International Summer School

Data Mining for Business Intelligence

Predicting Purchase Behavior in Social Network Ads

Amir Ali

Supervised By:
Prof Dr.Zenun Kastrati
Dr.Endrit Fetahu

Faculty of Computer Science

August 10, 2023

Abstract

This report explores the prediction of purchase behavior within the realm of social network ads, focusing on the influence of age, gender, and salary. Leveraging a dataset encompassing these demographic attributes, we conducted a thorough analysis employing various classification algorithms including Naive Bayes, Decision Trees, k-Nearest Neighbors, Support Vector Machines, and Multilayer Perceptrons. The examination encompasses data preprocessing, visualization, model implementation, and evaluation, providing insights into algorithm performance across accuracy, precision, recall, and f1-Score metrics. Our findings reveal the interplay between demographic factors and purchase decisions, showcasing the predictive capabilities of different algorithms. By understanding the strengths and limitations of each approach, this study equips businesses with actionable insights to optimize their social network ad campaigns for targeted and impactful engagement

Contents

1 Introduction

In today's ever-evolving landscape of digital marketing, social media platforms have become more than just virtual spaces for social interaction—they have evolved into powerful arenas for targeted advertising. Social network ads, featured prominently on platforms such as Facebook, Instagram, Twitter, and LinkedIn, have redefined the way businesses and organizations connect with their potential customers. These advertisements offer a unique opportunity to engage audiences based on a plethora of factors, including age, gender, location, interests, and behaviors, thus creating tailored campaigns that resonate deeply with individual preferences .

In this context, a critical question arises: Can we predict whether an individual will convert and make a purchase through a social network ad, relying on their age, gender, and salary as guiding factors? This query underscores the importance of leveraging data-driven insights to refine marketing strategies. By delving into predictive analysis, businesses can not only optimize their advertising endeavors but also gain a nuanced comprehension of the intricate dynamics that steer consumer actions on social media platforms .

This analysis aims to contribute to this discourse by employing a suite of classification algorithms, each playing a distinctive role in discerning patterns within the dataset. The algorithms we intend to implement include Naïve Bayes (NB), Decision Tree, k-Nearest Neighbors (kNN), Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP). These methodologies, renowned for their diverse strengths in classification tasks, will collectively enable us to construct predictive models capable of anticipating the likelihood of a purchase, predicated on the provided demographic information [1].

By exploring the performance of these algorithms, we seek to not only ascertain their efficacy in predicting purchasing behavior but also identify which method excels in this specific context. This endeavor carries profound implications for businesses aspiring to optimize their social network advertising strategies, offering insights into the interplay between demographic attributes and the potential outcomes of these targeted campaigns.

Subsequent sections of this report will delve into the distinct characteristics of the dataset, the methodologies employed, the results obtained through the application of each algorithm, and the consequential implications for businesses aiming to refine their marketing strategies. Through this comprehensive exploration, we aim to unravel the intricate relationship between demographics, consumer behavior, and the efficacy of social network ads, ultimately providing a robust foundation for future marketing campaigns.

2 Literature Review

The paper titled "An Overview of Digital Media and Advertising" by Dickey, I. J. and Lewis, W. F. [2] offers a clear and thorough look into social media advertising. This paper explains different types of ads used on platforms like social media: from regular display ads to sponsored posts and influencer marketing. The authors aim to help readers understand

how businesses promote themselves on social media, using these various ad formats to reach people effectively. They discuss how ads appear on social media, how sponsored posts work, and the rising trend of influencer marketing, where popular individuals help endorse products. Overall, this paper gives a simple but comprehensive view of how digital media is used for advertising, especially on social media platforms.

The paper titled "Targeting Consumers on Social Media: An Analysis of Facebook Advertising" by Ferreira, F. and Barbosa, B. [3] looks closely at how Facebook ads work to reach specific people. The authors study the tools Facebook provides for targeting certain groups, like using age, location, and interests. The paper aims to help us understand how businesses can use these tools to make their ads more effective and connect better with the right audiences on social media.

3 Data Collection

We got our data for this study from the Kaggle [4]. This data focuses on social network ads and helps us understand how people react to them. The dataset includes information like whether someone made a purchase after seeing an ad, their gender, age, and salary. Here's how the information looks in a simplified table format:

Customer ID	Gender	Age	Salary	Purchase
1	Male	25	30000	Yes
2	Female	30	50000	No
3	Male	35	75000	Yes
...	...	...	...	...

