ADS CLASSIFICATION USING SOCIAL NETWORK

A PROJECT REPORT

for

DATA MINING TECHNIQUES (SWE2009)

in

M. Tech (Integrated) Software Engineering

by

NITHISH KUMAR S (20MIS0024)

JEYADARSHAN A (20MIS0044)

SHRUTHI R (20MIS0438)

Under the Guidance of

Dr. SENTHILKUMAR N C

Associate Professor, SITE

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

School of Information Technology and Engineering

April, 2023

DECLARATION BY THE CANDIDATE

We hereby declare that the project report entitled **“ADS CLASSIFICATION USING SOCIAL NETWORK”** submitted by us to Vellore Institute of Technology; Vellore in partial fulfillment of the requirement for the award of the course **Data Mining Techniques (SWE2009)** is a record of bonafide project work carried out by us under the guidance of **Dr. Senthilkumar N C.** We further declare that the work reported in this project has not been submitted and will not be submitted, either in part or in full, for the award of any other course.

Place: Vellore

Signature

Date: 04th April, 2023.

Shruthi R

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

School of Information Technology & Engineering [SITE]

CERTIFICATE

This is to certify that the project report entitled **“ADS CLASSIFICATIN USING SOCIAL NETWORK”** submitted by **Nithish Kumar S (20MIS0024)** to Vellore Institute of Technology, Vellore in partial fulfillment of the requirement for the award of the course **Data Mining Techniques (SWE2009)** is a record of bonafide work carried out by them under my guidance.

Dr. Senthilkumar N C

Associate Professor, SITE

Ads Classification using Social Network

Abstract

Given the constant growth of investment in online advertising, it is essential to have tools that provide complete, accurate, verified, and real-time monitoring of the ads companies place on different websites and social networks to sell their products. For this reason, many advertisement companies provide a digital platform that allows easy and secure access to strategic information for advertising agencies, media, media centers, and advertisers so that they can verify that their advertising campaigns are being carried out successfully. Optimized advertising is then essential to maintain user engagement in the platform. However, a challenge that many companies face today is the lack of semi-automatic mechanisms to identify advertising from their clients and even detect competing advertisers who invest their money in other media. Currently, this identification is accomplished by many companies manually, representing a high cost in time and effort. Hence, this paper is a comparative study of machine learning algorithms to classify if a particular user will visit the ad or not.

Keywords – advertisement classification, machine learning, classification, effective advertisement.

I. INTRODUCTION

Effective advertising is key to successful businesses. Relevant and engaging advertisements are essential to reaching potential customers. Personalized and customized advertising approaches are needed to address different customers' needs. Although smart displays for ads are present, the market demands enhanced targeted advertisement approaches that can be displayed on smart screens in public places and upcoming smart cities. Artificial intelligence (AI) and deep machine learning approaches are growing significantly with an expected compound growth rate of 33.1% in the upcoming 15 years. Various industries such as healthcare, education, and marketing deploy AI techniques to enhance and optimize their applications. One of the common use cases for AI in marketing is in advertisement. For example, Facebook collects and analyzes users' data to model their interests and adapt advertisements' content accordingly. Similar approaches match user profiles on social media platforms to recommend the most suitable users for advertisement campaigns. A study conducted in applies AI on Instagram profile pictures to determine the gender of the user for targeted advertising. Deep Neural Network (DNN) algorithms are also used for effective advertisement recommendations. The use of machine learning and deep learning techniques improves the quality of advertising.

II. BACKGROUND

Social media has become a primary source of information for decision-makers, organizations, and scientists in today's fast-paced world. Indeed, because of the large volume of user-generated data available on social media, these online platforms are viewed as computable data sources that potentially mirror reality. It could be an authentic environment for the task of target customer's identification for marketing. The problem is identifying a group of social media users who are likely to be the potential target of those images. A target market is a group of potential customers to whom a business would like to sell its goods and services. Identifying the target market is one of the crucial stages in developing a marketing strategy for any business. This study aims to propose a system that uses social media analysis to discover potential customers.

III. LITERATURE SURVEY

A numerous machine learning algorithms have been discussed in the following research papers:

[1]. This paper uses data from Facebook to study the association of social media marketing content with user engagement. We find that inclusion of content related to brand personality is associated with higher levels of engagement (Likes, comments, shares) with a message. Intriguing content, such as deals and promotions, drive consumers' path to conversion (click-through). These results persist after incorporating corrections for the non – random targeting of Facebook's EdgeRank algorithm. The methodology we apply to content-code text is useful for future studies utilizing unstructured data such as advertising content or product reviews.

[2]. The massive amount of data generated by users using social media platforms is the result of the integration of their background details and daily activities. This enormous volume of generated data known as "big data" has been intensively researched recently. A review of the recent works is presented to obtain a broad perspective of the social media big data analytics research topic. We classify the literature based on important aspects. This study also compares possible big data analytics techniques and their quality attributes. Open research challenges in big data analytics are described as well.

[3]. This study was conducted to conceptualize advertising value and consumer attitudes towards advertisements. The research was developed to reveal the effect of the source of advertisements on credibility perception through the theoretical framework of Ducoffe's (1995) advertising value model. The research objective is to identify source derogation in terms of credibility to create advertising value and a positive attitude towards advertisements launched through the Facebook social network.

[4]. Social networks are consequence of the lack of understanding of secrecy and protection on OSN (online social networks) and media. According to studies, OSN users expose their private information such as email address, phone number etc. This paper has presented a high-level classification of recent OSN attacks for recognising the problem and analysing the blow of such attacks on World Wide Web.

[5]. The amount and variety of data generated through social media sites has increased along with the widespread use of these sites. The inclusion of personal information within these data makes it important to process the data. This process can be called intelligence and this meaningful information may be for commercial, academic, or security purposes.

[6]. The proposed algorithm in this study based on deep learning is designed to handle promotional image and message. Its performance is evaluated on the promotional advertising data provided by Wongnai. The accuracy of the proposed method can achieve satisfactory performance of 82.95% in testing data.

[7]. Social media (SM) represent beneficial channels for marketers, business promoters and consumers. To acquire continuous revenues and more active customers, key business players should understand the behavior and purchase preferences of buyers. To predict the buying decisions of purchasers, data about purchase intentions and desires have to be extracted with the help of data mining techniques. The purpose of this paper is to examine social media data analytics using machine learning tools; this new approach for developing a social media marketing strategy employs the Waikato

Environment for Knowledge Analysis (WEKA). WEKA is compared with other algorithms of interest and found to outperform its peers, especially with regard to parameters such as precision, recall, and F-measure, indicating that WEKA performs better than other approaches.

[8]. The machine learning classifier has practical implications since it allows an extensive competitor analysis to be conducted and is also able to influence social media campaigns. The results show that the machine learning classifier obtained has an excellent potential to measure effectiveness and analyses a significant amount of content with more efficiency.

[9]. This paper proposes the factors and characteristics of online classified ads including structured data and free text that may affect the effectiveness of the advertising in terms of the ability to close the sales are investigated. Classified advertising is one of the most popular online advertising especially for second-hand goods. One of the main functions of online classifieds is to help sellers sell their items quickly and easily. They have used k – means clustering integrated with decision tree classification in modelling to classify the ability to close the sales. The predictive efficiency of the models will be measured and compared using accuracy, precision, recall, and F-scores by using the 3 – fold Cross Validation method.

[10]. Major user response prediction models have to either limit themselves to linear models or require manually building up high-order combination features. The former loses the ability of exploring feature interactions, while the latter results in a heavy computation in the large feature space. To tackle the issue, we propose two novel models using deep neural networks (DNNs) to automatically learn effective patterns from categorical feature interactions and make predictions of users' ad clicks. To get our DNNs efficiently work, we propose to leverage three feature transformation methods, i.e., factorization machines (FMs), restricted Boltzmann machines (RBMs) and denoising auto-encoders (DAEs).

[11]. The goal of this paper is to develop a profile of Industry 4.0 job advertisements, using text mining on publicly available job advertisements, which are often used as a channel for collecting relevant information about the required knowledge and skills in rapid-changing industries. We searched website, which publishes job advertisements, related to Industry 4.0, and performed text mining analysis on the data collected from those job advertisements.

[12]. This study constructed a conceptual model based on cue utilization theory focusing on the effects of consumer perceptions to the personalized online ads on click-through intention. Empirical results based on data from a survey of 446 WeChat moments users in China showed that: (1) consumer's ad click-through intention increased as a result of employing a higher extent of product involvement, brand familiarity, visual attractiveness and information quality to consumer; (2) trust played a role of mediation in the processes of visual attractiveness and information quality affecting click-through intention; (3) the higher product involvement also stimulated the consumer's privacy concerns, which played negative moderating effects on the positive impacts of product involvement, brand familiarity and trust on click-through intention.

[13]. To date, existing methods for Ad click prediction, or click-through rate prediction, mainly consider representing users as a static feature set and train machine learning classifiers to predict clicks. Such approaches do not consider temporal variance and changes in user behaviors, and solely rely on given features for learning. In this paper, we propose two deep learning-based frameworks, LSTMcp and LSTMip, for user click prediction and user interest modelling. Our goal is to accurately predict the probability of a user clicking on an Ad and the probability of a user clicking a specific type of Ad campaign.

[14]. Social Media is becoming the next-level journalism as it reflects all the viral contents, social affairs, current conditions of our surroundings. This study proposes a system which will analyze Facebook data and will classify and detect the problems people are facing most frequently. The problems are categorized into 12 major classes addressing the socioeconomic aspects of Bangladesh

[15]. The classification and recommendation system for identifying social networking site users' interests plays a critical role in various industries, particularly advertising. Our proposed system provides insights into personalized SNS advertising research and informs marketers on making interest – based recommendations, ranked-order recommendations, and real – time recommendations.

[16]. This paper proposes systematic and purposeful endeavors utilized for impacting individuals for the political and religious gains. Machine Learning classifiers are used for classifying the text into binary classes (Propaganda and Non Propaganda). Support Vector Machine showed better results among all other traditional machine learning algorithms. They chose Logistic Regression, Multinomial Naive Bayes, Support Vector Machine and Decision Tree Algorithms for performing classification.

[17]. This paper proposes the screening of advertisements and user feature vectors, combining factor decomposition machine and neural network, a multi-label model is built, and through training data, a static "advertisement-user" classification matching push model is obtained. On the basis of the static model, three sub-models are respectively established for comprehensive evaluation to realize the "user-advertisement" household matching push. Combined with data set verification, it is concluded that the model is true and effective and can realize personalized recommendation.

