after exposure to advertisements. Our dataset consists of attributes such as User ID, gender, age, salary, and purchase history.

We employed several machine learning classification models including logistic regression, decision trees, random forests, SVM, Naive Bayes, KNN, and Perceptron to identify the most predictive factors and to model purchasing likelihood. Outliers were removed and missing data imputed through statistical methods to ensure data integrity. The highest predictive accuracy was achieved using a Random Forest model with 20 trees, which demonstrated a 95% accuracy rate. This model was integrated into a user-friendly interface developed with Gradio, facilitating real-time predictions. Our findings can assist companies in tailoring their marketing strategies to target demographics most likely to respond positively to advertisements, thereby enhancing the efficiency and effectiveness of their marketing campaigns on social platforms.

Keywords:- *Social Media Advertising, Consumer Behavior, Predictive Analytics, Machine Learning, Classification Algorithms, Random Forest, Marketing Strategy, Purchase Intentions, Data Preprocessing, Gradio Interface.*

I. INTRODUCTION

The digital age has transformed advertising, with social media emerging as a powerful platform for reaching a diverse and extensive audience. This paper explores the application of machine learning to predict consumer responses to social media advertisements based on age, gender, and estimated salary. Our study uses a dataset from Kaggle, which includes demographic and purchasing data, to train various classification models. The most effective model, a random forest with 20 trees, achieved an impressive 95% accuracy, underscoring its potential for predicting purchase decisions.

Our analysis not only refines predictive models in marketing but also helps companies tailor their advertising strategies to the preferences and behaviors of specific consumer segments. By integrating the predictive model into a Gradio-based interface, this study also demonstrates the practical application of our findings, making it possible for marketers to instantly predict and adapt to consumer responses. This approach bridges the gap between academic research and practical marketing needs, providing valuable insights for optimizing social media advertising strategies.

This paper aims to bridge the gap between theoretical machine learning techniques and practical marketing applications, offering valuable insights for both academic researchers and industry practitioners in the field of digital marketing and advertising analytics.

II. BASIC CONCEPTS

Pandas

The Python Pandas library is renowned for its powerful features in data manipulation and analysis. It offers various data structures and functions that facilitate effective data manipulation. Built on top of NumPy, Pandas is an essential tool for professionals in data science and analysis.

NumPy

NumPy, also referred to as Numerical Python, is extensively used for conducting scientific computations in Python. It provides a straightforward and intuitive method for handling arrays and matrices, along with a wide range of mathematical functions.

Matplotlib

Matplotlib, a popular data visualization library in Python, provides a variety of tools to create accurate and visually pleasing plots, charts, and figures. It offers extensive customization options, allowing users to create sophisticated visualizations suitable for various scientific and engineering applications. With a wide range of plotting functions, Matplotlib enables users to generate different types of charts, including line plots, scatter plots, bar plots, histograms, and more.

Sklearn

Scikit-learn, also known as Sklearn, is a popular machine learning library built on top of Python's scientific computing stack. It provides a wide range of algorithms for supervised and unsupervised learning tasks, including classification, regression, clustering, and dimensionality reduction methods. Scikit-learn is designed with a focus on usability and readability, making it accessible for users to apply machine learning techniques effectively.

Tensorflow

TensorFlow, created by the Google Brain team and provided as open source software, is a versatile machine learning library. It provides various tools and frameworks for building and deploying machine learning models for tasks like image recognition, natural language processing, and speech recognition. TensorFlow is designed to be flexible and extensible, making it well-suited for research and development purposes.

Seaborn

Seaborn is a statistical data visualization library in Python, built on top of Matplotlib. It simplifies the creation of complex visualizations with high-level functions. Seaborn offers a wide range of aesthetic choices and themes to enhance plot appearance. It excels in creating informative and visually appealing statistical graphics. With Seaborn, users can quickly generate various types of plots, such as scatter plots, bar plots, and heatmaps, with minimal code.

Gradio

Gradio is a Python library for creating interactive web-based interfaces for machine learning models. It allows users to deploy models with just a few lines of code, enabling easy sharing and collaboration. Gradio supports various input types, including text, images, and audio, making it versatile for different applications. It offers real-time inference and feedback, enhancing user interaction with deployed models.

III. Data Selection and Analysis:

- **Data Selection:**

  The dataset employed in our study is detailed and multi-faceted, capturing a broad spectrum of variables that could potentially influence consumer purchasing decisions on social media platforms.

  The attributes that are considered pivotal in determining whether an individual will make a purchase include age, gender, and estimated salary. However, it is essential to recognize that broader social and economic contexts may also impact these purchasing decisions.

  This dataset encompasses all necessary parameters and factors that could play a role in predicting consumer behavior towards advertisements.

  Here are what we consider to be the most critical attributes that determine the same: age, gender, and estimated salary, although broader socioeconomic factors also play a role.

  By incorporating a holistic approach, our study aims to provide deeper insights into the complex interplay between individual characteristics and external influences on purchasing decisions in the digital age.

- **Data description:**

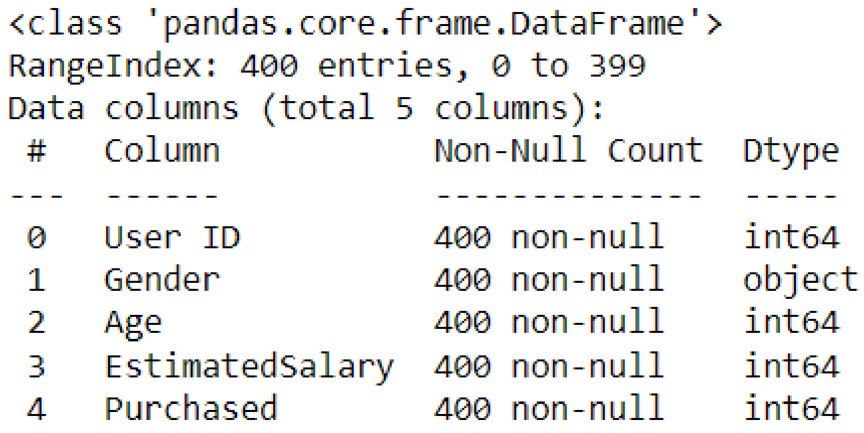
  id - Integer
  gender - Categorical: male, female or other
  age - Integer
  Estimated Salary - Integer
  Purchased - Categorical[0 or 1]

- **Model:**
  Since the output variable is purchase decision, whose values are either 0 (not purchased) or 1 (purchased), it is a binary classification problem. Therefore, we use several classification models including logistic regression, decision trees, and random forests among others.

Logistic regression is a type of regression we can use when the response variable is binary. It is particularly useful for understanding the impact of several independent variables on a single outcome variable.