Table 1: Dataset Attributes

In this table, each row represents an individual. "Customer ID" is a unique identifier for each person. "Gender" shows whether they are male or female. "Age" is their age in years. "Salary" indicates their yearly salary in a certain currency. Finally, "Purchase" tells us whether they made a purchase after seeing the ad, with "Yes" or "No" as the values. This data will help us analyze how different factors like age, gender, and salary influence people's decision to purchase products advertised on social networks.

4 Exploratory Data Analysis

4.1 Gender Distribution

After analyzing the "Gender" column in the dataset, it was determined that the distribution of genders is fairly equal, with a similar number of males and females represented. This gender balance is substantiated by both the pie chart and the histogram plot see the figure 1, where the segments or bars corresponding to "Male" and "Female" are approximately the same size. This signifies that the dataset contains a comparable number

of individuals from both genders, which is a favorable condition for conducting unbiased analyses and drawing conclusions applicable to both male and female participants.

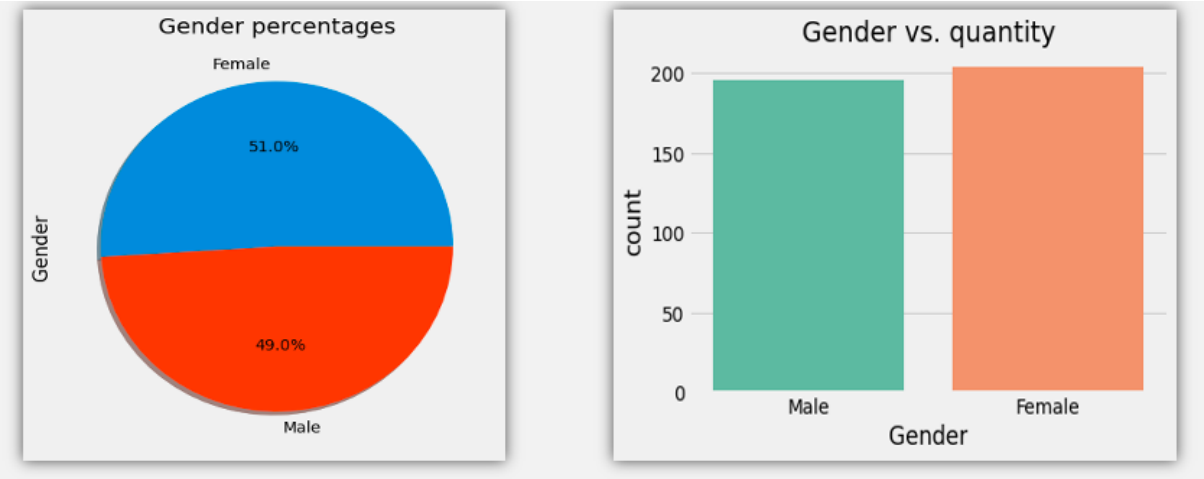

Figure 1: Gender Distribution

4.2 Age Distribution

The age distribution was visualized using a Distplot see the figure 2 , showcasing the distribution pattern of participants' ages within the dataset. The curve of the Distplot indicates that the age data follows a relatively normal distribution centered around 37 years. This implies that the dataset is populated with a significant number of individuals around this age range. Notably, a prominent concentration of participants falls within the range of 35 to 45 years, signifying a demographic cluster where most of the individuals studied are situated. This insight sheds light on the predominant age group within the dataset and provides valuable context for understanding the age dynamics influencing subsequent analyses.

