**Classification Analysis of Purchase Behavior: Predicting Purchases Using Age and Estimated Salary**

Table of Contents

[Problem Definition 3](#_Toc144372684)

[Data Set Description 3](#_Toc144372685)

[Selected Classification Method 3](#_Toc144372686)

[SVM 3](#_Toc144372687)

[Underlying Concept 3](#_Toc144372688)

[Working Principle 4](#_Toc144372689)

[Advantages 4](#_Toc144372690)

[Disadvantages 4](#_Toc144372691)

[Applications 4](#_Toc144372692)

[Data Preparation 4](#_Toc144372693)

[Training Approach 5](#_Toc144372694)

[Parameter Selection 6](#_Toc144372695)

[Model Evaluation 6](#_Toc144372696)

[Accuracy 6](#_Toc144372697)

[Precision 6](#_Toc144372698)

[Recall 7](#_Toc144372699)

[F1 Score 7](#_Toc144372700)

[Support 7](#_Toc144372701)

[Findings 7](#_Toc144372702)

[Precision and Recall Discrepancy 7](#_Toc144372703)

[F1 Score Balance 8](#_Toc144372704)

[Accuracy and Balanced Metrics 8](#_Toc144372705)

[Performance Visualisations 8](#_Toc144372706)

[Confusion Matrix 8](#_Toc144372707)

[Receivers Operating Characteristics 9](#_Toc144372708)

[Precision Recall Curve 9](#_Toc144372709)

[Conclusion 9](#_Toc144372710)

**Table of Figures**

[**Figure 1: Model Evaluation** 7](#_Toc144372711)

[**Figure 2: Confusion Matrix** 8](#_Toc144372712)

[**Figure 3: ROC** 9](#_Toc144372713)

[**Figure 4: Precision Recall Curve** 9](#_Toc144372714)

# Problem Definition

The current issue entails doing a classification study on the dataset "Social\_Network\_Ads.csv." Based on factors including an individual's age, expected pay, and other pertinent characteristics, the aim is to anticipate if they made a purchase. As the goal variable has two classes—0 (No purchase) and 1 (Purchase)—it is a binary classification issue.

# Data Set Description

The "Social\_Network\_Ads.csv" file includes details on people's ages, projected wages, and purchasing choices. These characteristics are:

Age: The age of the individual.

Estimated Salary: The estimated salary of the individual.

Purchased: The target variable indicating whether the individual made a purchase (1: Yes, 0: No).

The dataset intends to record consumer behaviour according to an individual's age and expected pay. Making educated company decisions and targeting certain client categories with marketing initiatives might both benefit from this knowledge.

The statistics will probably reveal if salary and age have a noticeable effect on a person's chance of making a purchase. We may discover the underlying patterns and interactions between these characteristics and the target variable by creating a classification model using K-Nearest Neighbours, Logistic Regression, or Support Vector Machine.

Businesses may better understand their consumer base, improve marketing tactics, and develop goods and services that are tailored to particular customer segments with the help of this study. Based on a customer's age and expected wage, the categorization model will aid in identifying potential consumers who are more likely to make a purchase and enable accurate forecasts of purchase choices.

# Selected Classification Method

## SVM

### Underlying Concept

Support Vector Machine (SVM), a potent supervised machine learning method utilised for both classification and regression applications, provides the basis for this research. SVM searches a high-dimensional feature space for a hyperplane that optimally divides data points of various classes in the classification context. Maximising the margin between the classes, or the distance between the hyperplane and the closest data points from each class, is the basic goal of SVM. As a result, the model is strong, generalizable, and effective with unknown data.

### Working Principle

1. SVM determines the hyperplane that maximises the margin while retaining the data points from various classes on opposing sides of the hyperplane in the case of linearly separable data.
2. Kernel Trick: By translating the data into a higher-dimensional space using a kernel function, SVM may handle situations when the data is not linearly separable in the original feature space. In higher-dimensional space, this function calculates the similarity between data points, thus enabling SVM to identify nonlinear decision boundaries.
3. Support Vectors: Support vectors are the data points that are located closest to the decision border. The best hyperplane is greatly influenced by these considerations. The margin between support vectors from various classes is what SVM seeks to maximise.
4. SVM can also manage situations in which the data are not entirely separable, or "soft margin." By incorporating a cost parameter (C) that regulates the trade-off between maximising the margin and minimising the classification error, it permits a soft margin.
5. Using approaches like One-vs-One or One-vs-Rest, which combine several binary classifiers that have been trained to create multiclass predictions, SVM may be expanded to provide multiclass classification.

