Overview

Online social networks generate vast amounts of user data that can be leveraged for targeted advertising. Predicting which users are likely to purchase a product after seeing a social network advertisement is valuable for optimizing marketing strategies. An accurate prediction model enables marketers to allocate advertising budgets more efficiently and maximize return on investment, The objective is to evaluate and compare several machine learning algorithms in terms of their ability to predict whether a user will purchase the advertised product, using features such as the user's gender, age, and estimated salary. The dataset (Kaggle)

Dataset

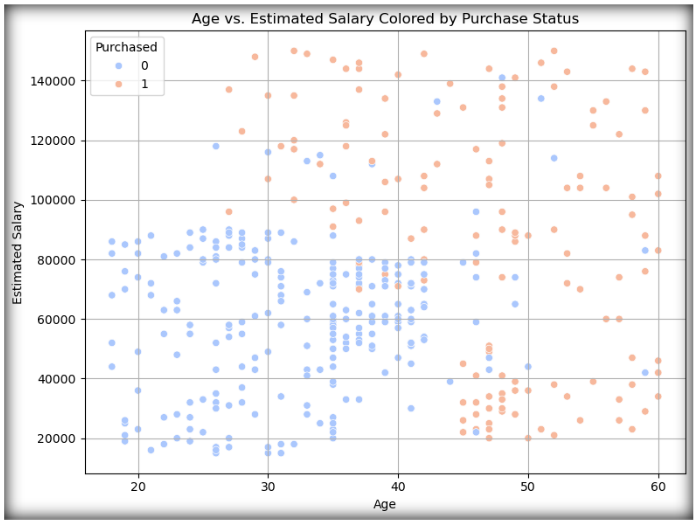
The Social Network Ads dataset consists of 400 observations of users from a social networking platform, each with several attributes and a binary outcome, The data columns include User ID, Gender, Age, EstimatedSalary, and Purchased. The User ID is a unique identifier for each user (an integer code) and is not a predictive feature, so it is not going to be used during the modeling. The Gender feature is a categorical variable with values "Male" or "Female," representing the user's sex. In this dataset, the gender distribution is roughly balanced (about 49% male and 51% female). The Age of users is given in years (ranging approximately from 18 to 60 years old), and the EstimatedSalary is an annual salary, The target variable **Purchased** is a binary indicator (Boolean) of whether the user purchased the product after seeing the advertisement (1 = *Yes*, 0 = *No*). In the dataset, about 35.8% of the users (143 out of 400) made a purchase, whereas 64.2% did not, indicating a moderately imbalanced but not severe class distribution. There are no missing values in this dataset.

Data preprocessing

we performed basic preprocessing to prepare the data. The categorical **Gender** feature was label-encoded into numeric form (e.g., Female = 0, Male = 1) to be usable by machine learning algorithms. Next, we applied feature scaling to the input variables. **Age** and **EstimatedSalary** have different value ranges and units, so to ensure comparability and to help certain algorithms converge, we standardized these features using a **StandardScaler** (z-score normalization). All feature values (including the binary gender feature) were transformed to have mean 0 and unit variance. Standardization is an important step for models sensitive to feature magnitude, such as K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Artificial Neural Networks (ANN), After encoding and scaling, the dataset is ready for the training and evaluation of various classification models.

Data visualization

Scatter Plot: we used a scatter plot to visualize how the two features age and estimated salary relate to the Purchase decision. The plot reveals a noticeable separation based on age, where most users who made a purchase (class 1) are aged 35 and above, regardless of their salary. On the other hand, Estimated Salary alone does not appear to be a strong predictor, as both classes are scattered across all salary ranges, including higher income levels. This suggests that Age has a stronger influence on purchasing behavior than salary in this dataset.



Methodology

We split the processed data into a training set and a testing set to build and evaluate our models. Using an 75/25 train-test split, 300 samples were used for training and 100 for testing (with a fixed random seed for reproducibility). The stratification was approximately preserved due to the random sampling, ensuring the training and test sets have a similar proportion of positive purchases.