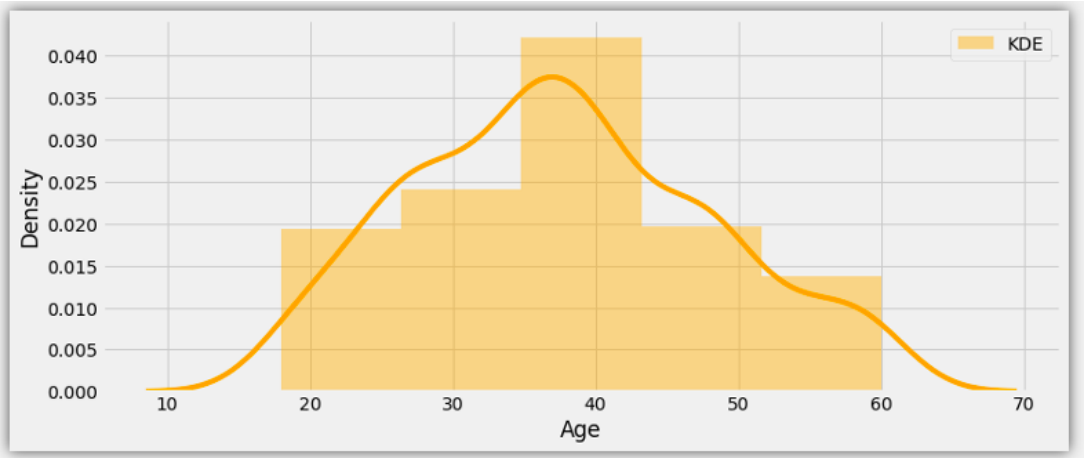

Figure 2: Age Distribution

4.3 Relationship between Age and Salaries

The scatter plot was employed to explore the potential relationship between participants' ages and their corresponding salaries. Despite initial expectations, the scatter plot see the figure 3 reveals a lack of discernible correlation between age and salary. Points are spread across the plot without exhibiting a clear trend or direction. This suggests that within the dataset, no apparent linear relationship exists between age and salary. As such, the age of an individual does not appear to consistently predict their salary, highlighting the importance of considering other factors that might contribute to variations in salary.

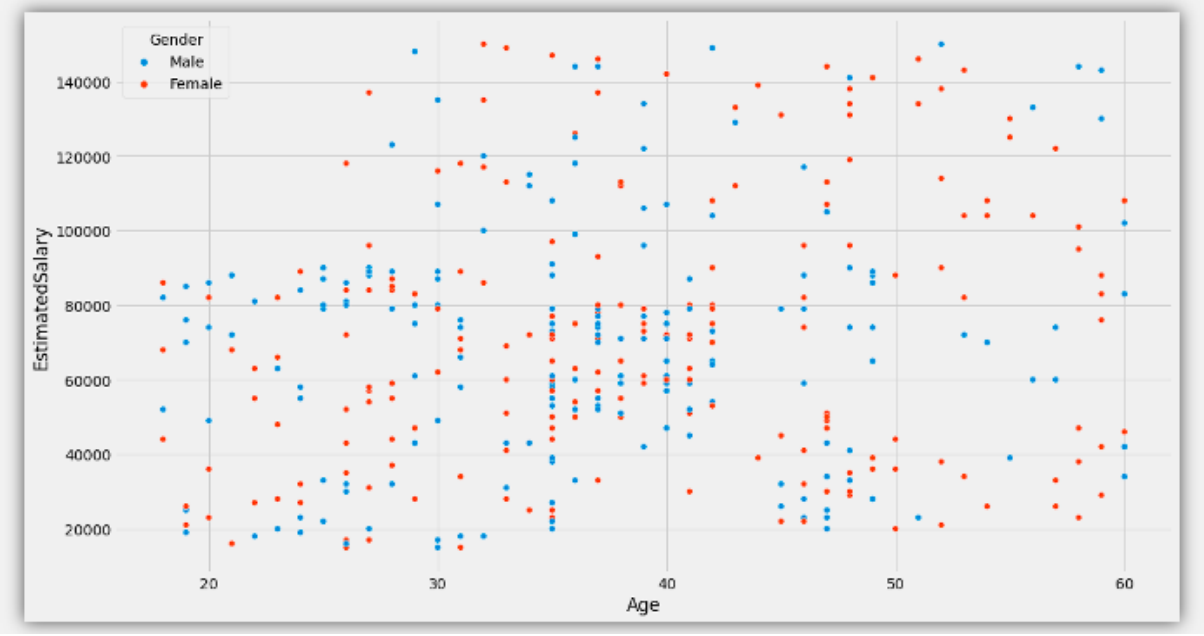

Figure 3: Relationship between Age and Salary

4.4 Age Distributed Per Gender

The histogram in figure 6 provides an insightful perspective into the distribution of ages within each gender category. By dividing the data into 'Male' and 'Female' groups, we gain a clearer understanding of how age is distributed among individuals of different genders. Notably, the histogram demonstrates a relatively balanced representation of analyzed people per gender across different age groups. However, a slight variation emerges, revealing a slightly higher number of men around the age of 35.