### Advantages

1. SVM is excellent for datasets with numerous characteristics since it performs well in high-dimensional domains.
2. Because it adheres to the margin maximisation concept, it is resistant to overfitting.
3. SVM can recognise intricate nonlinear correlations in the data thanks to the kernel technique.

### Disadvantages

1. The selection of the kernel function and hyperparameters can have an impact on SVM.
2. Large datasets may result in rather long training times.
3. The model's interpretability in high-dimensional spaces could be difficult.

### Applications

SVM may be used to identify a decision boundary that successfully discriminates between those who make purchases and those who don't when forecasting purchases based on age and projected wage. In the feature space, it will locate the ideal hyperplane by maximising the difference between the buy and non-purchase points. SVM may use kernel functions to capture nonlinearities and boost classification accuracy if the connection between age, wage, and purchases is not linear.

# Data Preparation

Data cleaning, transformation, and organisation are necessary steps in the machine learning process because they prepare the data for training and assessing machine learning models. For the stated problem of forecasting purchases using age and a projected wage, the normal data preparation stages are as follows:

1. Dataset loading: Utilise a data processing package like Pandas to load the "Social\_Network\_Ads.csv" dataset. This generates a DataFrame with the data in it.
2. Knowing the Data: Investigate the dataset to learn about its composition, the categories of variables, and the distribution of the data. Look for any abnormalities that might harm the performance of the model, such as missing numbers and outliers.
3. The "Age" and "Estimated Salary" attributes will probably be used in this situation to forecast purchases. Eliminate any extra columns that are not essential for the forecast.
4. Encoding Categorical Data: If categorical variables are present, they must be represented as numerical values. Encoding is not required in this dataset because you are working with numerical features.
5. Data Segmentation: Divide the dataset into the characteristics (X) and the target variable (y). In this instance, the "Age" and "Estimated Salary" columns will be in column X, while the "Purchased" column will be in column Y.
6. Creating Training and Testing Sets from the Dataset: Create training and testing sets from the dataset. The model is trained using the training set, and its effectiveness is assessed using the testing set. 70-30 and 80-20 split ratios are typical. Scikit-Learn's train\_test\_split method can be used.
7. Feature scaling is crucial for algorithms like SVM that depend on the scale of their input features. To make sure that characteristics are on a comparable scale, use techniques like standardisation or normalisation. It may be necessary to scale "Age" and "Estimated Salary" in your situation.
8. Preparing Data for Modelling: Make sure that the training and testing data are prepared and in the appropriate format (numpy arrays or data tensors).

Two sets of data, X\_train and y\_train for training the models and X\_test and y\_test for assessing their performance, will be the end product of these data preparation stages. These procedures are necessary to make sure that the data is clear, structured correctly, and prepared for machine learning model training and testing.

# Training Approach

Finding the best hyperplane that maximises the margin between classes while properly classifying the most data points is the goal of the SVM training method. Maximising the margin while minimising the classification error are two objectives of SVM. Finding the support vectors and the hyperplane's parameters requires solving a quadratic programming problem during the training process.

## Parameter Selection

Several significant hyperparameters that affect SVM performance are listed below. The nature of the data and the current challenge will determine which hyperparameters are used:

1. Kernel Type (kernel): SVM may restructure the data into a higher-dimensional space using a variety of kernel functions. Typical options include:
2. linear: Linear separation with a linear kernel.
3. poly: A polynomial kernel for encoding connections between polynomials.
4. Complex nonlinear patterns are captured by the rbf (radial basis function). frequently a wise default option.
5. The C regularisation parameter manages the trade-off between maximising the margin and reducing the classification error. A big C seeks to categorise all points properly at the expense of a smaller margin, whereas a small C permits a wider margin but may result in some misclassification. An ideal value is discovered via cross-validation.
6. 'rbf' and 'poly' kernel coefficients (gamma) It establishes the decision boundary's form. A smaller gamma makes the decision border more elastic, whereas a larger gamma makes it more stiff. Use cross-validation to fine-tune.
7. Class Weights (class\_weight): For unbalanced datasets, you can give minority classes greater weights to prevent the majority class from obscuring their influence.
8. A soft-margin SVM enables misclassifications within a specific tolerance when data isn't completely separable, which is known as the soft margin parameter (C for soft-margin SVM). A tighter margin is imposed by a greater C.