[18]. This paper proposes a greedy algorithm and a genetic algorithm to find near-optimal combinations of conceptual nodes in polynomial time, with the genetic algorithm nearly matching the optimal solution. The study of click-through rate (CTR) prediction of advertisements, in environments like online social media, is of much interest. Prior works build machine learning (ML) using user-specific data to classify whether a user will click on an advertisement or not. ML models are trained using the advertisement data to perform CTR prediction with conceptual node combinations. They observe that simple ML models can exhibit the high Pearson correlation coefficients w.r.t. click predictions and real click values

[19]. This paper proposes a methodology for identification of a segment of customers to whom advertisement of a particular product can be shown. One of the important aspects of this work is data. To create a database of the customer, datasets from different sources/websites are collected, cleaned and pre – processed to prepare the customer database. DBSCAN (Density-based spatial clustering of applications with noise) clustering algorithm is then used to form customer segments. KNN (K – Nearest Neighbor) algorithm and performance metrics like Silhouette Coefficient are used for the smooth conduct and working of the clustering algorithm. The customer database and the labels generated after clustering are used to build and train a Logistic Regression Classification Model. The classification model is then used to predict the classes/labels for the product. The labels predicted by the Classification Model are used to extract a target customer segment.

[20]. The authors propose a creative advertising system (CAS) for the generation and testing of advertising creative ideas, founded on artificial intelligence (AI) principles. The proposed system emerges from a conceptual framework where advertising creativity is more broadly defined as a search process, the outcomes of which should be evaluated based on a set of rules. This broader definition provides a generative perspective and extends current approaches to advertising creativity that are mainly based on outcome measures such as perceived novelty and appropriateness (value).

[21]. Many enterprises choose to publish advertisements on short video platforms for higher revenue with lower cost in order to lure users into using their products, but these ads might be risky. In this study, the crawler on the mobile Internet is realized through the network packet capture technology. The short video platform advertisements are collected and classified to provide data support for the follow-up advertising monitoring work. Finally, the ads are divided into two kinds, regular and risky through discrimination, which provides data support for the follow-up supervision.

[22]. This paper proposes a need for an intelligent method to detect clickbait and fake advertisements on social networks. Several machine learning methods have been applied for this detection purpose. However, the obtained performance (accuracy) only reached 87% and still needs to be improved. The proposed approach includes three main phases: data collection, data preparation, and machine learning model training and testing phases. The ML models were evaluated, and the overall performance is reported in this paper. The experimental results show that the Support Vector Machine (SVM) with the top 10% of ANOVA F-test features (user-based features (UFs) and content-based features (CFs)) obtained the best performance and achieved 92.16% of detection accuracy.

[23]. This paper proposes a keyword-based advertising search framework to provide instant access to the relevant advertisements from online English newspapers in a category of reader's choice. First, an image extraction algorithm is proposed which can identify and extract the images from online newspapers. The proposed 'Adv_Recognizer' model separates advertisement and non-advertisement images with an accuracy of around 97.8%. and the proposed 'Adv_Classifier' model classifies the advertisements in four predefined categories exhibiting an accuracy of around 73.5%.

[24]. To predict the user engagement rate, we extract the significant attributes of posts and introduce an adaptive hybrid convolutional model based on FW-CNN-LSTM. We cluster the selected data based on the weight and significance of their attributes using the FCM and XGBoost algorithms and then apply CNN- and LSTM-based methods to select similar features. Using accuracy, recall, F-measure, and precision metrics, we compared our algorithm to standard techniques such as SVM, Logistic regression, Naïve Bayes, and CNN. According to the findings, hashtag, brand ID, movie title, and actors achieve the highest scores, and the values for actual training time in various data ratios are relatively linear, which confirms the scalability of the proposed model for large datasets.

[25]. A quantitative research approach was used to achieve study objectives and examine the hypothesized research framework by using a customized survey questionnaire in the retailing sector. A total of 255 valid responses were considered for further analysis by using SmartPLS3 software to conduct the key analyses. The results revealed the significant effect and role of all digital marketing channels on the consumers buying decisions, with the moderated role of the eWOM on the effect of digital marketing channels on consumer buying decisions.

[26]. Top social media sites are becoming effective marketing tools, perhaps taking the place of more conventional options like TV ads or brochures. The internet is a key marketing tool that may be utilized to increase brand awareness, draw in clients, and establish credibility. Social media is used by social scientists and business professionals all around the world to study how individuals interact with their environment. It all comes down to analyzing the ads/commercials to determine whether or not your target market will really purchase the goods. This is a fantastic application of data science in marketing. This article shows how to categorize your target audience by analyzing social media marketing.

[27]. This paper proposes a real-time customer detection and classification system at the supermarket. The goal of this proposed Internet of Things (IoT) system is to automatically show the suitable advertising clips to many customers at the right time. They built a classification model using deep learning with a large amount of data. The dataset is collected from reality and labelled with five different object classes. The data is trained on YOLOv4 and YOLOv4-tiny models. The models are deployed on the embedded system with the Jetson Nano device as the processor. We compare the accuracy and speed of the two models on the same embedded system, analyse the results, and choose the best model according to the specific system requirements.

[28]. This paper proposes an innovative and sustainable advertisement display system tailored to bystanders' liking based on age and gender using Convolutional Neural Networks. The bystander's face is detected using a webcam attached to the advertising screen and fed into the trained Convolutional Neural Network to identify the age and gender. The classification results are stored in a real-time database server and retrieved on a website that enables the client to customize advertisements to the targeted groups. The proposed system achieved accuracy of 91.7% and 95.5% for age and gender, respectively.

[29]. This paper proposes a novel image-based personalized advertisement recommendation system named IPARS to identify target customers in social media using image processing and machine learning techniques for an online advertisement. Assume having a set of advertising images. In IPARS, a given social network is first converted into a weighted bipartite graph where the nodes are the users and keywords. Then, another bipartite graph is formed by decomposing the advertising images into their objects, labels, concepts, and sentiments. An algorithm is proposed to search the social graph and identify and rank the best group of users.

[30]. Street-level images are highly suited to predict building functions as the building façades provide clear hints. Social media image platforms contain billions of images, including but not limited to street perspectives. This study proposes a filtering pipeline to yield high-quality, ground-level imagery from large-scale social media image datasets.

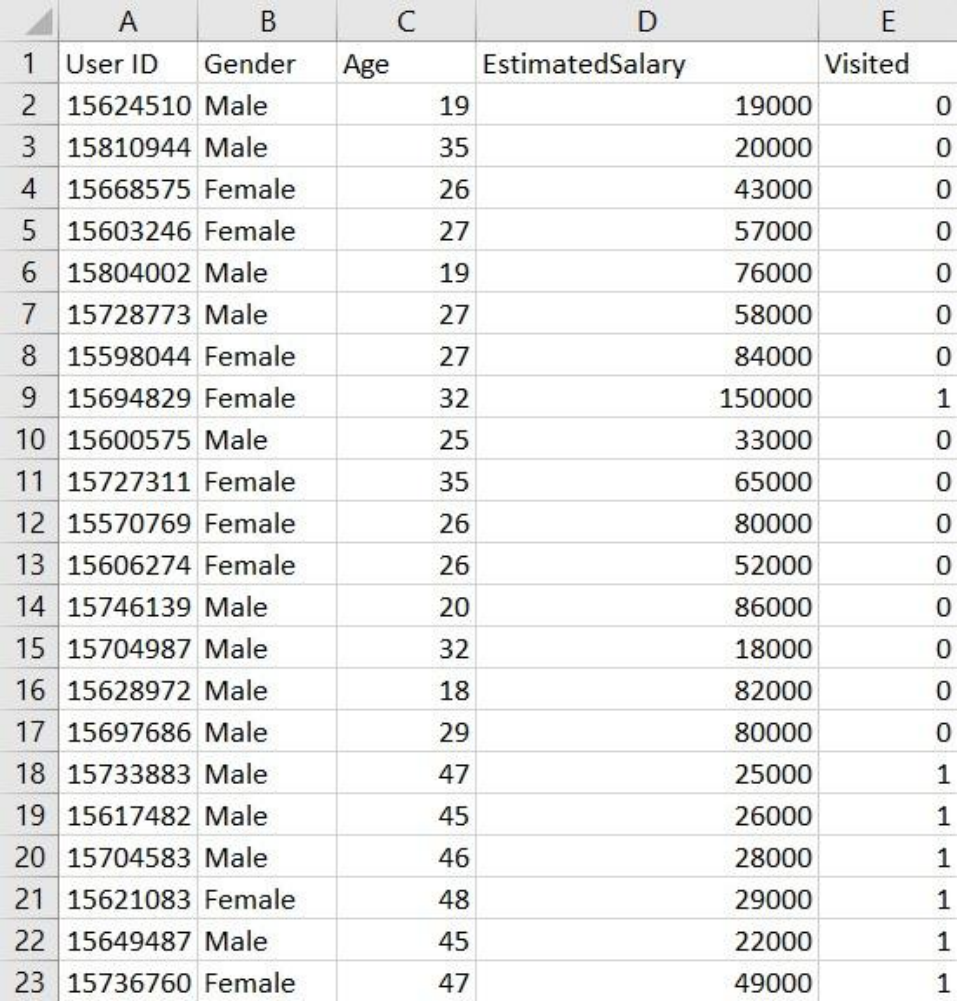
IV. DATASET DESCRIPTION & SAMPLE DATA

As per the literature survey, the datasets were collected from various sources / websites. These datasets are taken according to and in reference to the type of sector selected for the model. These datasets were semi – automatically cleaned and pre – processed to remove unwanted, correct and duplicate data. The final dataset is termed as a customer database and includes 401 records and 5 features.