This observation suggests that, while there is a fairly even distribution of analyzed individuals within each gender across various age ranges, a relatively larger number of men around 35 years of age can be identified. The histogram effectively visualizes these trends, enabling us to identify potential age-related patterns that might contribute to gender-specific behaviors or preferences within the context of the dataset.

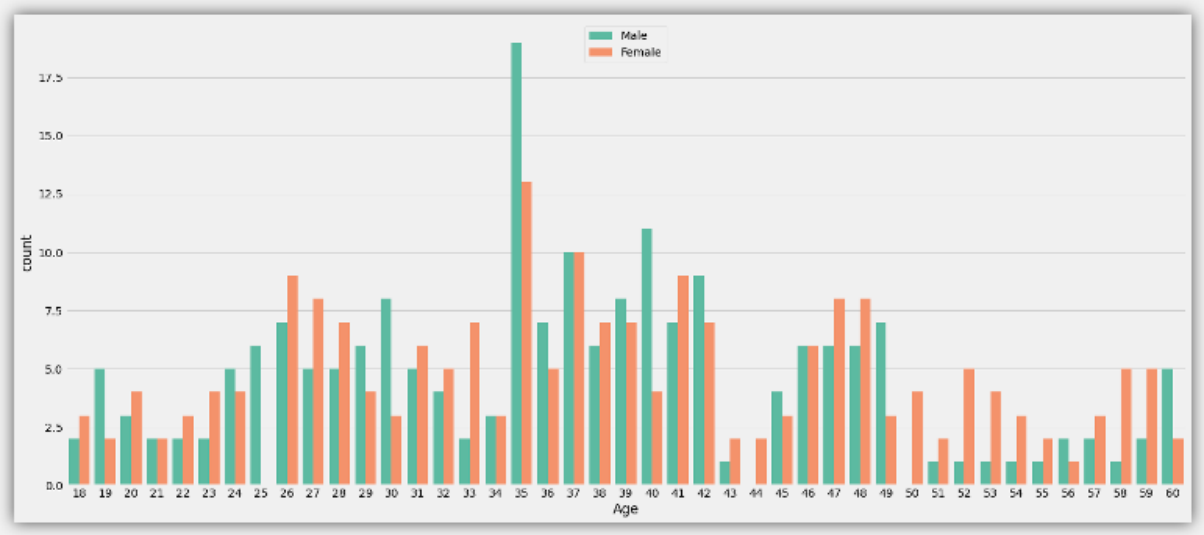

Figure 4: Age Distributed Per Gender

4.5 Purchased Based on Age

The histogram depicted in figure 5 offers a compelling depiction of how purchase behavior aligns with participants' ages. By segmenting the data into 'Purchased' and 'Not Purchased' categories, we gain valuable insights into the age dynamics associated with these behaviors.

Upon analyzing the histogram, a distinctive trend becomes apparent. The majority of individuals who made purchases fall within the age range of 50 to 60 years. This concentration implies that the product being considered seems to resonate most strongly with individuals in this age bracket. Additionally, the histogram reveals that the product's appeal diminishes for individuals outside this age range.