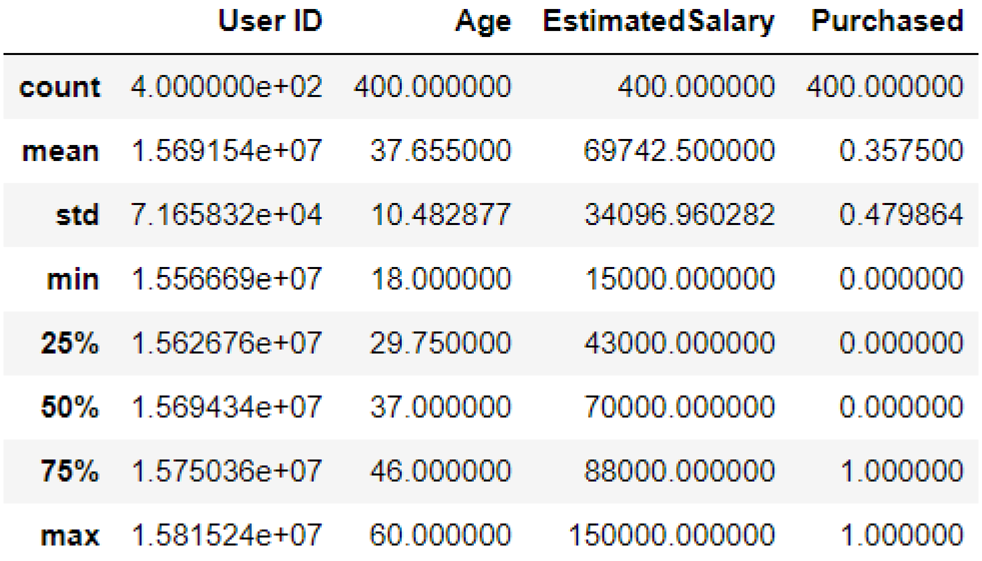
To evaluate the quality of our models, particularly logistic regression, we create a confusion matrix, which is a 2 × 2 table that shows the predicted values from the model versus the actual values from the test dataset. This matrix helps in assessing the performance of the model in terms of its sensitivity and specificity, providing clear insights into the accuracy of our predictions.