The features are as follows:

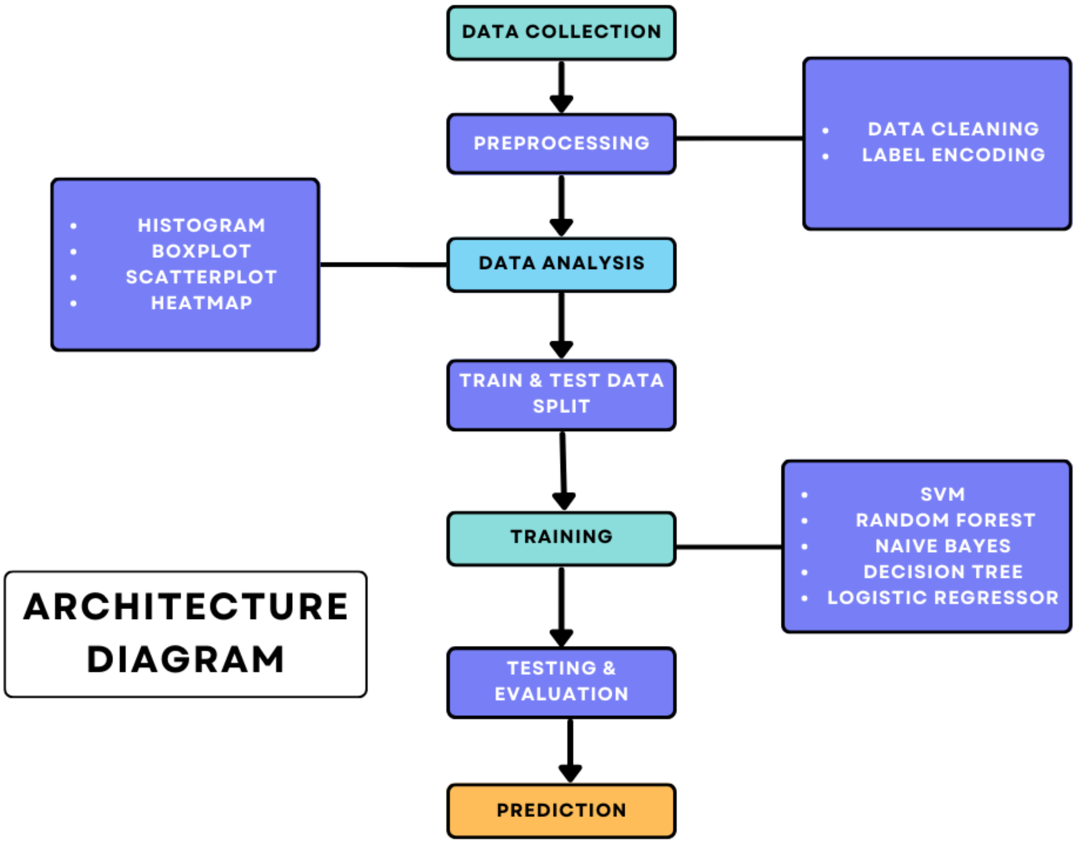
- 'User ID': unique identification for consumer
- 'Age': customer age in years
- 'Estimated Salary': Avg. Income of consumer
- 'Gender': Whether consumer was male or female
- 'Purchased': 0 or 1 indicated clicking on Ad

Among these 5 features, we consider Gender, Age, Estimated Salary and Visited details to train the model and accurately classify the advertisement visit for the given stakeholder. The sample of the dataset is as follows:



	A	B	C	D	E
1	User ID	Gender	Age	EstimatedSalary	Visited
2	15624510	Male	19	19000	0
3	15810944	Male	35	20000	0
4	15668575	Female	26	43000	0
5	15603246	Female	27	57000	0
6	15804002	Male	19	76000	0
7	15728773	Male	27	58000	0
8	15598044	Female	27	84000	0
9	15694829	Female	32	150000	1
10	15600575	Male	25	33000	0
11	15727311	Female	35	65000	0
12	15570769	Female	26	80000	0
13	15606274	Female	26	52000	0
14	15746139	Male	20	86000	0
15	15704987	Male	32	18000	0
16	15628972	Male	18	82000	0
17	15697686	Male	29	80000	0
18	15733883	Male	47	25000	1
19	15617482	Male	45	26000	1
20	15704583	Male	46	28000	1
21	15621083	Female	48	29000	1
22	15649487	Male	45	22000	1
23	15736760	Female	47	49000	1

V. PROPOSED ALGORITHM WITH FLOWCHART



The datasets are collected from various sources. Then, we pre – process the raw data through two sub steps and we analyse to understand the available data using statistical and visual techniques. Then, we split the available data. We train and validate the model using the training data. Then, we test and evaluate the model using the testing data, which can be used for classification.

DATA COLLECTION

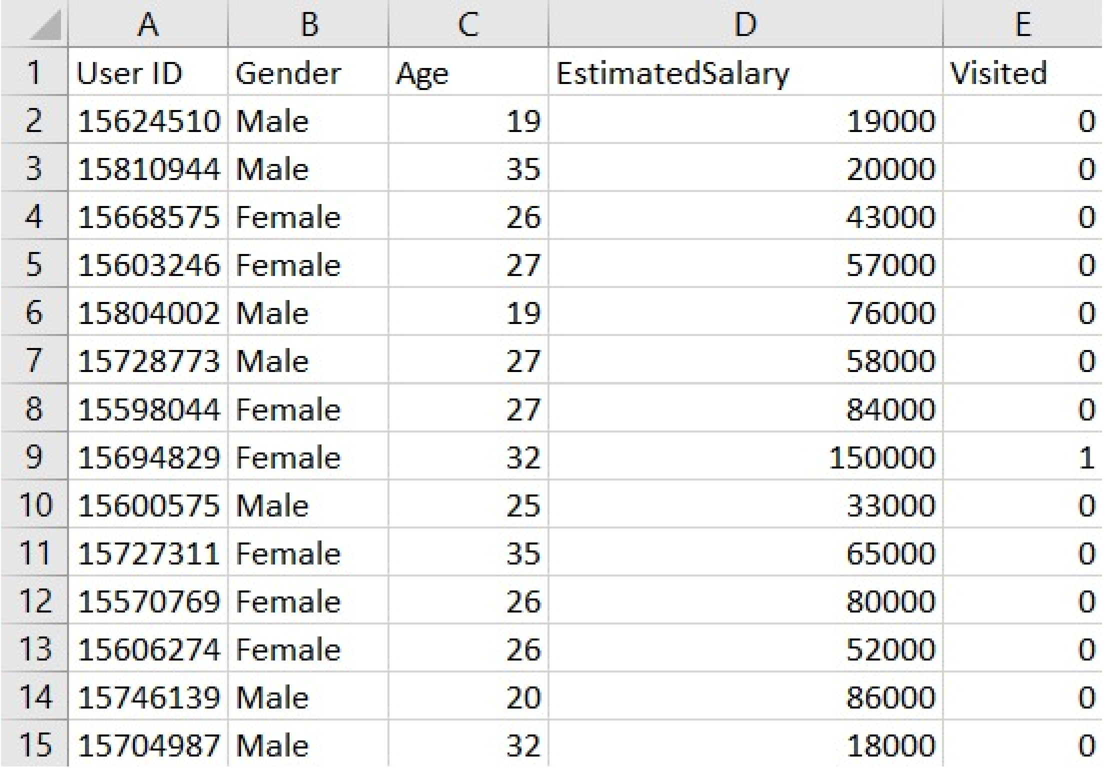
As per the literature survey, the datasets were collected from various sources / websites. These datasets are taken according to and in reference to the type of sector selected for the model. These datasets were semi – automatically cleaned and pre – processed to remove unwanted, correct and duplicate data. The final dataset is termed as a customer database and includes 401 records and 5 features.

The features are as follows:

- 'User ID': unique identification for consumer
- 'Age': customer age in years

- 'Estimated Salary': Avg. Income of consumer
- 'Gender': Whether consumer was male or female
- 'Purchased': 0 or 1 indicated clicking on Ad

Among these 5 features, we consider Gender, Age, Estimated Salary and Visited details to train the model and accurately classify the advertisement visit for the given stakeholder. The sample of the dataset is as follows:



	A	B	C	D	E
1	User ID	Gender	Age	EstimatedSalary	Visited
2	15624510	Male	19	19000	0
3	15810944	Male	35	20000	0
4	15668575	Female	26	43000	0
5	15603246	Female	27	57000	0
6	15804002	Male	19	76000	0
7	15728773	Male	27	58000	0
8	15598044	Female	27	84000	0
9	15694829	Female	32	150000	1
10	15600575	Male	25	33000	0
11	15727311	Female	35	65000	0
12	15570769	Female	26	80000	0
13	15606274	Female	26	52000	0
14	15746139	Male	20	86000	0
15	15704987	Male	32	18000	0

PREPROCESSING

In this step, we preprocess the raw data through two sub steps, they are as follows:

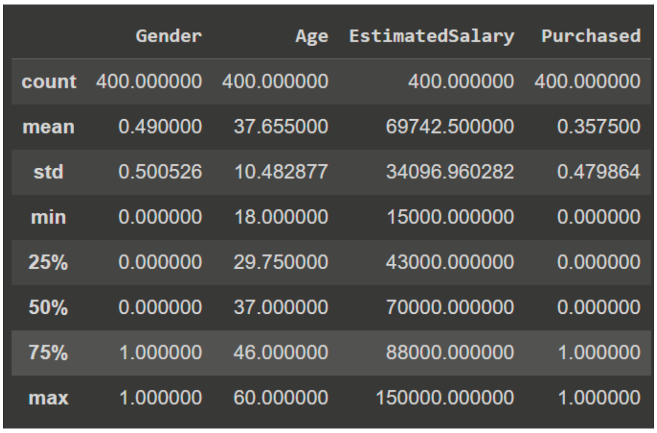
- Data cleaning
- Label Encoding

We remove the unnecessary columns, which is User ID, as it is a categorical variable so it will not be useful for building the model. We also encode the Gender column using 0 and 1, where 1 is Male and 0 is Female. We also remove any row with null value in order to ensure the sanity of the data. We further transform the raw data using Standard Scaler to make it better organized.