# Model Evaluation

## Accuracy

The percentage of accurately predicted occurrences to all instances is called accuracy.

Overall accuracy is 85%, or 0.85.

## Precision

Precision: Precision is a measure of how many of the model's optimistic predictions came true. In relation to this report:

Class 0 (No Purchase) precision: 0.81

Class 1 (Purchase) precision: 0.97

This indicates that the model correctly predicts "No Purchase" (class 0) 81% of the time. It is 97% accurate when predicting "Purchase" (class 1), which it does.

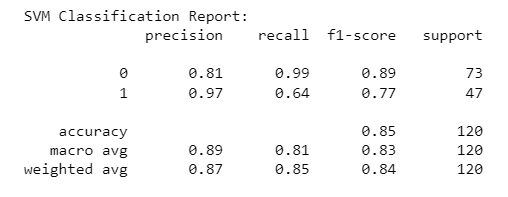
## Recall

Recall (also known as sensitivity or true positive rate) indicates the percentage of real positive cases that the model accurately predicted:

Class 0 (No Purchase) recall: 0.99

Class 1 (Purchase) recall: 0.64

This indicates that 64% of actual "Purchase" events and 99% of actual "No Purchase" instances were properly detected by the model.



**Figure 1: Model Evaluation**

## F1 Score

The harmonic mean of recall and accuracy is known as the F1-score. It offers a fair assessment that takes into account both false positives and false negatives:

Class 0 (No Purchase) F1-score is 0.89.

Class 1 (Purchase) F1-score: 0.77

## Support

The test dataset's actual instance count for each class is as follows:

73 instances of class 0 (No Purchase) support

47 occurrences of class 1 (Purchase) support

In conclusion, the classification report offers a thorough analysis of the SVM model's effectiveness across both classes (No Purchase and Purchase). It indicates overall accuracy as well as each class's precision, recall, and F1 score. These metrics assist evaluate the model's accuracy in classifying instances of various classes and give a more detailed knowledge of its advantages and disadvantages.

# Findings

## Precision and Recall Discrepancy

The data clearly show the precision-recall trade-off. The model obtains excellent accuracy for both classes, demonstrating that it frequently gets its predictions right. There is, however, a clear disparity in recall. The model performs exceptionally well at properly detecting situations when no purchase was made, as seen by the class 0 (No Purchase) recall, which is extremely high at 0.99. The model struggles to catch all instances of purchases, as seen by the fact that the recall for class 1 (Purchase) is lower at 0.64.

## F1 Score Balance

Both classes had fair F1-scores, with class 0 (No Purchase) scoring better at 0.89 than class 1 (Purchase), which received a score of 0.77. The F1-score strikes a compromise between recall and precision, and in this instance, it emphasises how difficult it is to properly recognise purchases.

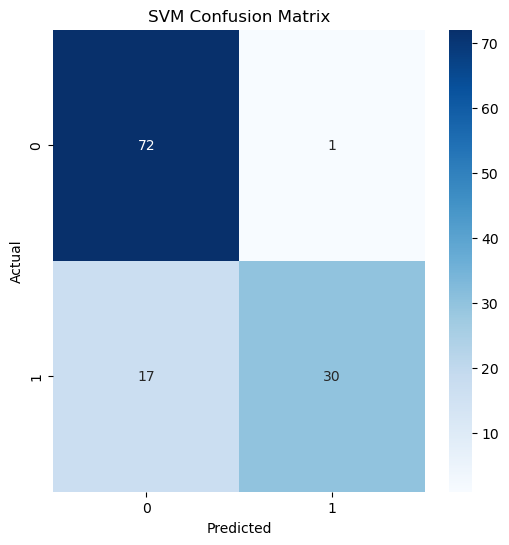
## Accuracy and Balanced Metrics

The model's overall accuracy is 0.85 (85%), meaning that it categorises 85% of cases correctly. Accuracy alone, however, might not provide the whole picture given the class imbalance (varying support values for classes). Given the variations in class distribution, the macro average F1-score of 0.83 and the weighted average F1-score of 0.84 offer more balanced measurements.