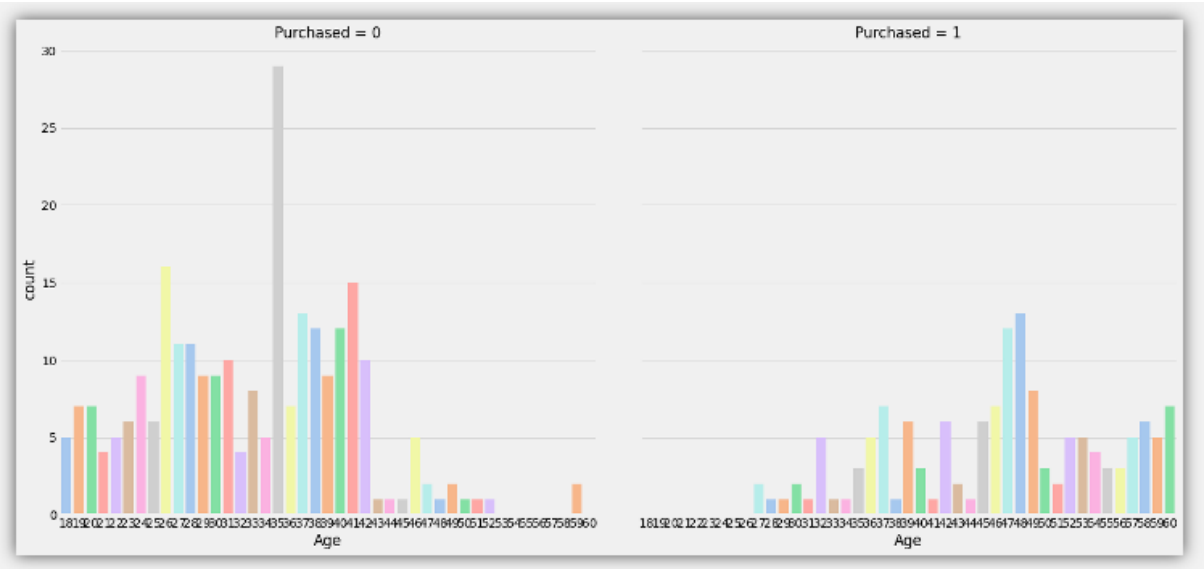

Figure 5: Purchased based on Age

5 Data Preprocessing

5.1 Handling Missing Values

At the outset of my analysis, I identified missing values within the dataset, specifically in the 'Age' and 'Gender' columns. To address this data gap, I employed a dual imputation strategy. For the 'Age' column, I utilized the mean imputation method, which involved replacing the missing age values with the calculated mean age of the available dataset. This approach not only maintained the overall age distribution but also ensured that the dataset remained complete. Additionally, for the 'Gender' column, I applied the mode imputation technique, whereby I filled the missing gender entries with the mode—the most frequently occurring gender category in the dataset. By implementing these imputation strategies, I ensured that the dataset's integrity was preserved, allowing subsequent analyses to incorporate both age and gender variables for a comprehensive exploration.

5.2 Label Encoding

I proceeded to perform label encoding on the 'Gender' column in the dataset. Label encoding is a technique used to transform categorical data, such as 'Male' and 'Female' in this case, into numerical values. This enables the data to be processed by machine learning algorithms that require numerical inputs. In this process, 'Male' might be encoded as 0, and 'Female' as 1, for instance. This transformation allows for seamless incorporation of gender as a feature in subsequent analyses, contributing to a more comprehensive understanding of how it influences various outcomes within the dataset.

5.3 Feature Scaling

Following the label encoding step, I proceeded to implement feature scaling on the 'Age' and 'Salary' columns using the StandardScaler method. Feature scaling is essential to ensure that the data is on a comparable scale, enhancing the effectiveness of various machine learning algorithms. With the StandardScaler, I standardized these two columns by subtracting the mean and dividing by the standard deviation. This process effectively centered the data around zero mean and unit variance, allowing for fair comparison and preventing certain features from dominating others due to their scale differences.

The StandardScaler transformation represented mathematically as follows:

$$x_{scaled} = \frac{x - \mu}{\sigma}$$

5.4 Split the Data

Following the preprocessing steps, I partitioned the Social Networks Ads data into training and testing sets. Utilizing a common practice, I allocated 80 percent of the data

for training purposes and reserved the remaining 20 percent for testing. This division ensures that the machine learning model is trained on a substantial portion of the dataset while maintaining an independent subset for evaluating its performance. This separation aids in assessing the model's ability to generalize beyond the training data and provides valuable insights into its predictive capabilities on unseen data points.

6 Methodologies

6.1 Naive Bayes

Naive Bayes is a probabilistic classification technique that leverages the Bayes theorem, assuming independence among the features. By utilizing scikit-learn's implementation, I created a Naive Bayes classifier that learned from the training data's patterns to predict whether an individual would make a purchase based on age, gender, and salary. The trained model was then evaluated using the reserved testing data to gauge its accuracy and predictive performance on new, unseen instances.