DATA ANALYSIS

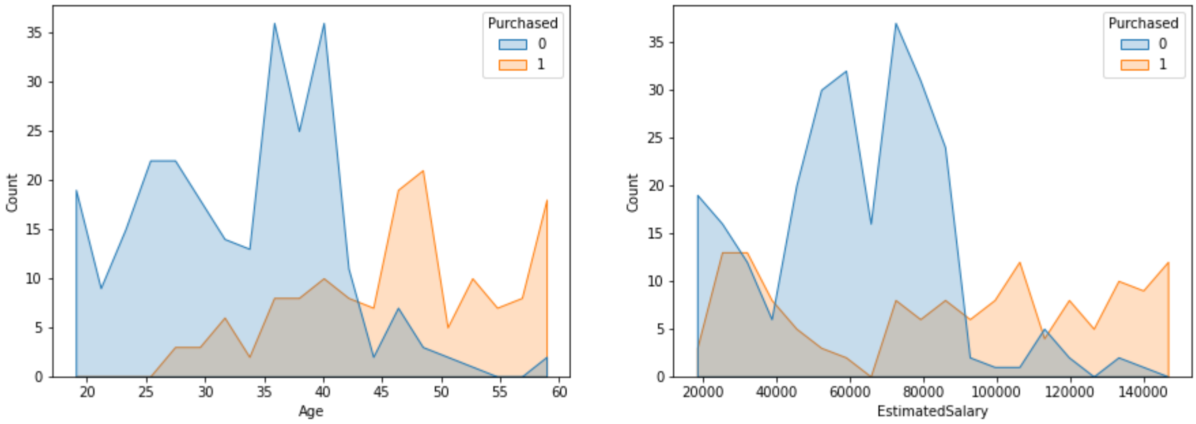
In this step, we analyse and understand the available data using statistical and visual techniques. We achieve the same using the following methods:

- Statistics

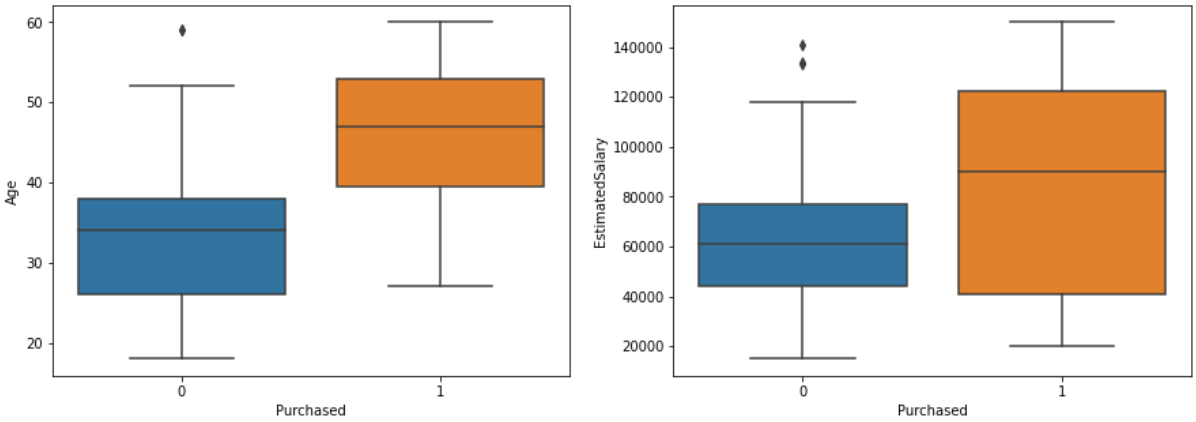


	Gender	Age	EstimatedSalary	Purchased
count	400.000000	400.000000	400.000000	400.000000
mean	0.490000	37.655000	69742.500000	0.357500
std	0.500526	10.482877	34096.960282	0.479864
min	0.000000	18.000000	15000.000000	0.000000
25%	0.000000	29.750000	43000.000000	0.000000
50%	0.000000	37.000000	70000.000000	0.000000
75%	1.000000	46.000000	88000.000000	1.000000
max	1.000000	60.000000	150000.000000	1.000000

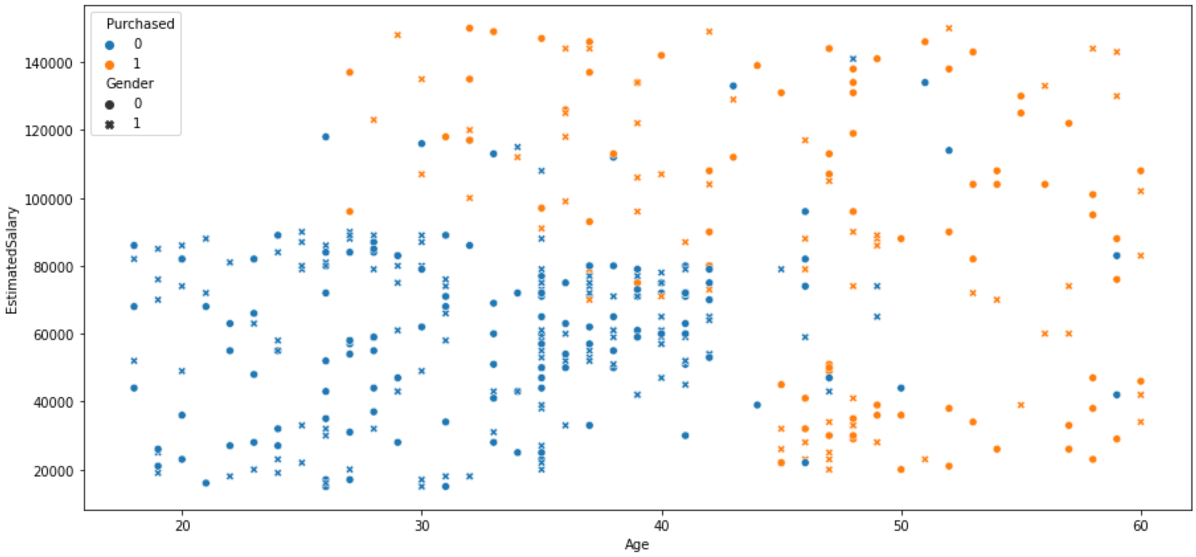
- Histogram



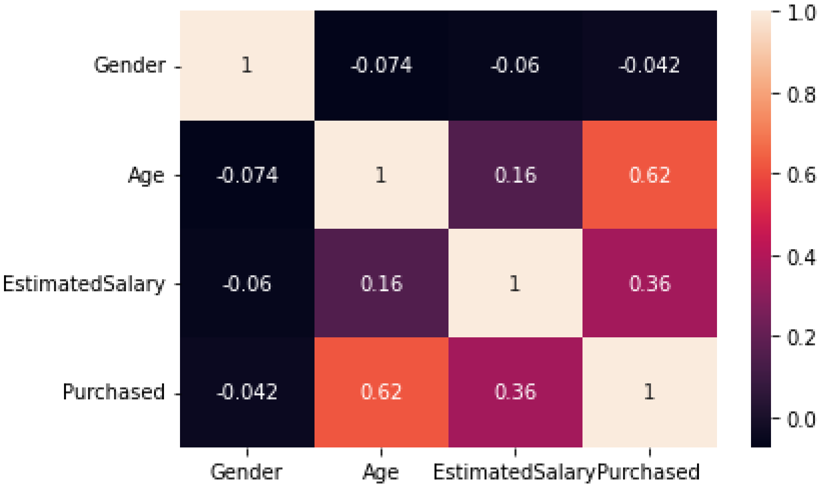
- Boxplot



- Scatterplot



- Heatmap



MODEL TRAINING

In this step, we split the available data into training and testing sets, so that we can use the training set to train and validate the model, whereas the testing set can be used to test and evaluate the model efficiently, the ratio of training and testing set is 60 : 40. We train the model using the following classification algorithms:

- Random Forest
- Decision Tree
- KNN
- Naive Bayes

- Logistic Regression
- SVM
- Grid Search CV
- AdaBoost
- Gradient Boosting
- XGB
- XGBRF
- LGBM

Once we train these models using the training data, we validate the models and understand the goodness of the model in terms of accuracy.

TESTING

In this step, we evaluate the trained models using the testing set and compare the models in terms of various metrics in order to better understand the efficiency of the models. The metrics are as follows:

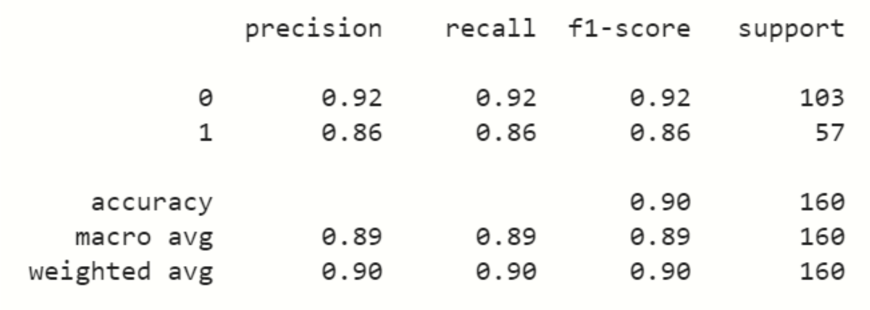
- Accuracy
- Precision
- Recall
- F1 Score
- Support
- Mean Squared Error

We can also visualize the performance of the models using the Confusion Matrix.

PREDICTION

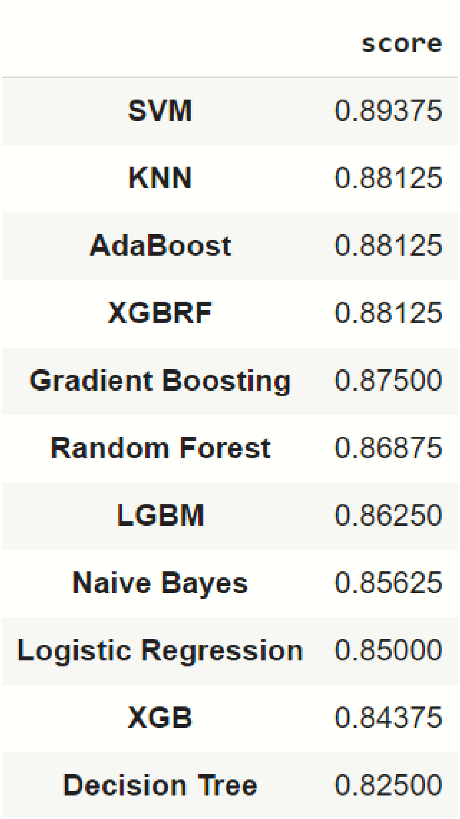
In this step, we can perform prediction using the developed model and be able to classify the user accurately with the least possible error. With the help of this model, we will be able to identify the potential customers and broadcast the ads efficiently. This will help any organization / company to reach the customer and enhance their service and ultimately optimize cost and methodology for marketing and advertising strategy.