IV. Exploratory Analysis:



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 400 entries, 0 to 399
Data columns (total 5 columns):
 #   Column           Non-Null Count  Dtype 
---  ------           --------------  ----- 
 0   User ID          400 non-null    int64 
 1   Gender           400 non-null    object
 2   Age              400 non-null    int64 
 3   EstimatedSalary  400 non-null    int64 
 4   Purchased        400 non-null    int64 
```

Fig. 1: The datatypes of the variables



	User ID	Age	EstimatedSalary	Purchased
count	4.000000e+02	400.000000	400.000000	400.000000
mean	1.569154e+07	37.655000	69742.500000	0.357500
std	7.165832e+04	10.482877	34096.960282	0.479864
min	1.556669e+07	18.000000	15000.000000	0.000000
25%	1.562676e+07	29.750000	43000.000000	0.000000
50%	1.569434e+07	37.000000	70000.000000	0.000000
75%	1.575036e+07	46.000000	88000.000000	1.000000
max	1.581524e+07	60.000000	150000.000000	1.000000

Fig. 2: The Statistical computations

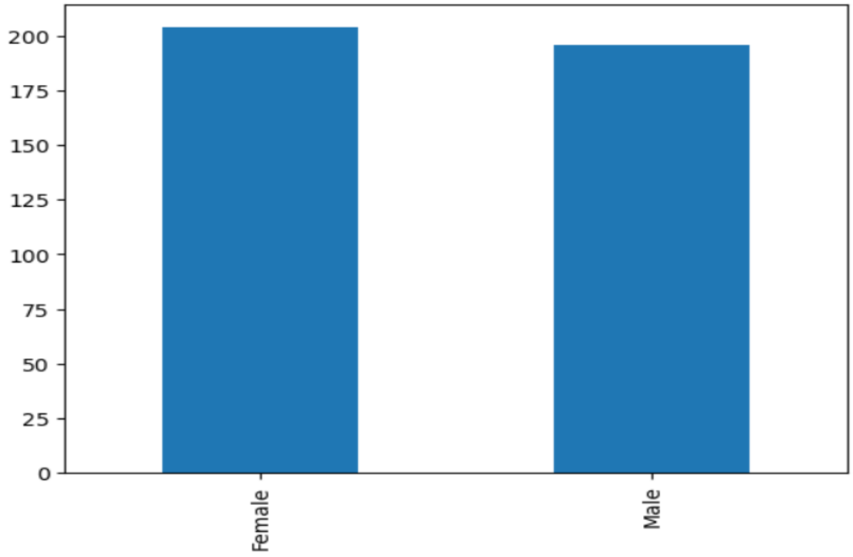


Fig. 3: Bar plots of the Gender count

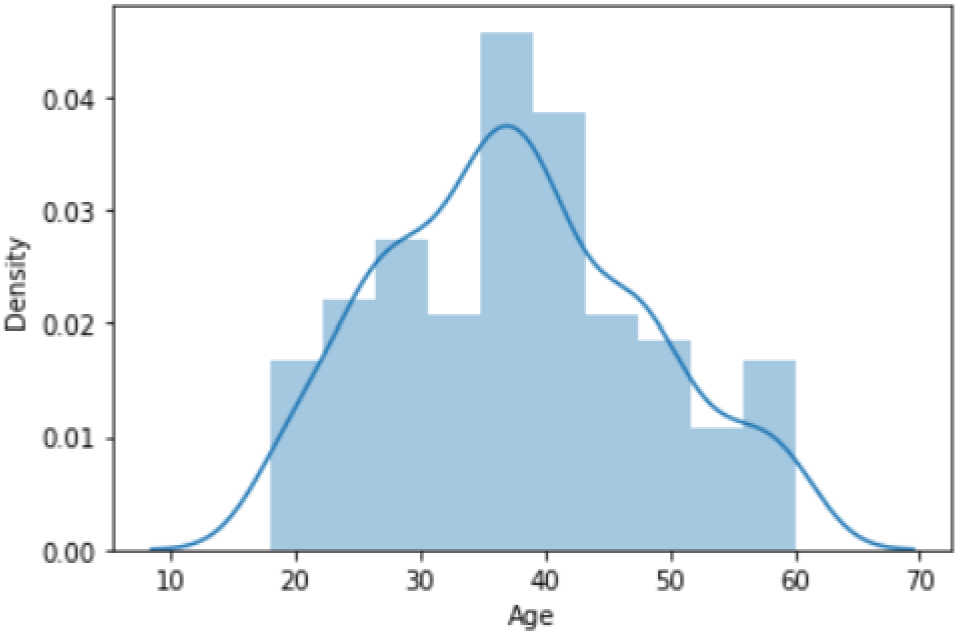


Fig. 4: Age density plot

V. Data Visualisation

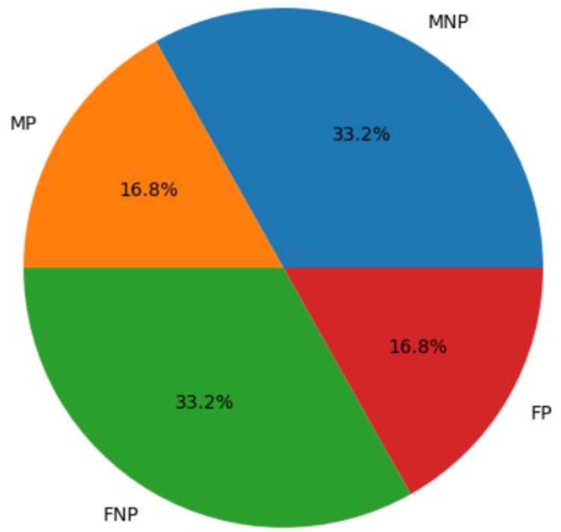


Fig. 5: pie chart of language labels

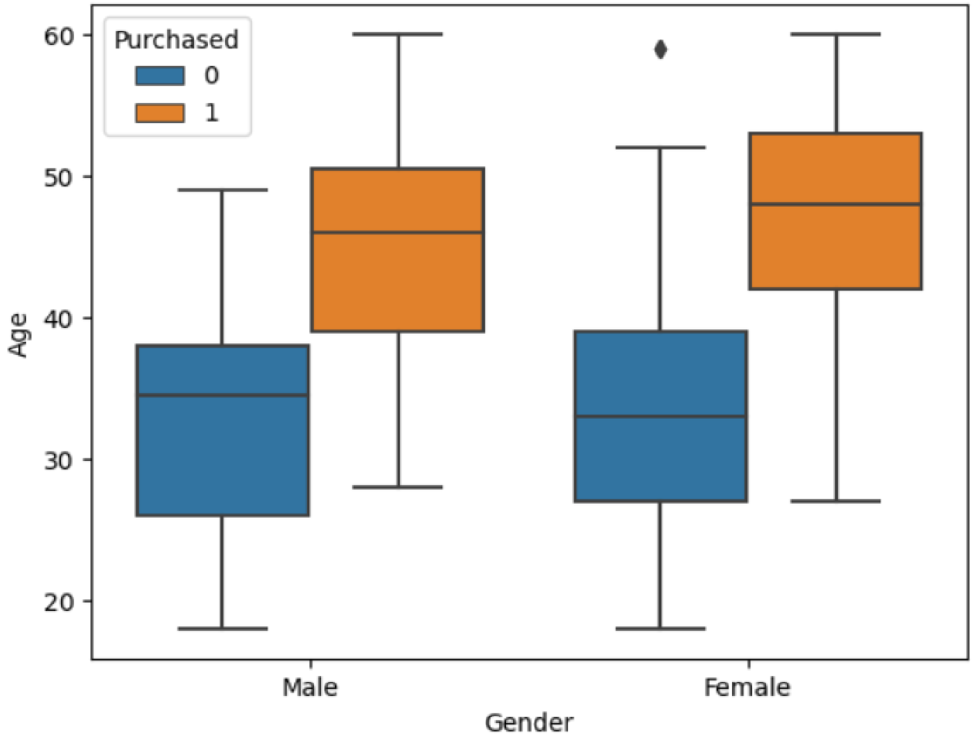


Fig. 6: Box plot

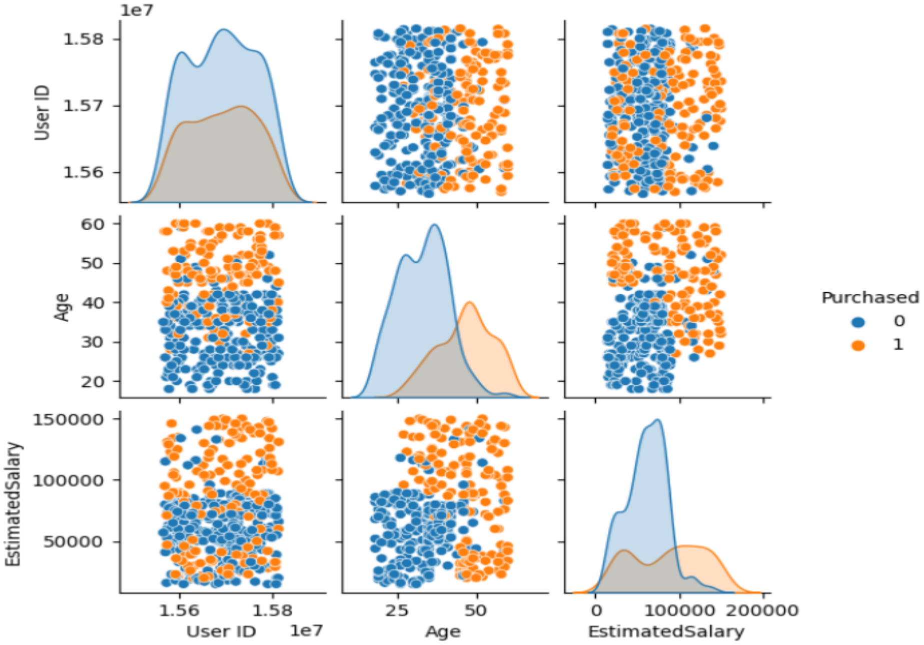


Fig. 7: pair plot

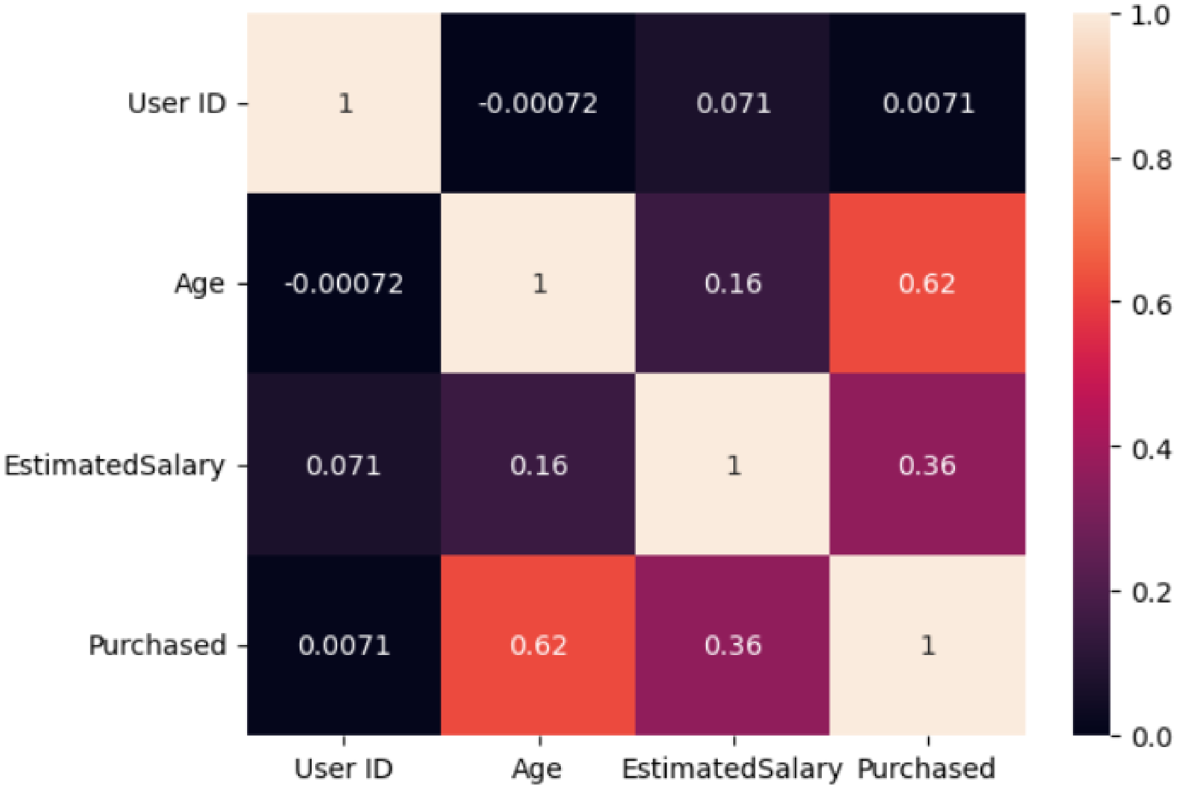


Fig. 8: Heat map

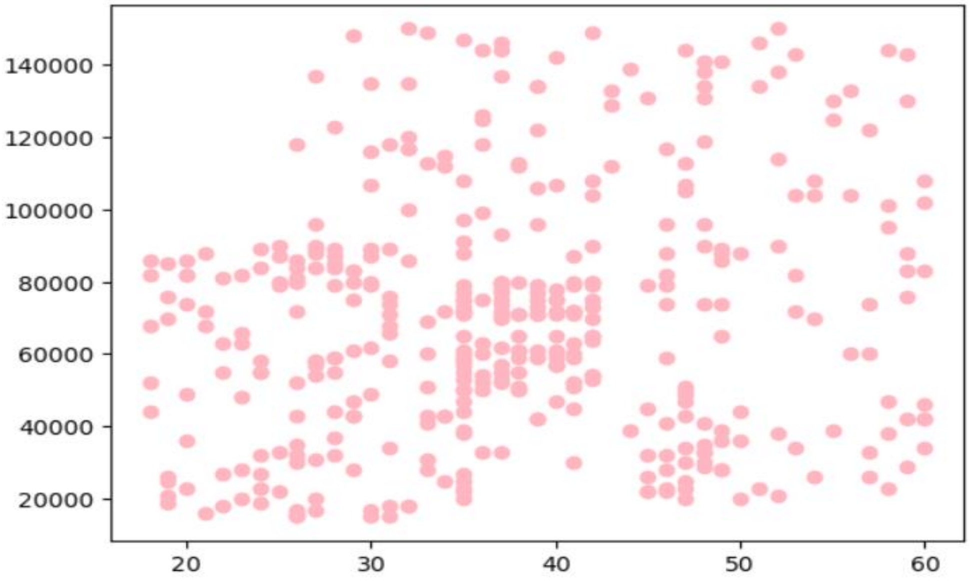


Fig. 9: scatter Plots

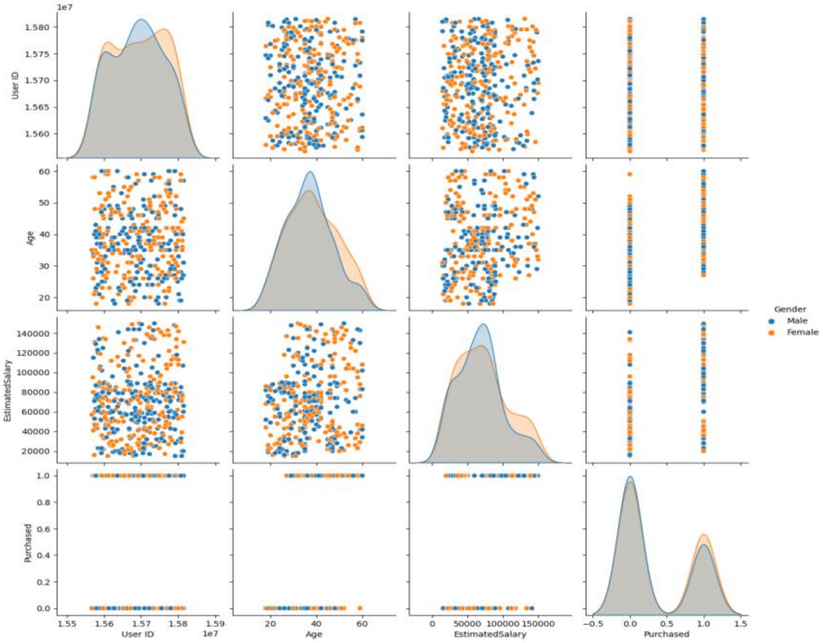


Fig. 10: Pairplot

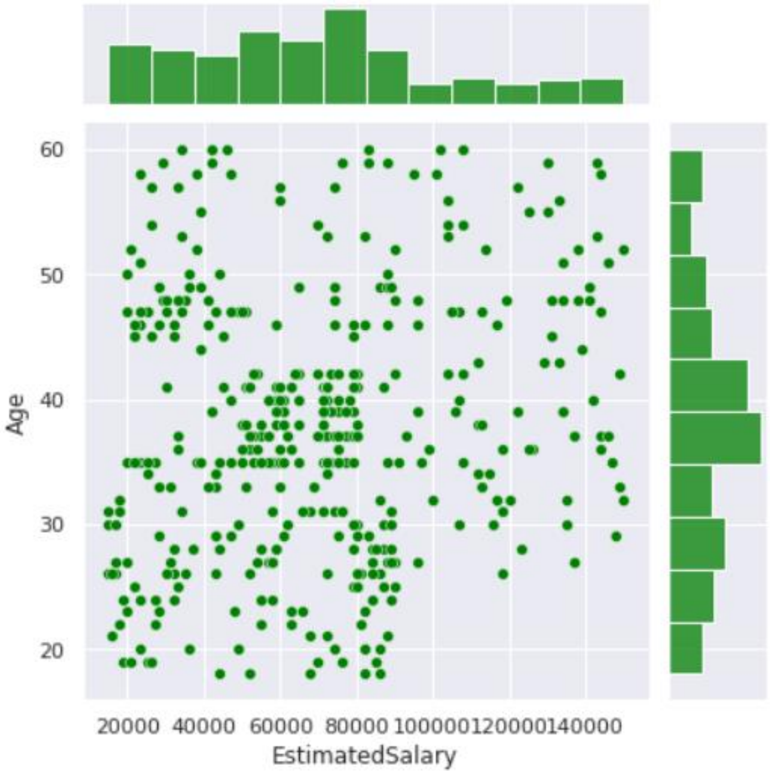


Fig. 11: Joint plot

VI.Feature Engineering:

Feature Engineering for Predictive Analysis of Social Media Advertisement Effectiveness

Feature Selection and Data Preprocessing:

For our study, the dataset included several key demographic and socioeconomic factors such as age, gender, and estimated salary. We considered these factors critical in influencing a user's decision to purchase a product after viewing an advertisement on social media. The dataset initially consisted of additional information like User ID, which was removed as it does not contribute to the model's predictive capability.

1. Handling Missing Data:
Before the model building process, we addressed the challenge of missing data in our dataset. We used multiple imputation techniques depending on the nature of the data:
- For continuous variables like estimated salary, we imputed missing values using the median of the data, given its robustness to outliers.
- For categorical data such as gender, we applied mode imputation to fill in the gaps, ensuring that the most frequent category is used as a substitute.

2. Outlier Detection and Removal:

Outliers can skew the results of predictive modeling. We employed box plots to visually inspect for outliers, particularly in the salary and age variables. Where outliers were identified, they were removed using standard deviation and IQR (Interquartile Range) methods to maintain the integrity of our predictions.

3. Feature Transformation:

We applied several transformations to better fit our models:

- Age and Salary Normalization: These features were normalized to ensure they contribute equally to the model's performance. Normalization helps in handling highly skewed data and makes the training process more stable and faster.
- Encoding Categorical Data: The gender variable was transformed using one-hot encoding, converting it into numerical values to facilitate the modeling process.

4. Feature Creation:

We explored creating new features that might capture additional nuances in the data:

- Age Groups: We categorized age into groups (e.g., 18-25, 26-35, etc.) to capture generational effects on purchasing behavior.
- Income Bracket: Salary was categorized into brackets such as low, medium, and high to simplify the model's understanding of economic status.

Feature Importance Evaluation:
Using the models such as Random Forest and Decision Trees, we also conducted a feature importance analysis to identify which variables most significantly predict the purchase decision. This analysis helped refine our feature set and focus on the most influential factors.

Model Re-evaluation:
Post-feature engineering, we re-evaluated our models to assess the impact of our changes. Improvement in model accuracy post-feature engineering indicates the effectiveness of the transformations and feature selections applied.
This detailed approach to feature engineering aimed to optimize our predictive model's performance, ensuring that it accurately reflects the influences of various factors on consumer purchasing decisions. Through this meticulous preprocessing, the model could generate more reliable and insightful results, aiding in crafting targeted marketing strategies.

VII. Models Used:

Logistic Regression:

Logistic regression is a statistical model that is commonly used for binary classification problems, where the outcome is typically binary (e.g., yes/no, 0/1). It estimates the probability of an event occurring by fitting data to a logistic curve. This is done by transforming the output using the logistic sigmoid function to return a probability value between 0 and 1. Logistic regression is easy to implement and interpret, making it a popular choice for many binary classification tasks, such as predicting whether a customer will buy a product or not based on certain predictors like age and salary.

Accuracy of the model : 0.67

Decision Tree:

A decision tree is a machine learning model used primarily for classification and regression tasks. It works by splitting the data into subsets based on the value of input features, which forms a tree-like structure of decisions. Each node in the tree represents a feature in the dataset, each branch represents a decision rule, and each leaf node represents the outcome. Decision trees are intuitive and easy to visualize, making them useful for understanding how decisions are made. They can handle both numerical and categorical data but are prone to overfitting if not properly pruned or if the tree is allowed to grow too deep.

Accuracy of the model : 0.8196666666666667

Random Tree:

Random Forest is an ensemble machine learning algorithm that operates by constructing a multitude of decision trees at training time and outputting the class that is the majority vote of the individual trees for classification tasks, or the average prediction of the trees for regression tasks. This method improves upon the standard decision tree algorithm by adding randomness in two key ways:

1. Bootstrap Aggregating (Bagging): Each tree in a Random Forest is built from a sample drawn with replacement (i.e., a bootstrap sample) from the training set. This means that each tree is built on slightly different data. Some observations may be repeated in one sample and absent in another, making the trees diverse.

2. Feature Randomness: When splitting a node during the construction of a tree, the choice of the split is no longer the best among all features. Instead, the split that is picked is the best among a random subset of the features. This ensures that the trees in the forest are de-correlated, making the average of the ensemble less variable and hence more reliable.