6.2 Decision Tree

A Decision Tree is a versatile and intuitive classification method that partitions the data into subsets based on features, effectively creating a tree-like structure of decisions. By analyzing the training data, the Decision Tree algorithm 'learns' to make decisions that optimally separate the data points, aiming to achieve the highest accuracy in predictions. This approach not only provides insights into feature importance but also aids in comprehending decision paths leading to specific outcomes.

Using scikit-learn's implementation, I crafted a Decision Tree classifier tailored to the dataset. This classifier learned from the training data's patterns, capturing relationships between age, gender, salary, and purchase behavior. The trained Decision Tree was then put to the test using the testing data, evaluating its accuracy and predictive prowess on new instances. By opting for a Decision Tree,

6.3 k-Nearest Neighbors

kNN is an instance-based classification technique that operates on the principle of proximity. When predicting an outcome for a new data point, kNN considers the 'k' closest data points in the training set and determines the majority class among them. This approach takes advantage of local patterns in the data and can be especially effective in capturing non-linear relationships.

Implementing kNN using scikit-learn, I trained a kNN classifier on the training data, with 'k' chosen as an appropriate value. The trained classifier leveraged the proximity of data points based on age, gender, and salary to predict purchase behavior. Subsequently,

I evaluated the kNN model's performance using the testing data, assessing its accuracy and predictive capability on unseen instances.

6.4 Support Vector Machine

SVM is a powerful technique that seeks to find a hyperplane that best separates data points belonging to different classes. It accomplishes this by maximizing the margin between the classes, effectively identifying the optimal decision boundary.

By implementing SVM using scikit-learn, I developed a classifier tailored to the training data. This classifier learned from the training instances' distribution in the feature space, aiming to find the most suitable hyperplane to distinguish purchase behaviors based on age, gender, and salary.

6.5 Multilayer Perceptrons

MLPs are a type of artificial neural network characterized by multiple layers of interconnected nodes, or neurons. These layers include an input layer, one or more hidden layers, and an output layer. By leveraging non-linear activation functions, MLPs can capture intricate relationships between input features and target outputs.

By utilizing scikit-learn's MLPClassifier, I constructed a neural network customized for the dataset. The network's architecture was defined by specifying the number of hidden layers, the number of neurons in each layer, and the activation functions.

7 Results Evaluation

When assessing how well a classification model predicts outcomes, we use specific metrics to measure its accuracy and the types of errors it makes.

7.1 Accuracy

Measures overall correct predictions, suitable when classes are balanced.

The Accuracy is calculated as:

$$Accuracy = \frac{NumberofCorrectPredictions}{TotalNumberofPredictions}$$

7.2 Precision

Measures correctly predicted positive cases, minimizes false positives.

The Precision is calculated as:

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives}$$

7.3 Recall

Measures actual positive cases predicted correctly, minimizes false negatives.

The Recall is calculated as:

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives}$$

7.4 F1-Score

Balances precision and recall, useful for imbalanced classes.

The F1-Score is calculated as:

$$F1{-}Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

By understanding and utilizing these metrics, we can effectively evaluate our classification model and make informed decisions.

7.5 Discussion of Results

From the evaluation, we can observe from Table 2 that the Naive Bayes algorithm achieved the highest accuracy 92 Percent, showcasing its effectiveness in predicting purchase behavior based on age, gender, and salary. It also achieved a high precision of 93 percent, indicating that when it predicts a positive outcome, it's quite accurate. The Decision Tree exhibited balanced performance with an accuracy of 86 percent, while k-Nearest Neighbors showed reasonable accuracy but lower recall. SVM struggled to capture recall, highlighting that it may not be the best fit for this specific dataset. MLPs, while exhibiting the lowest accuracy, presented a competitive f1-Score, reflecting its capability to capture more complex relationships.

Classifier	Accuracy	Precision	Recall	f1-Score
NB	0.92	0.93	0.83	0.88
DT	0.86	0.78	0.78	0.80
kNN	0.81	0.80	0.64	0.71
SVM	0.75	0.83	0.40	0.54
MLPs	0.63	0.61	0.64	0.65

Table 2: Classification Performance Metrics

These results provide valuable insights into the strengths and weaknesses of each algorithm in predicting purchase behavior within the context of social network ads. Depending on the specific objectives and trade-offs, we can make informed decisions on selecting the most suitable algorithm for the task at hand.