VI. EXPERIMENTS RESULTS



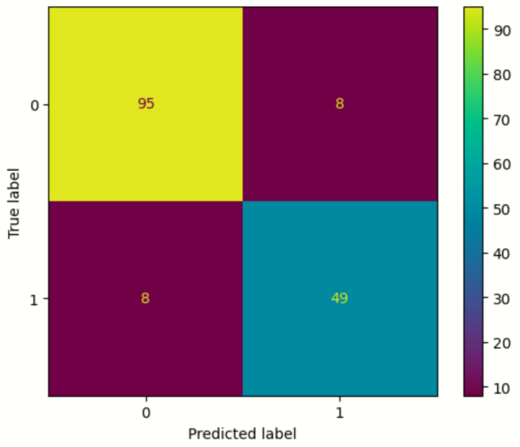
```
              precision    recall  f1-score   support

           0       0.92      0.92      0.92       103
           1       0.86      0.86      0.86        57

    accuracy                           0.90       160
   macro avg       0.89      0.89      0.89       160
weighted avg       0.90      0.90      0.90       160
```

Classification Report



	score
SVM	0.89375
KNN	0.88125
AdaBoost	0.88125
XGBRF	0.88125
Gradient Boosting	0.87500
Random Forest	0.86875
LGBM	0.86250
Naive Bayes	0.85625
Logistic Regression	0.85000
XGB	0.84375
Decision Tree	0.82500

Score Comparison Table



Confusion Matrix

VII. COMPARATIVE STUDY

We can see that the SVM algorithm has performed the best with an accuracy of 0.89375. Support Vector Machines (SVM) are powerful algorithms that can work well for both linear and non-linear classification problems. SVM tries to find the best separating hyperplane between the classes in the data. SVM can be used for both binary and multi-class classification problems. The second-best algorithm in terms of accuracy is KNN, AdaBoost, and XGBRF, all with an accuracy of 0.88125. KNN or K – Nearest Neighbors is a non – parametric algorithm used for both classification and regression problems. In KNN, a new data point is classified based on the majority class of its K – nearest neighbors. AdaBoost or Adaptive Boosting is an ensemble learning algorithm that combines weak learners to create a strong classifier. XGBRF is an implementation of the gradient boosting algorithm that is optimized for speed and performance. Gradient Boosting, Random Forest, and LGBM algorithms have accuracy in the range of 0.86250 – 0.87500. Gradient Boosting and Random Forest are ensemble learning algorithms that use decision trees as their base learners. Gradient Boosting tries to minimize the loss function by iteratively adding decision trees, while Random Forest creates a forest of decision trees and uses their output to make the final classification. LGBM or Light Gradient Boosting Machine is a gradient boosting algorithm that is optimized for speed and efficiency. Naive Bayes, Logistic Regression, XGB, and Decision Tree algorithms have the lowest accuracy in the range of 0.82500 – 0.85625. Naive Bayes is a probabilistic algorithm based on Bayes' theorem, used for both classification and regression problems. Logistic Regression is a linear algorithm used for binary classification problems. XGB or Extreme Gradient Boosting is an implementation of the gradient boosting algorithm that is optimized for accuracy and speed. Decision Tree is a non – parametric algorithm used for both classification and regression problems. Overall, SVM has outperformed all other algorithms in terms of accuracy, followed by KNN, AdaBoost, and XGBRF.

VIII. CONCLUSION AND FUTURE WORK

The Classification report shows the precision, recall, and f1 – score for a binary classification problem with two classes, 0 and 1. The support column shows the number of instances for each class in the dataset, and the weighted average provides an overall performance metric. From the table we can infer that the classifier has an overall accuracy of 0.90, which means that 90% of the predictions are correct. The precision for class 0 is 0.92, which means that out of all the instances predicted as class 0, 92% were correct. The recall for class 0 is also 0.92, which means that out of all the instances that actually belong to class 0, 92% were correctly classified. The f1 – score for class 0 is 0.92, which is the harmonic mean of precision and recall for that class. Similarly, for class 1, the precision is 0.86, which means that out of all the instances predicted as class 1, 86% were correct. The recall for class 1 is also 0.86, which means that out of all the instances that actually belong to class 1, 86% were correctly classified. The f1 – score for class 1 is 0.86. The macro average of precision, recall, and f1 – score is 0.89, which is the average of the performance metrics for both classes. The weighted average of precision, recall, and f1 – score is also 0.90, which takes into account the support for each class. Overall, the classifier has good performance with an accuracy of 0.90 and high precision and recall for both classes. It is always recommended to evaluate multiple algorithms and compare their performance metrics to select the best algorithm for a given problem. While accuracy is an important metric, it is not the only factor to consider when evaluating classification algorithms. Other factors to consider include precision, recall, f1 – score, and the computational complexity of the algorithm, now that we have a detailed result, we can conclude that Support Vector Machines (SVM) is the most accurate algorithm with an accuracy of 0.89375, which clearly states that SVM is the most suitable model and popular choice for advertisement classification because they are particularly effective in handling high-dimensional datasets with complex features. In advertising classification, the input data can include various types of features such as user demographics, browsing behaviour, and other contextual factors. SVM can handle this type of high-dimensional data effectively by finding the best hyperplane to separate the data into different classes. So, SVM's ability to handle high-dimensional data and capture complex relationships between variables makes it an effective algorithm for advertisement classification

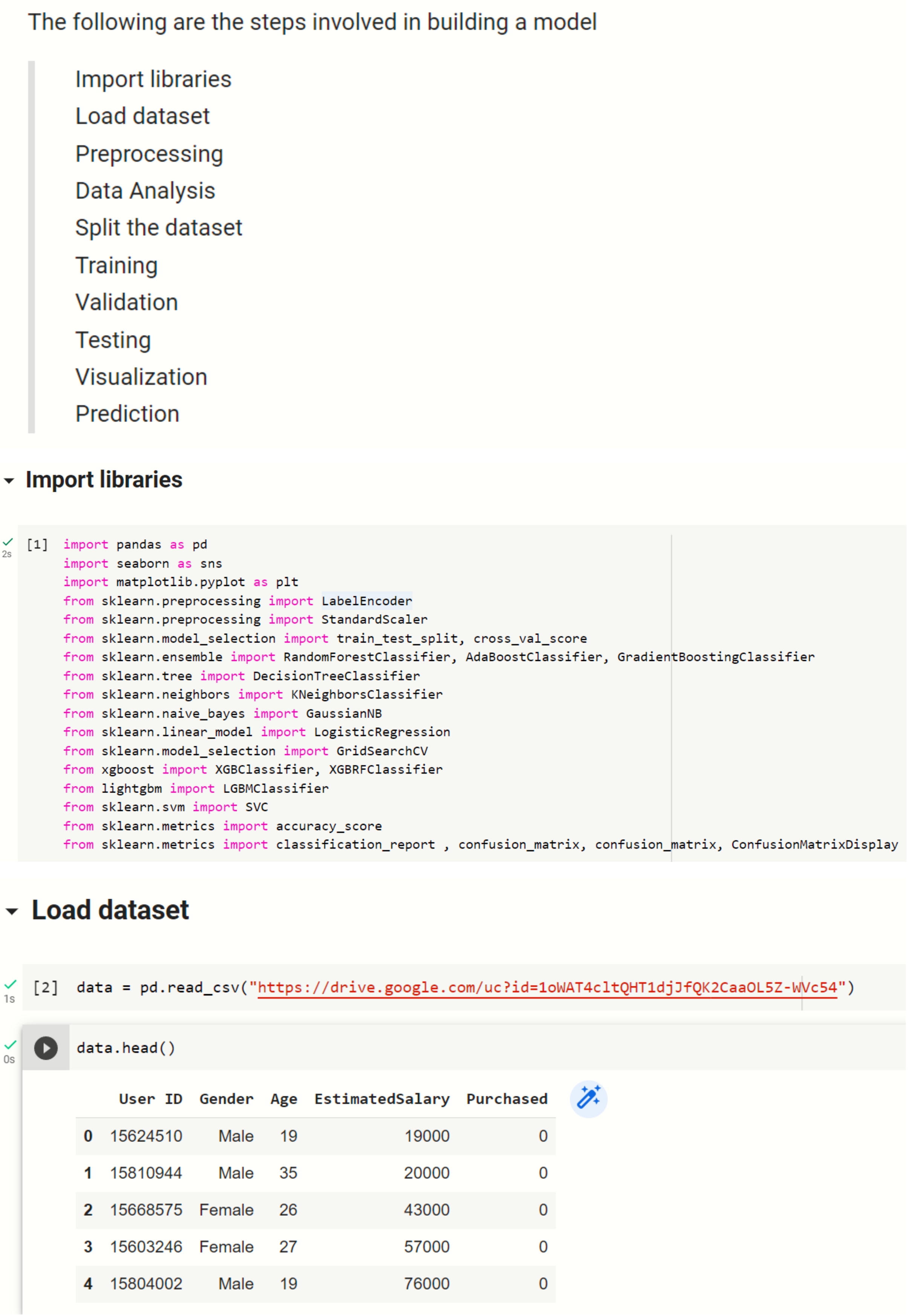
IX. REFERENCES

1. Lee, D., Hosanagar, K., & Nair, H, Advertising Content and Consumer Engagement on Social Media: Evidence from Facebook, 2018.
2. Norjihan Abdul Ghani, Suraya Hamid and Ejaz Ahmed, Computers in Human Behaviour, Social media big data analytics, Volume 101, Pages 417-428, 2019.
3. Mahmud Akhter Shareef, Bhasker Mukerji, Yogesh K. Dwivedi, Nripendra P. Rana and Rubina Islam, Social media marketing: Comparative effect of advertisement sources, Journal of Retailing and Consumer Services, Volume 46, 2019.
4. Somya Ranjan Sahoo and Brij Bhooshan Gupta, Classification of various attacks and their defence mechanism in online social networks: a survey, Enterprise information systems, April, 2019.
5. Serkan Savas and Nurettin Topaloglu, Data Analysis through social media according to the classified crime, Turkish Journal of Electrical Engineering and Computer Sciences,27(1):407-420, 2019.
6. E. Phaisangittisagul, Y. Koobkrabee, K. Wirojborisuth, T. Ratanasrimetha and S. Aummaro, "Target Advertising Classification using Combination of Deep Learning and Text model," 2019 10th International Conference of Information and Communication Technology for Embedded Systems (IC-ICTES), 2019.
7. B. Senthil Arasu, B. Jonath Backia Seelan, N. Thamaraiselvan, A machine learning-based approach to enhancing social media marketing, Computers & Electrical Engineering, Volume 86, 2020.
8. Gustavo Nogueira de Sousa, Gustavo R T De Almeida and Fabio Lobato, Social Network Advertising Classification Based on Content Categories, Business Information Systems Workshops, 2019.
9. K. Pornsawangdee and U. Taetragool, Pattern Recognition of Effective Online Classified Advertisement, 2019 IEEE 2nd International Conference on Knowledge Innovation and Invention (ICKII), Seoul, Korea (South), pp. 429-432, 2019.
10. Weinan Zhang, Tianming Du and Jun Wang, Deep Learning over Multi – Field Categorical Data, Conference: European Conference on Information Retrieval, 2020.
11. Mirjana Pejic-Bach, Tine Bertoncel, Maja Meško, Živko Krstić, Text mining of industry 4.0 job advertisements, International Journal of Information Management, Volume 50, 2020.