The combination of these two mechanisms helps Random Forest to perform better than a single decision tree, making it less prone to overfitting while maintaining high accuracy.
Random Forests can be used for both classification and regression tasks and they are easy to use because they require very few tuning parameters and can perform quite well with their default settings.

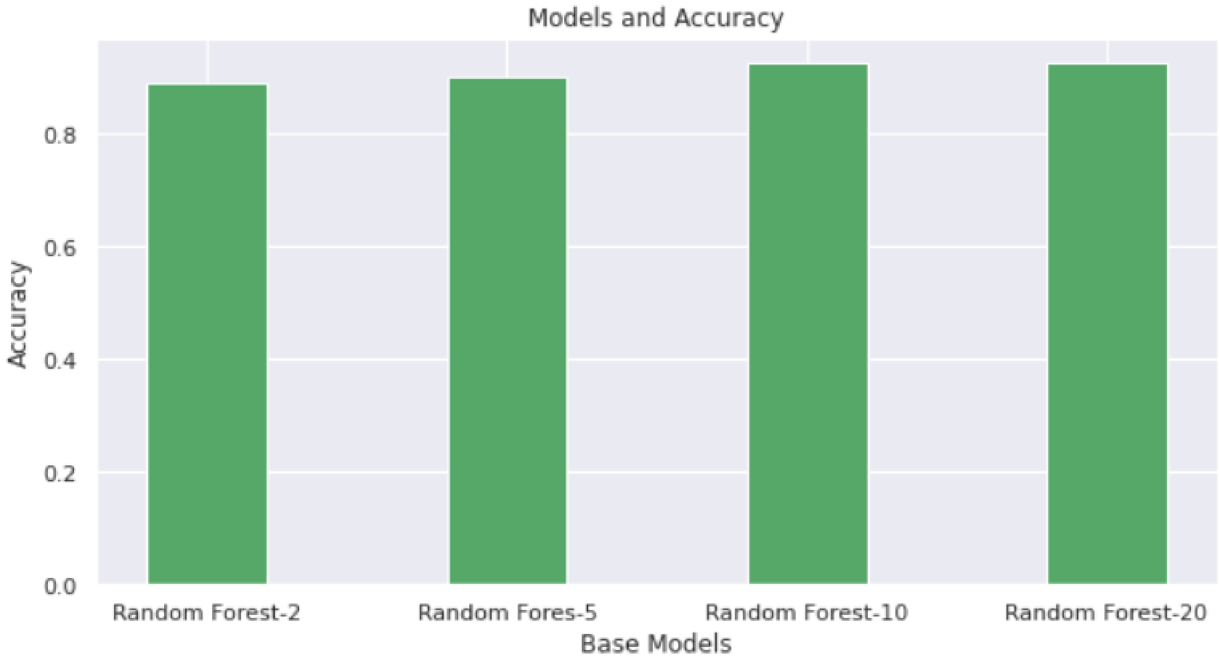


Fig. 12: Models and Accuracy

Accuracy of the model with 2 trees : 0.8916666666666667

Accuracy of the model with 5 trees : 0.9

Accuracy of the model with 10 trees : 0.9416666666666667

Accuracy of the model with 20 trees : 0.925

Support Vector Machine:

A Support Vector Machine (SVM) is a powerful supervised machine learning model used primarily for classification and regression tasks. It works by finding a hyperplane that best separates different classes in a high-dimensional space. SVM enhances classification accuracy by maximizing the margin between data points of different classes, using kernels to handle linear and non-linear separations effectively. This makes SVM highly effective for complex datasets with clear class delineation.

Support Vector Machine (SVM) is particularly favored for its ability to handle non-linear data through the use of kernel functions. This flexibility allows it to perform well across various applications, from image recognition to bioinformatics. While SVM can be computationally intensive due to its requirement for parameter tuning and support for large datasets, its high accuracy and robustness against overfitting make it a preferred choice in environments where precision is critical.
Accuracy of the model : 0.7833333333333334

Gaussian Naive Bayes:

Gaussian Naive Bayes is a variant of Naive Bayes that is especially useful for continuous data and assumes that the features follow a normal distribution. This model applies Bayes' theorem, with the naive assumption of conditional independence between every pair of features given the class label. Gaussian Naive Bayes is straightforward to implement, requires a small amount of training data to estimate the necessary parameters, and is highly effective for a wide range of classification tasks.

Gaussian Naive Bayes is well-suited for high-dimensional data due to its simplicity and efficiency. It excels in text classification and medical diagnosis, despite the independence assumption potentially limiting its accuracy. This model is favored for its quick application and robust performance, especially when data aligns well with Gaussian distributions, making it a practical choice for initial model testing and rapid analytics.
Accuracy of the model : 0.9166666666666667

K- Nearest Neighbours:

The K-Nearest Neighbors (KNN) algorithm is a straightforward and effective non-parametric method used for classification and regression. It classifies a data point based on the majority class of its nearest neighbors, with the crucial parameter being the number of neighbors, k. KNN is easy to implement and understand, making it suitable for situations where model transparency is important.

However, KNN struggles with efficiency in large datasets as it requires computing the distance to every other point, leading to high computational costs and slower prediction times. It is also sensitive to noise and data imbalance, which can affect its accuracy. Despite these challenges, KNN is favored for its interpretability and effectiveness in smaller datasets where computational demands are manageable.

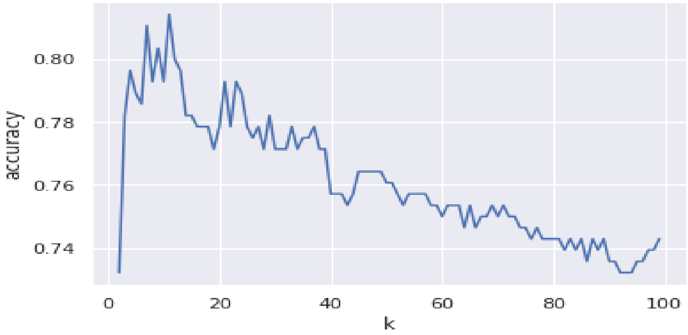


Fig. 13: k vs Accuracy