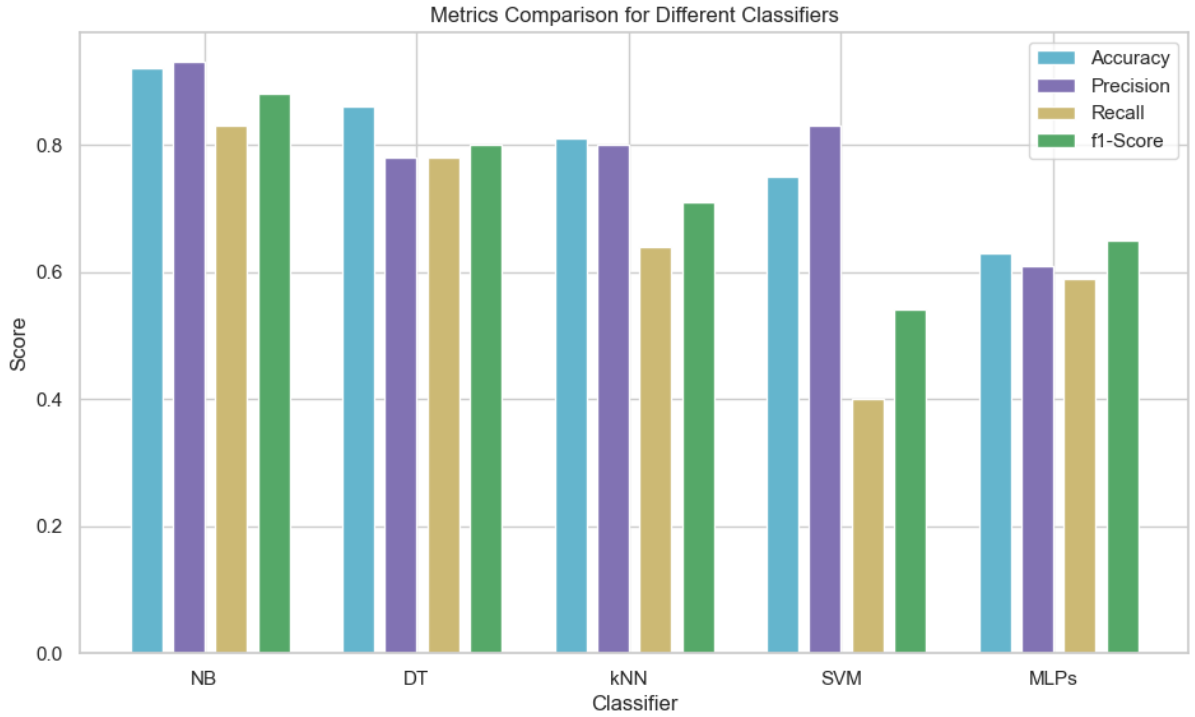

Figure 6: Result Evaluation

8 Conclusion

In summary, this analysis delved into predicting purchase behavior based on age, gender, and salary in the context of social network ads. Through a balanced exploration of classification algorithms including Naive Bayes, Decision Trees, k-Nearest Neighbors, Support Vector Machines, and Multilayer Perceptrons, a comprehensive understanding of their strengths and weaknesses was gained. The results highlight the predictive prowess of Naive Bayes, the interpretability of Decision Trees, and the potential complexities captured by k-Nearest Neighbors. The findings offer actionable insights for refining marketing strategies to target specific age groups and emphasize the significance of algorithm selection in optimizing social network ad campaigns for enhanced engagement and effectiveness.

References

[1] T. Balaji, C. S. R. Annavarapu, and A. Bablani, "Machine learning algorithms for social media analysis: A survey," *Computer Science Review*, vol. 40, p. 100395, 2021.

[2] I. J. Dickey and W. F. Lewis, "An overview of digital media and advertising," *E-marketing: Concepts, methodologies, tools, and applications*, pp. 31–61, 2012.

[3] F. Ferreira and B. Barbosa, "Consumers' attitude toward facebook advertising," *International Journal of Electronic Marketing and Retailing*, vol. 8, no. 1, pp. 45–57, 2017.

[4] Kaggle, "Social networks ads," *https://www.kaggle.com/datasets/rakeshrau/social-network-ads*, 2018.