12. Chuanpeng Yu, Zhengang Zhang, Chunpei Lin, Yenchun Jim Wu, Can data-driven precision marketing promote user ad clicks? Evidence from advertising in WeChat moments, Industrial Marketing Management, Volume 90, 2020.
13. Zhabiz Gharibshah, Xingquan Zhu, Arthur Hinline and Michael Conway, Deep Learning for User Interest and Response Prediction in Online Display Advertising, Data Science and Engineering, 2020.
14. P. Bhowmik, M. Sohrawordi, U. A. M. Ehsan Ali, M. N. Hasan and P. K. Roy, Analysis of Social Media Data to Classify and Detect Frequent Issues Using Machine Learning Approach,2020 2nd International Conference on Advanced Information and Communication Technology (ICAICT), Dhaka, Bangladesh, pp. 394-399, 2020.
15. Taekeun Hong, Jin A Choi and Kiho Lim, Enhancing Personalized Ads Using Interest Category Classification of SNS Users Based on Deep Neural Networks, Sensors, 2020.
16. A. M. Ud Din Khanday, Q. Rayees Khan and S. T. Rabani, Analysing and Predicting Propaganda on Social Media using Machine Learning Techniques, 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), Greater Noida, India, 2020.
17. Yan Sun, Qian Wu and Wendi Li, A Push Model of Advertisement Classification Matching Based on Machine Learning, IOP Conference Series Materials Science and Engineering, 2020.
18. N. Hudson, H. Khamfroush, B. Harrison and A. Craig, Smart Advertisement for Maximal Clicks in Online Social Networks Without User Data," 2020 IEEE International Conference on Smart Computing (SMARTCOMP), 2020.
19. A. Garg, N. Kapil and S. Shukla, Prediction of the Target User Segment for Paid Advertisement, IEEE 8th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON), 2021.
20. Demetrios Vakratsas and Shane Wang, Artificial Intelligence in Advertising Creativity, 2021, Journal of Advertising, 2020.
21. Yingshuo Jiang, Li Shi, Zhaoxin Zhang and Yongdong Xu, Mobile Terminal Advertising Video Acquisition and Classification, Academic Journal of computing and Information Science, 2021.
22. Mohammed Al-Sarem, Faisal Saeed and Zeyad Ghaleb Al-Mekhlafi, An Improved Multiple Features and Machine Learning-Based Approach for Detecting Clickbait News on Social Networks, Applied sciences 11(20):9487, 2021.

23. Pooja Jain, Kavita Taneja and Harmunish Taneja, Convolutional Neural Network Based Intelligent Advertisement Search Framework for Online English Newspapers, Recent Patents on Engineering, Vol 15, 2021.
24. Ebadi Jokandan, Seyed Mohsen, Bayat, Peyman and Mehdi, Targeted Advertising in Social Media Platforms Using Hybrid Convolutional Learning Method besides Efficient Feature Weights, Journal of Electrical and Computer Engineering p1-17.17p, 2022.
25. Barween Al Kurdi, Muhammad Turki Alshurideh, Haitham M. Alzoubi and Iman Akour, The role of digital marketing channels on consumer buying decisions through eWOM in the Jordanian markets, International Journal of Data and Network Science,6(4): P1175-1186, 2022.
26. Shittu Olumide Ayodeji, Machine Learning for Classifying Social Media Ads, 2022.
27. Dang Phuc, Dau Hieu, Nguyen Hoang, Tran Khoa, Apply Deep Learning in Real-time Customer Detection and Classification System for Advertisement Decision Making at Supermarket, In proceedings of the 11th International Conference on smart cities and green ICT systems, Volume 1: SMARTGREENS, 2022.
28. M. Alhalabi, N. Hussein, E. Khan, O. Habash, J. Yousaf and M. Ghazal, Sustainable Smart Advertisement Display Using Deep Age and Gender Recognition,2021 International Conference on Decision Aid Sciences and Application (DASA), Sakheer, Bahrain, 2021.
29. Farzaneh Jouyandeh, Pooya Moradian Zadeh, IPARS: An Image-based Personalized Advertisement Recommendation System on Social Networks, Procedia Computer Science, Volume 201, 2022.
30. Eike Jens Hoffmann, Karam Abdulahhad, Xiao Xiang Zhu, Using social media images for building function classification, Cities, Volume 133, 2023.

Appendix



The following are the steps involved in building a model

> Import libraries
> Load dataset
> Preprocessing
> Data Analysis
> Split the dataset
> Training
> Validation
> Testing
> Visualization
> Prediction

Import libraries

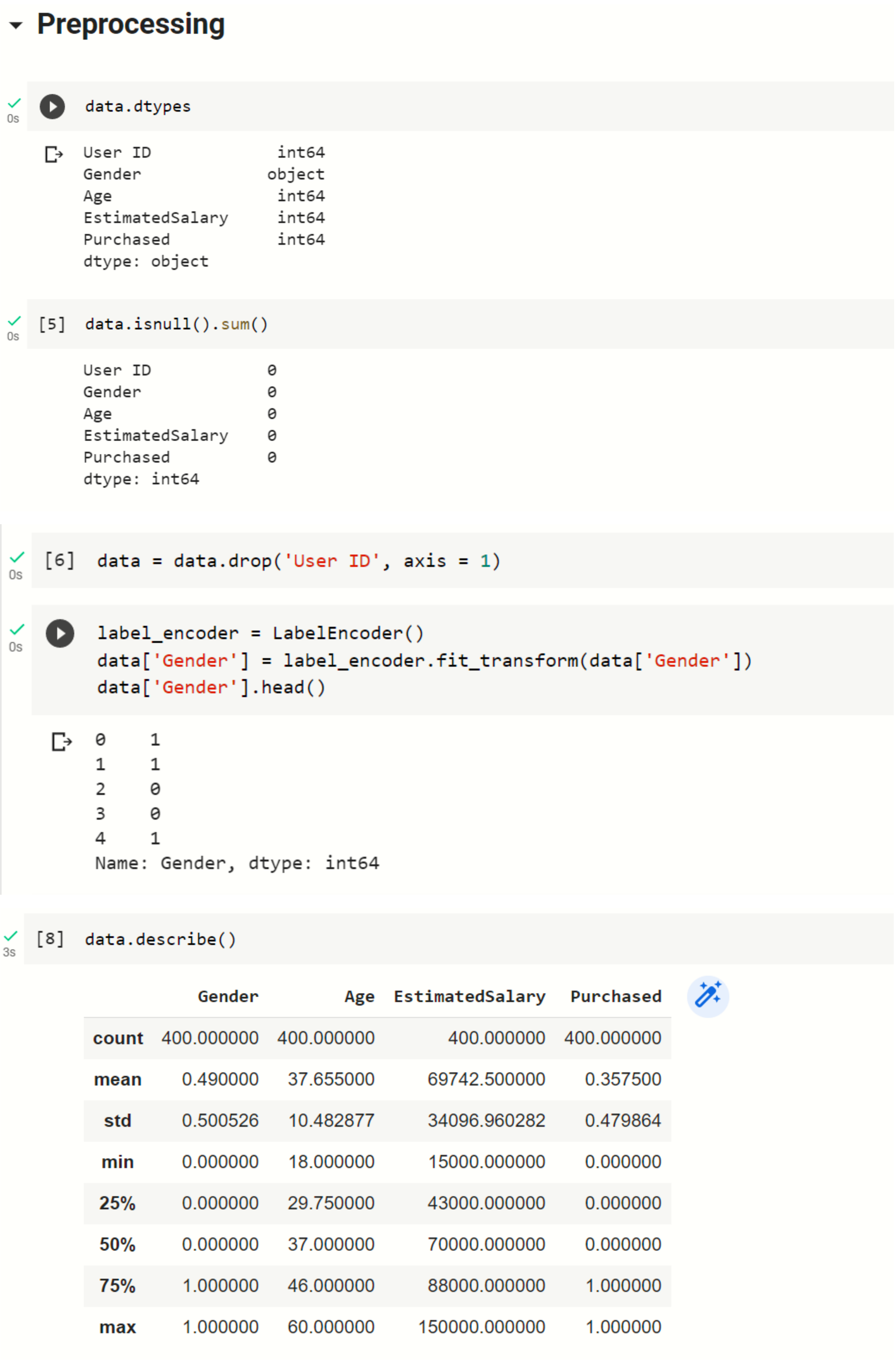
```
[1]  import pandas as pd
     import seaborn as sns
     import matplotlib.pyplot as plt
     from sklearn.preprocessing import LabelEncoder
     from sklearn.preprocessing import StandardScaler
     from sklearn.model_selection import train_test_split, cross_val_score
     from sklearn.ensemble import RandomForestClassifier, AdaBoostClassifier, GradientBoostingClassifier
     from sklearn.tree import DecisionTreeClassifier
     from sklearn.neighbors import KNeighborsClassifier
     from sklearn.naive_bayes import GaussianNB
     from sklearn.linear_model import LogisticRegression
     from sklearn.model_selection import GridSearchCV
     from xgboost import XGBClassifier, XGBRFClassifier
     from lightgbm import LGBMClassifier
     from sklearn.svm import SVC
     from sklearn.metrics import accuracy_score
     from sklearn.metrics import classification_report , confusion_matrix, confusion_matrix, ConfusionMatrixDisplay
```

Load dataset

```
[2]  data = pd.read_csv("https://drive.google.com/uc?id=1oWAT4cltQHT1djJfQK2CaaOL5Z-WVc54")
```

```
data.head()
```