K value for max accuracy = 11
Accuracy of the model : 0.825

Preceptron:

The Perceptron is a simple type of artificial neural network used as a linear classifier for binary classification tasks. It combines input features with weights and a bias to produce an output; if the weighted sum exceeds a threshold, it outputs one class, otherwise, it outputs the alternative. The Perceptron adjusts its weights based on prediction errors during training, continuing until it achieves satisfactory accuracy or reaches a set number of iterations. Due to its straightforward mechanism, the Perceptron works well with linearly separable data but is inadequate for non-linear datasets, where more complex models are needed. Despite these limitations, its simplicity makes the Perceptron valuable for educational purposes and initial binary classification tasks.

Accuracy of the model : 0.6666666666666667

K- Means Clustering:

K-means clustering is an unsupervised learning algorithm that efficiently partitions a dataset into K distinct, non-overlapping clusters. It begins by randomly selecting K points as initial centroids and iterates through two main steps: assignment and update. During the assignment phase, each data point is assigned to the nearest centroid based on Euclidean distance. In the update phase, centroids are recalculated as the mean of all points in their cluster, ensuring clusters are as homogenous as possible.

This process repeats until the centroids stabilize, signifying convergence. K-means is favored for applications like market segmentation and document clustering due to its simplicity. However, it requires pre-specifying the number of clusters and can be sensitive to initial centroid placements, often necessitating multiple runs to achieve optimal clustering. Despite these challenges, K-means remains a popular choice for data analysis tasks requiring quick and straightforward cluster identification.

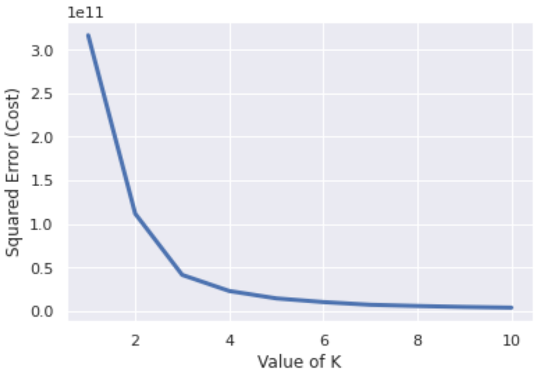


Fig. 14: Value of K and Squared Error

Number of clusters : 3

Accuracy of the model : 0.675

Agglomerative clustering:

Agglomerative clustering is a type of hierarchical clustering used to group objects in a dataset into clusters based on their similarity. It's a "bottom-up" approach where each data point initially represents a single cluster, and pairs of clusters are successively merged until a desired structure is achieved. This algorithm works by calculating the distance between every pair of clusters using various metrics such as Euclidean or Manhattan distance and merging the closest pairs. Popular linkage criteria for defining the distance between clusters include single linkage (minimum distance), complete linkage (maximum distance), average linkage (average distance), and Ward's method (minimum variance).

The process iterates by updating the distance matrix after each merger and continues until all points are merged into a single cluster or until a specified number of clusters is reached. Agglomerative clustering is particularly useful for data that exhibits a hierarchical structure and is widely used in fields such as biological data analysis, image segmentation, and document clustering. However, its computational complexity can be a limitation for very large datasets, and the choice of linkage criteria can significantly affect the outcome, making it essential to match the criterion with the specific characteristics of the data.

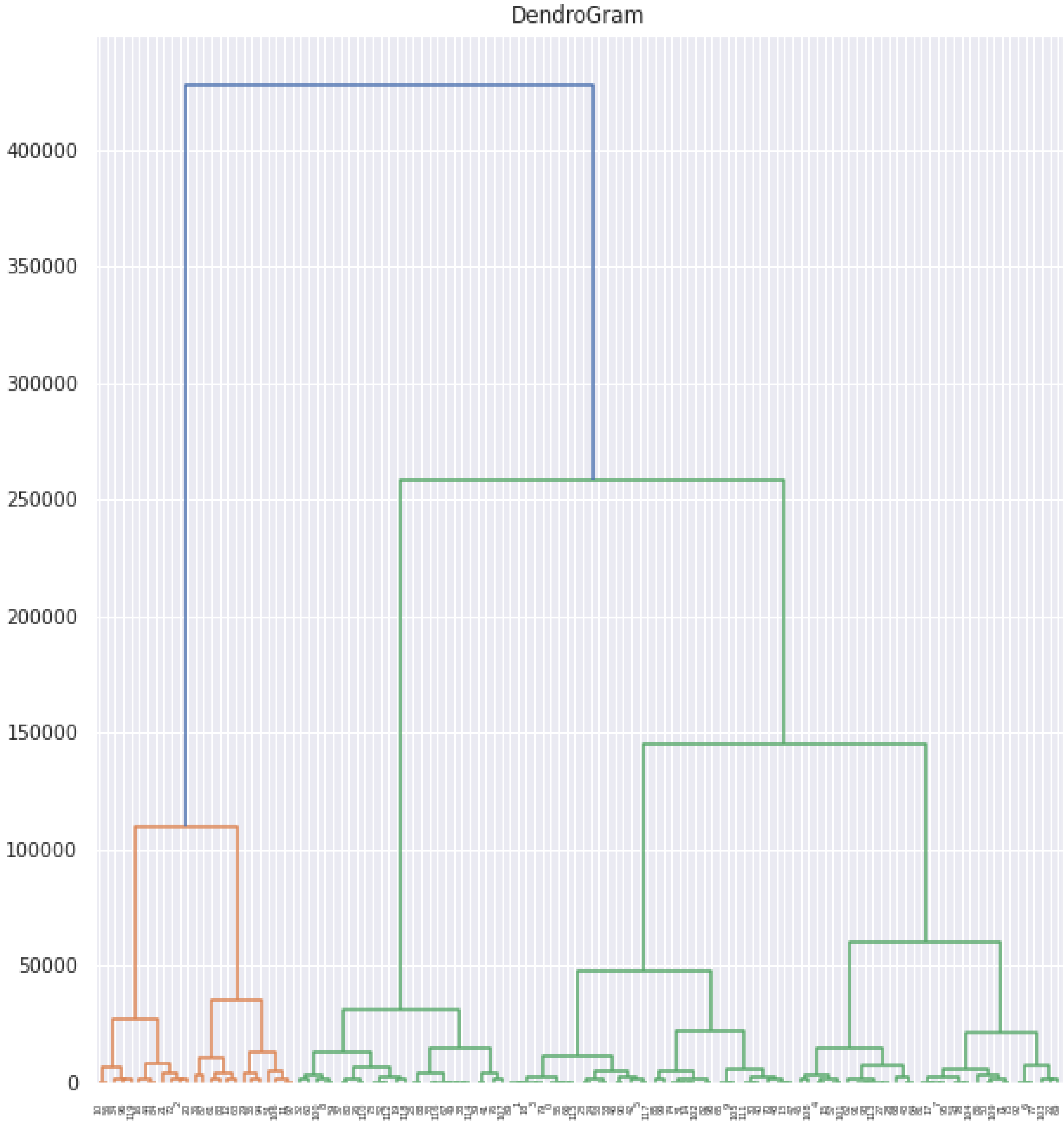
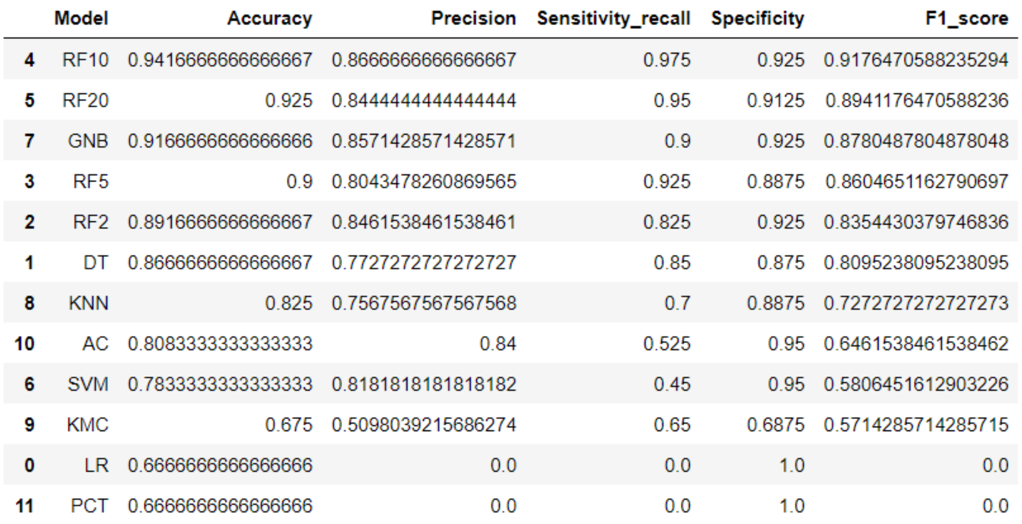


Fig. 15: Dendrogram

Accuracy of the model : 0.8083333333333334

VIII. Combined Analysis:



	Model	Accuracy	Precision	Sensitivity_recall	Specificity	F1_score
4	RF10	0.9416666666666667	0.8666666666666667	0.975	0.925	0.9176470588235294
5	RF20	0.925	0.8444444444444444	0.95	0.9125	0.8941176470588236
7	GNB	0.9166666666666666	0.8571428571428571	0.9	0.925	0.8780487804878048
3	RF5	0.9	0.8043478260869565	0.925	0.8875	0.8604651162790697
2	RF2	0.8916666666666667	0.8461538461538461	0.825	0.925	0.8354430379746836
1	DT	0.8666666666666667	0.7727272727272727	0.85	0.875	0.8095238095238095
8	KNN	0.825	0.7567567567567568	0.7	0.8875	0.7272727272727273
10	AC	0.8083333333333333	0.84	0.525	0.95	0.6461538461538462
6	SVM	0.7833333333333333	0.8181818181818182	0.45	0.95	0.5806451612903226
9	KMC	0.675	0.5098039215686274	0.65	0.6875	0.5714285714285715
0	LR	0.6666666666666666	0.0	0.0	1.0	0.0
11	PCT	0.6666666666666666	0.0	0.0	1.0	0.0

Fig. 16: Metrics of all models

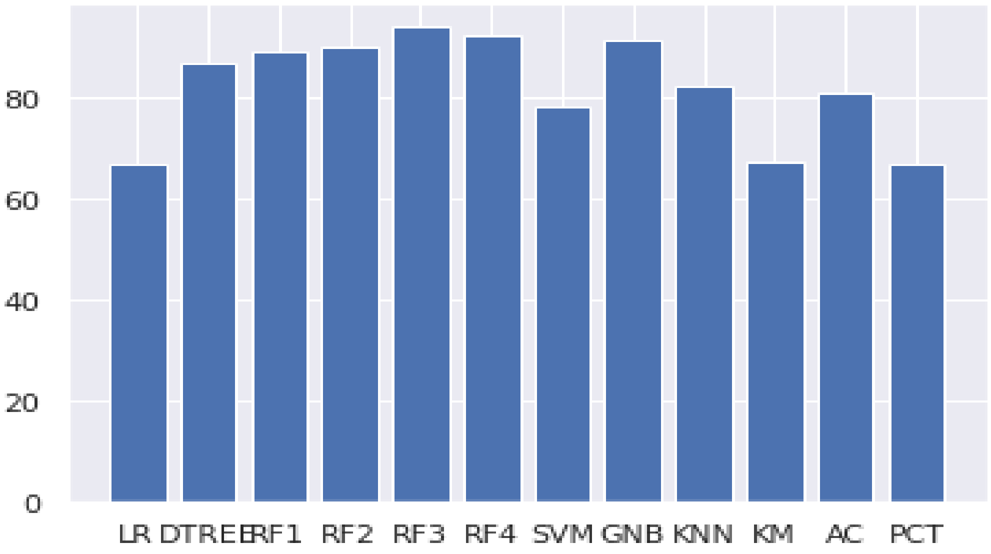


Fig. 17: Metrics of all models

VIII. Conclusion:

This research has explored the use of machine learning techniques to predict consumer purchasing decisions based on social media advertising, utilizing demographic and socioeconomic factors such as age, gender, and estimated salary. Through comprehensive data analysis and the application of various classification models, we have demonstrated that these factors significantly influence consumer behavior.

Our findings indicate that the random forest model, particularly with 10 trees, provided the highest accuracy at 94.1%. This robust model not only underscores the potential of machine learning in marketing analytics but also highlights the precision with which these tools can predict consumer responses to advertisements. The integration of the model into a Gradio interface for real-time predictions further emphasizes the practical application of our research in real-world settings, allowing for dynamic interaction and accessibility to users.

This study contributes to the growing field of digital marketing by providing insights that can help companies tailor their advertising strategies more effectively. By understanding which demographic groups are more likely to respond to certain types of advertisements, marketers can allocate resources more efficiently and increase the overall effectiveness of their campaigns.

However, while the results are promising, they also underscore the need for continuous improvement and expansion in research methodologies. Future work should consider integrating additional data points, exploring advanced predictive models, and addressing ethical considerations related to consumer privacy and data security. Such efforts will enhance the capability of predictive models and ensure they are used responsibly and effectively in the evolving landscape of digital marketing.

In conclusion, the intersection of machine learning and marketing presents an exciting frontier for both technological innovation and strategic business applications. Our research not only adds to the academic literature but also serves as a guide for practitioners seeking to leverage the power of data analytics in enhancing the impact of social media advertising.

IX. Future Scope:

Future Scope of Research in Predictive Analysis of Social Media Advertising Effectiveness

The research presented in this paper opens several avenues for further exploration and enhancement in the domain of predictive analytics for social media advertising. Here are some potential directions for future research:

1. Integration of Additional Data Sources:
Future studies could incorporate more diverse data sources, such as social media engagement metrics (likes, comments, shares), browsing history, and interaction patterns. These data points could provide deeper insights into user behavior and preferences, enabling more nuanced and accurate predictions.

2. Advanced Machine Learning Models:
While traditional models like logistic regression and random forests have proven effective, there is scope to explore advanced machine learning and deep learning techniques. Models such as neural networks or ensemble methods that combine several algorithms could potentially improve accuracy and robustness.

3. Real-Time Prediction Systems:
Developing real-time predictive models that can adjust and respond as new data comes in can significantly enhance the responsiveness of advertising strategies. Implementing streaming analytics to process data in real-time and immediately adjust ad targeting strategies could be a game-changer.

4. Cross-Platform Advertising Analysis:
Exploring the effectiveness of advertisements across different social media platforms can provide insights into how platform-specific features and user demographics influence ad performance. Comparative studies across platforms like Facebook, Instagram, Twitter, and LinkedIn could help optimize cross-platform advertising strategies.

5. Sentiment Analysis and Natural Language Processing (NLP):
Incorporating sentiment analysis to gauge the mood and opinions expressed in user comments and feedback on social media can provide additional layers of understanding about consumer attitudes. NLP techniques can be employed to analyze text data from social media posts to better understand the contexts in which products are discussed.

6. Ethical and Privacy Considerations:
As data analytics ventures deeper into personalization and targeted advertising, it becomes crucial to address privacy concerns and ethical implications. Future research should also focus on developing models that respect user privacy and comply with evolving data protection laws.

7. Economic Impact Studies:
Further research could also explore the economic impacts of predictive advertising, assessing not just direct sales impacts but broader economic considerations such as brand loyalty, consumer trust, and long-term engagement.

8. User-Centric Models:
Developing models that focus on user satisfaction and experience rather than just sales metrics could help in creating more sustainable business practices. This approach would align more closely with modern marketing ethics, which emphasize customer centricity.

9. Automated AI-driven Marketing Tools:
Research into fully automated AI-driven tools that can not only predict but also autonomously execute marketing strategies based on predictive insights could revolutionize the field. This could include automated content creation, dynamic ad placement, and self-optimizing campaigns.

10. Longitudinal Studies:
Conducting longitudinal studies to observe how consumer behavior changes over time in response to advertising strategies could provide insights into the long-term effectiveness of different marketing tactics.

By exploring these areas, future research can extend the boundaries of current knowledge on predictive analytics in advertising, leading to more effective and efficient marketing strategies that are responsive to consumer needs and market dynamics.

References:

1. kaggle.com/datasets/akram24/social-network-ads/data
2. Witten, I. H., & Frank, E. (2005). Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann
3. Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53-65
4. Han, J., Pei, J., & Kamber, M. (2011). Data Mining: Concepts and Techniques. Morgan Kaufmann.