Modeling

We used the following machine learning models to train and evaluate the dataset:

Random Forest Classifier: A powerful ensemble method that builds many decision trees and outputs the mode of their predictions.

Decision Tree classifier: A binary decision tree that splits the data based on feature thresholds to create a flowchart-like model for classification.

K-Nearest Neighbors (KNN): distance-based classifier that labels a test sample based on the majority class of its *k* nearest neighbors in the feature space (with *k*=5 by default)

Support Vector Machine (SVM): A classifier that finds the optimal hyperplane to separate the classes in a high-dimensional space.

Naive Bayes: A probabilistic classifier based on Bayes theorem with the assumption that features are independent distributed.

Logistic Regression: A linear classifier that models the probability of purchase as a logistic function of the features.

Artificial Neural Network (ANN): A feed-forward multilayer perceptron implemented via MLPClassifier.

Model Evaluation

Each model was trained on the training set (using the scaled input features) and then used to predict the outcomes on the independent test set. We evaluated the classification performance using several standard metrics: **accuracy**, **precision**, **recall**, and **F1-score**. In our binary classification context, precision, recall, and F1 were particularly examined for the positive class (Purchase = 1), since that class (actual purchasers) is of primary business interest.

Illustration (1)

Model	Accuracy	Precision (1)	Recall (1)	F1-score (1)
Decision Tree	90%	89%	84%	86%
Random Forest	93%	88%	95%	91%
K-Nearest Neighbors	92%	87%	92%	89%
Support Vector Machine	93%	88%	95%	91%
Naïve Bayes	93%	94%	86%	90%
Artificial Neural Net	92%	87%	92%	89%
Logistic Regression	88%	93%	73%	82%

Results

All models were successfully trained on the social network ads data, and their performance was evaluated on the 100-user test set. Overall, the predictive models achieved strong performance, with most classifiers attaining high accuracy above 90% on the test data. From the above results, we observe that the Random Forest, SVM, and Naïve Bayes classifiers achieved the highest overall accuracy on the test set, about 93%. These top models show an excellent balance of precision and recall. In particular, Random Forest and SVM both reached a recall of 95%, correctly identifying 95% of the actual purchasers, while still maintaining a precision of 88% (meaning a relatively low false positive rate). Naïve Bayes, on the other hand, achieved the highest precision of all models (94%), indicating that when it predicts a user will purchase, it is correct 94% of the time, though its recall (86%) was slightly lower than the others at this accuracy level. Following close behind, KNN and the ANN each obtained about 92% accuracy. Both had a recall around 92% and precision around 87%, indicating their performance on the positive class was quite good. Logistic Regression showed the lowest accuracy at 88%. The logistic model had high precision (93%) but significantly lower recall (73%), suggesting it was conservative in predicting purchases – it predicted fewer positives and thus missed quite a few actual purchasers (high false negative rate).

Conclusion

I selected the Social Network Ads dataset because it deals with realistic digital marketing behavior.
The dataset provides valuable insights into how user demographics affect online purchasing.
It simulates a scenario that marketers face daily: predicting which users are likely to buy.
This makes it highly relevant in today's data-driven advertising industry.
With only a few simple features (age, gender, salary), it’s both clean and effective.
It’s also perfectly suited for binary classification problems using classic models.
No missing values made preprocessing easier and more focused on modeling.
Real-world importance lies in its ability to improve ad targeting and budget efficiency.
For example, knowing which age groups are more likely to purchase helps tailor ads.
The model can guide who should see the ad and who should not, saving money.
I used label encoding and scaling to standardize the features for fair comparison.
After training, the best-performing models were Random Forest and SVM.
Both achieved 93% accuracy, 95% recall, and 88% precision.
This high recall is crucial—it means they detected nearly all potential buyers.
Naive Bayes had the highest precision (94%), making its positive predictions very reliable.
However, it missed more real purchasers, so it’s less useful if recall is the goal.
Logistic Regression had the weakest recall (73%), showing it was too conservative.
From the data, I observed that age plays a stronger role than salary in purchasing.
Younger users rarely made purchases regardless of their income.
Overall, this task showed how simple demographic data can drive real marketing decisions.