	User ID	Gender	Age	EstimatedSalary	Purchased
0	15624510	Male	19	19000	0
1	15810944	Male	35	20000	0
2	15668575	Female	26	43000	0
3	15603246	Female	27	57000	0
4	15804002	Male	19	76000	0



Preprocessing

```
data.dtypes
```

```
User ID             int64
Gender             object
Age                 int64
EstimatedSalary     int64
Purchased           int64
dtype: object
```

```
[5] data.isnull().sum()
```

```
User ID            0
Gender             0
Age                0
EstimatedSalary    0
Purchased          0
dtype: int64
```

```
[6] data = data.drop('User ID', axis = 1)
```

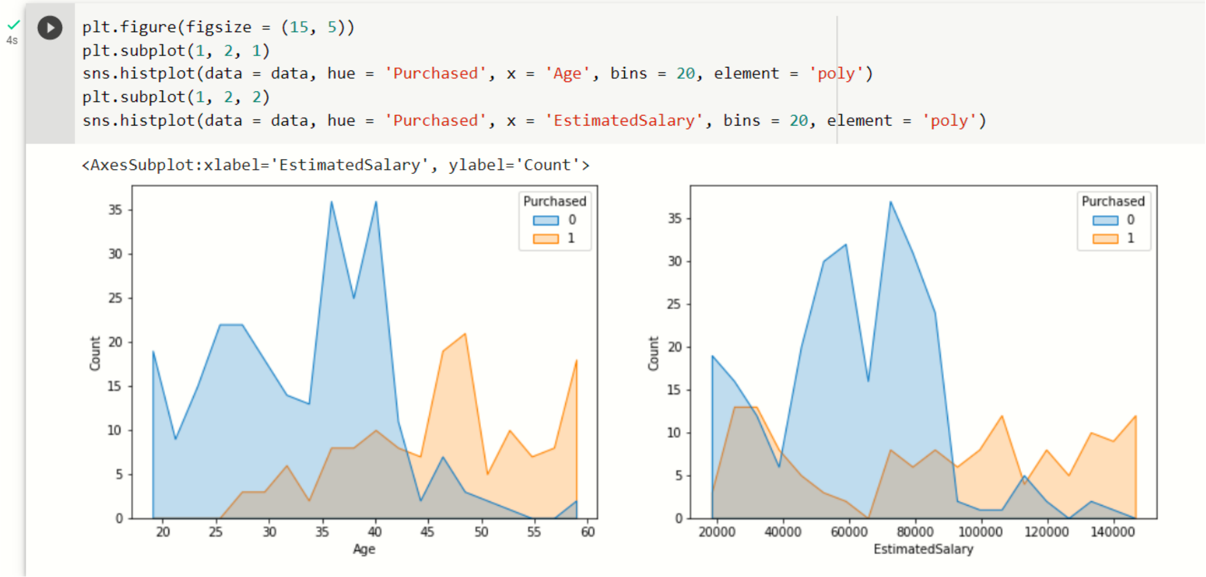
```
label_encoder = LabelEncoder()
data['Gender'] = label_encoder.fit_transform(data['Gender'])
data['Gender'].head()
```

```
0    1
1    1
2    0
3    0
4    1
Name: Gender, dtype: int64
```

```
[8] data.describe()
```

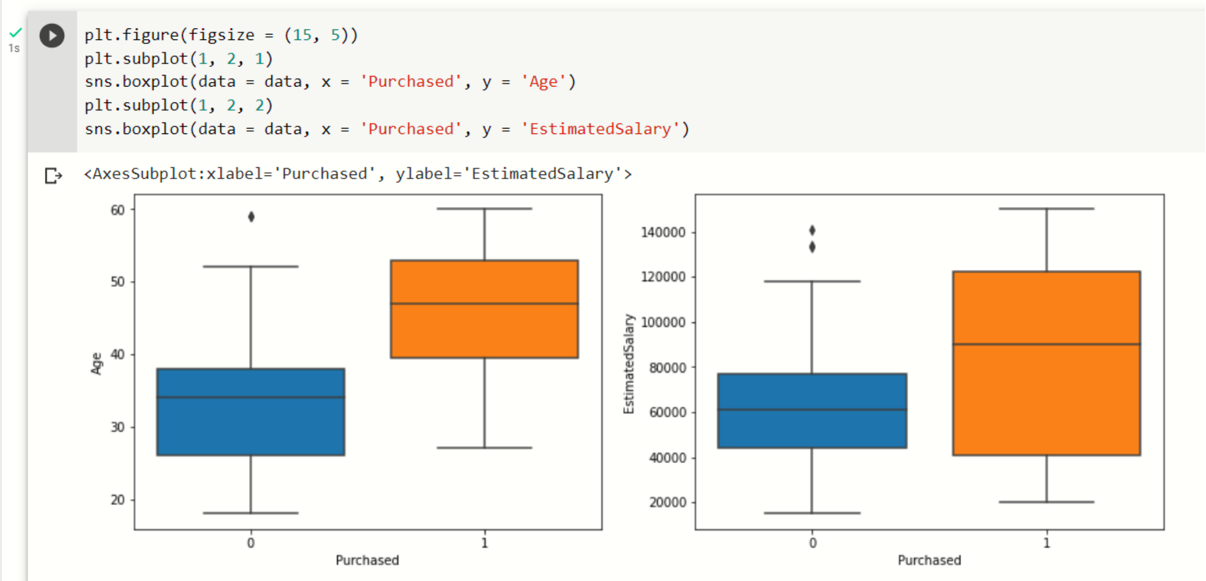
	Gender	Age	EstimatedSalary	Purchased
count	400.000000	400.000000	400.000000	400.000000
mean	0.490000	37.655000	69742.500000	0.357500
std	0.500526	10.482877	34096.960282	0.479864
min	0.000000	18.000000	15000.000000	0.000000
25%	0.000000	29.750000	43000.000000	0.000000
50%	0.000000	37.000000	70000.000000	0.000000
75%	1.000000	46.000000	88000.000000	1.000000
max	1.000000	60.000000	150000.000000	1.000000

Data Analysis



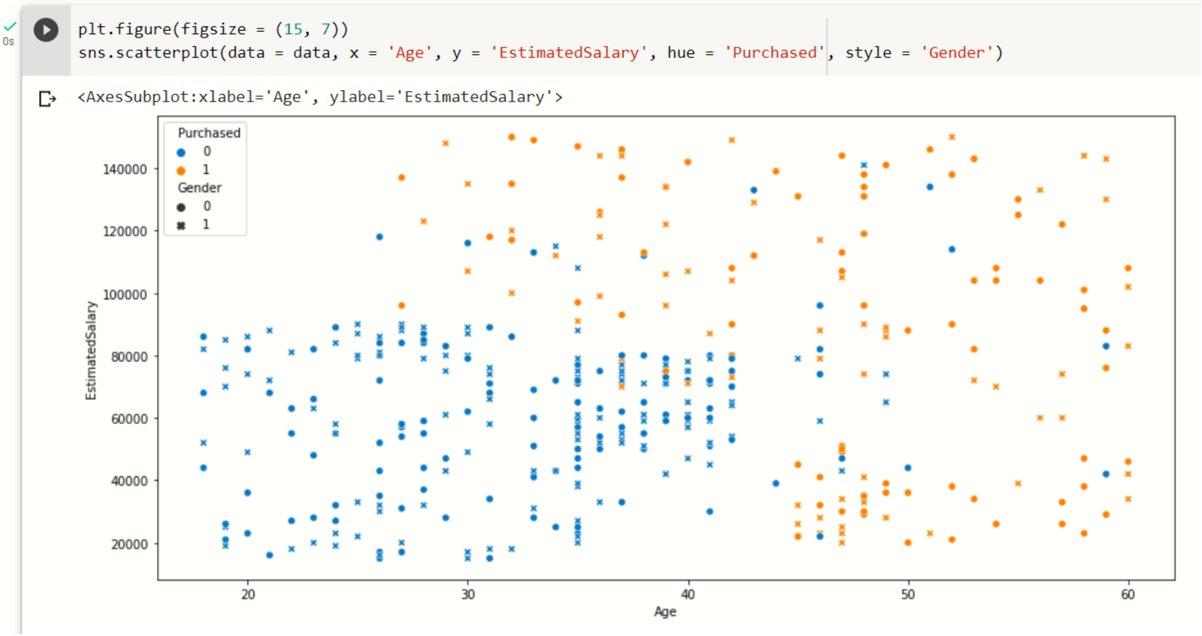
```
plt.figure(figsize = (15, 5))
plt.subplot(1, 2, 1)
sns.histplot(data = data, hue = 'Purchased', x = 'Age', bins = 20, element = 'poly')
plt.subplot(1, 2, 2)
sns.histplot(data = data, hue = 'Purchased', x = 'EstimatedSalary', bins = 20, element = 'poly')
```

```
<AxesSubplot:xlabel='EstimatedSalary', ylabel='Count'>
```