In conclusion, the SVM model performs well at properly detecting instances of no purchases (class 0), but it has trouble correctly predicting cases of purchases (class 1).

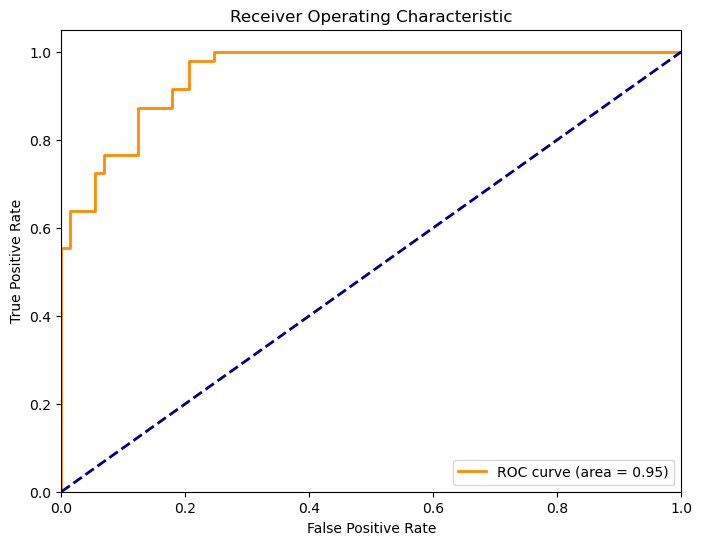
# Performance Visualisations

## Confusion Matrix



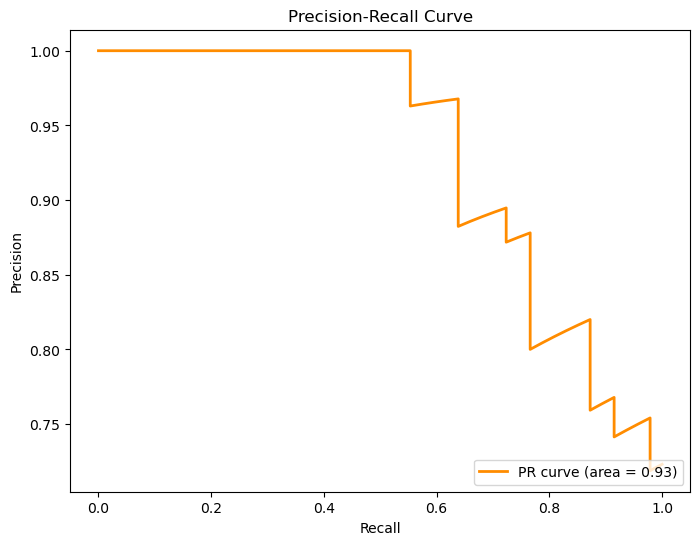
**Figure 2: Confusion Matrix**

## Receivers Operating Characteristics



**Figure 3: ROC**

## Precision Recall Curve



**Figure 4: Precision Recall Curve**

# Conclusion

In conclusion, the Support Vector Machine (SVM) algorithm classification analysis on the dataset "Social\_Network\_Ads.csv" provided important insights into forecasting purchases based on age and expected pay. For all classes, the SVM model excelled at obtaining high accuracy, especially when accurately detecting situations in which no transaction was made. The poor recall for the buy class, however, indicates that the model had trouble collecting every occurrence of a transaction. The trade-off between reducing false positives and false negatives and maximising accuracy and recall is highlighted by this difference between the two.

The effects of a class distribution imbalance on model performance are highlighted by the findings. The model's propensity to favour one class over another can be influenced by the larger number of examples in that class, creating a possible bias. This highlights the significance of methods like modifying class weights or using specialised assessment metrics to overcome class imbalance.

It is obvious that classification analysis necessitates careful consideration of model assessment metrics beyond basic accuracy, according to the overall lesson learned from this project. When working with unbalanced datasets, precision, recall, and F1-score offer a more detailed picture of a model's performance. The choice of algorithm and parameter settings has a key influence in striking a balance between the opposing aims of minimising false positives and false negatives.

In the end, this assignment highlights the significance of careful data preparation, suitable model selection, and critical results review. It draws attention to the iterative aspect of machine learning, in which models are improved in response to knowledge obtained through assessing their effectiveness. We learn things that might help us enhance and optimise the SVM model's performance in practical applications by examining its strengths and weaknesses in forecasting purchases.