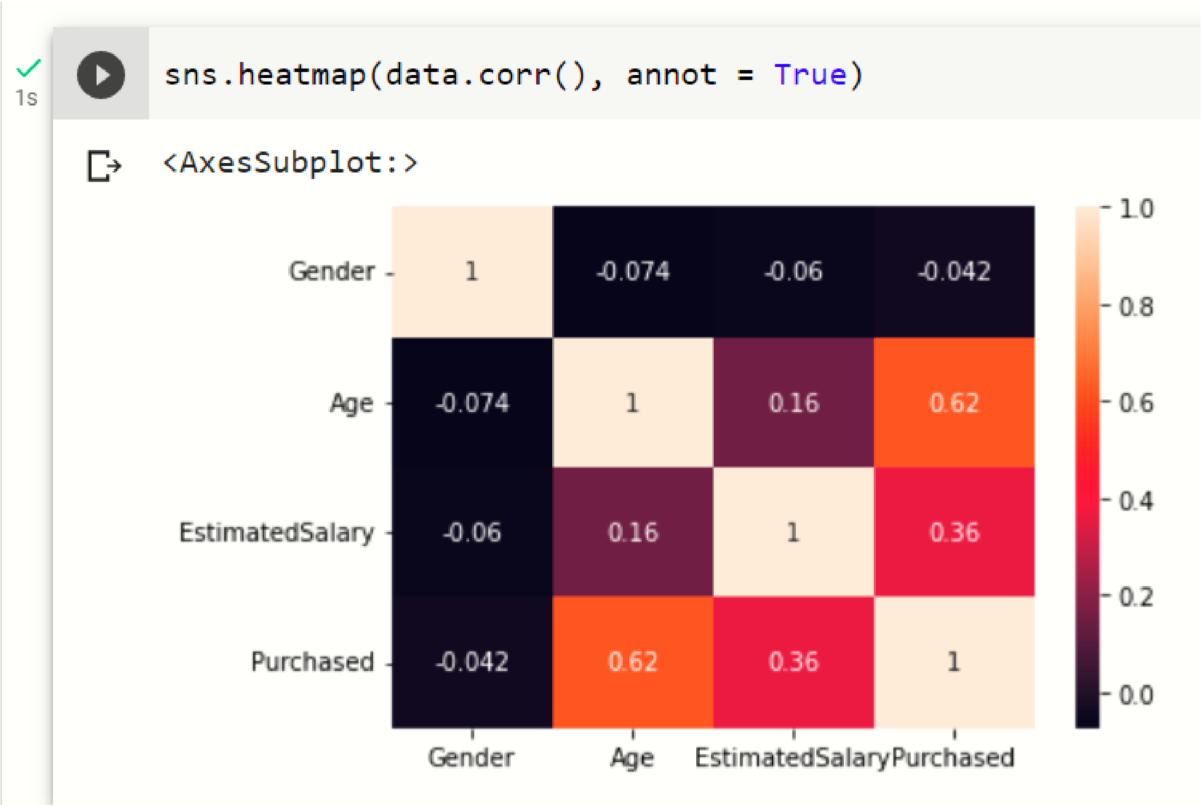
```
plt.figure(figsize = (15, 5))
plt.subplot(1, 2, 1)
sns.boxplot(data = data, x = 'Purchased', y = 'Age')
plt.subplot(1, 2, 2)
sns.boxplot(data = data, x = 'Purchased', y = 'EstimatedSalary')
```

```
<AxesSubplot:xlabel='Purchased', ylabel='EstimatedSalary'>
```



```
plt.figure(figsize = (15, 7))
sns.scatterplot(data = data, x = 'Age', y = 'EstimatedSalary', hue = 'Purchased', style = 'Gender')
```

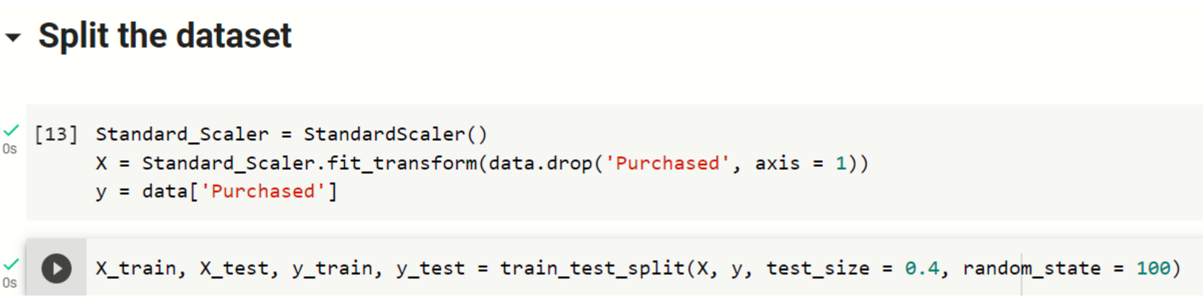
```
<AxesSubplot:xlabel='Age', ylabel='EstimatedSalary'>
```



```
sns.heatmap(data.corr(), annot = True)
```

```
<AxesSubplot:>
```

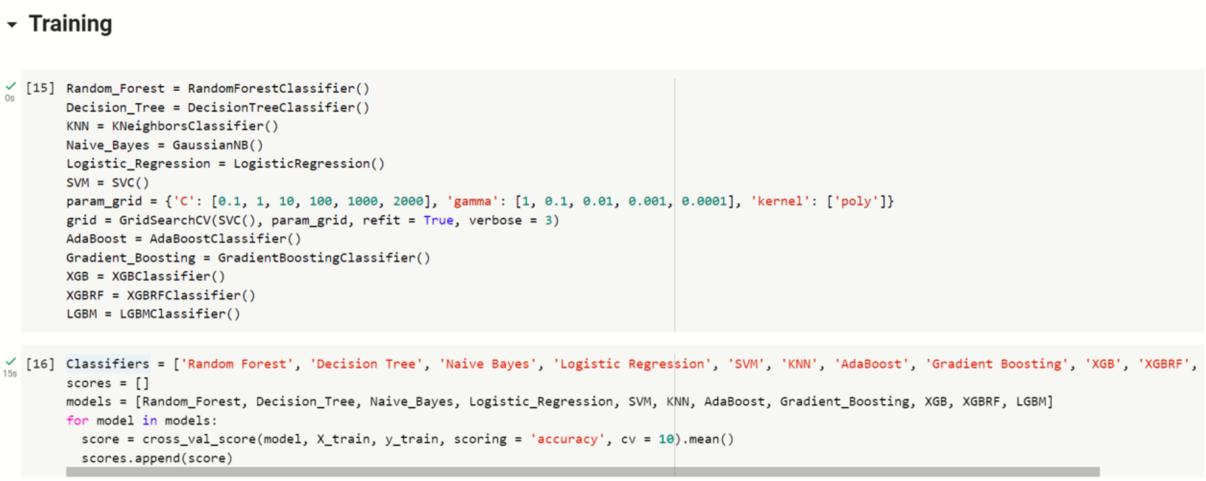
	Gender	Age	EstimatedSalary	Purchased
Gender	1	-0.074	-0.06	-0.042
Age	-0.074	1	0.16	0.62
EstimatedSalary	-0.06	0.16	1	0.36
Purchased	-0.042	0.62	0.36	1



Split the dataset

```
[13] Standard_Scaler = StandardScaler()
     X = Standard_Scaler.fit_transform(data.drop('Purchased', axis = 1))
     y = data['Purchased']
```

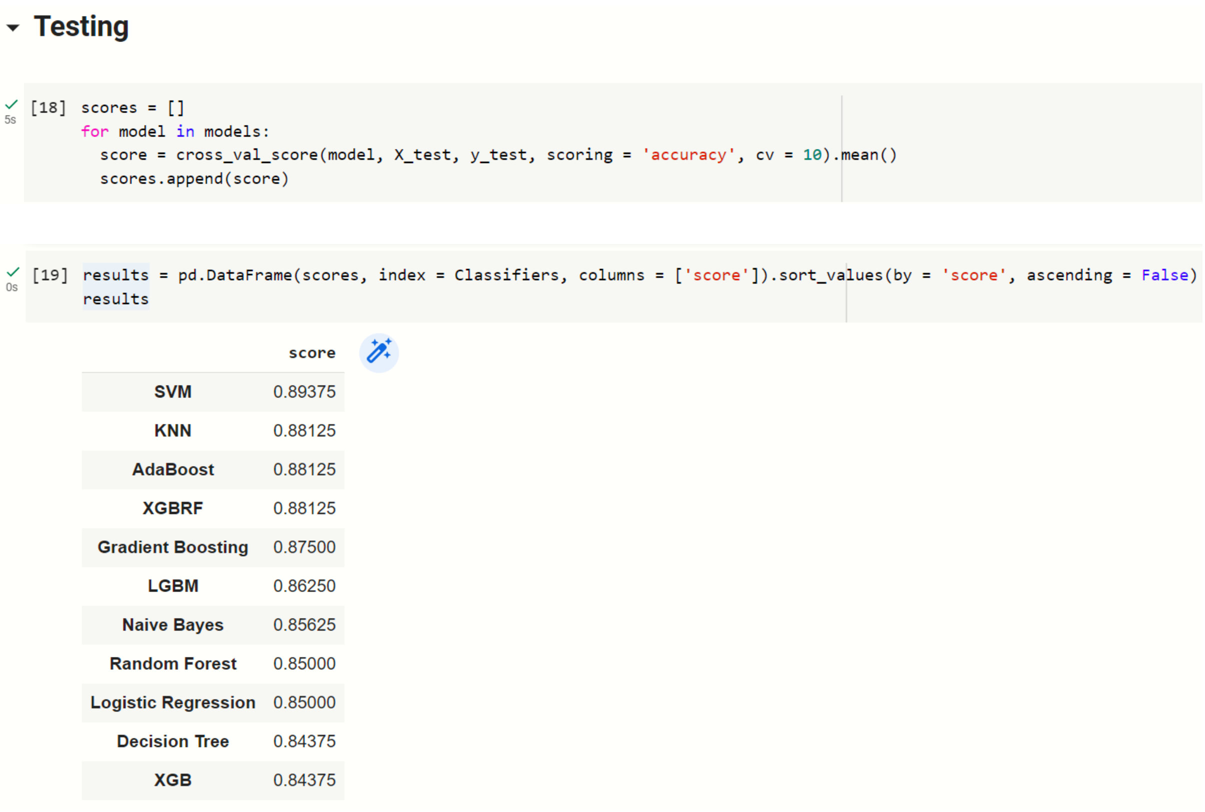
```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.4, random_state = 100)
```



Training

```
[15] Random_Forest = RandomForestClassifier()
     Decision_Tree = DecisionTreeClassifier()
     KNN = KNeighborsClassifier()
     Naive_Bayes = GaussianNB()
     Logistic_Regression = LogisticRegression()
     SVM = SVC()
     param_grid = {'C': [0.1, 1, 10, 100, 1000, 2000], 'gamma': [1, 0.1, 0.01, 0.001, 0.0001], 'kernel': ['poly']}
     grid = GridSearchCV(SVC(), param_grid, refit = True, verbose = 3)
     AdaBoost = AdaBoostClassifier()
     Gradient_Boosting = GradientBoostingClassifier()
     XGB = XGBClassifier()
     XGBRF = XGBRFClassifier()
     LGBM = LGBMClassifier()
```

```
[16] Classifiers = ['Random Forest', 'Decision Tree', 'Naive Bayes', 'Logistic Regression', 'SVM', 'KNN', 'AdaBoost', 'Gradient Boosting', 'XGB', 'XGBRF',
     scores = []
     models = [Random_Forest, Decision_Tree, Naive_Bayes, Logistic_Regression, SVM, KNN, AdaBoost, Gradient_Boosting, XGB, XGBRF, LGBM]
     for model in models:
       score = cross_val_score(model, X_train, y_train, scoring = 'accuracy', cv = 10).mean()
       scores.append(score)
```

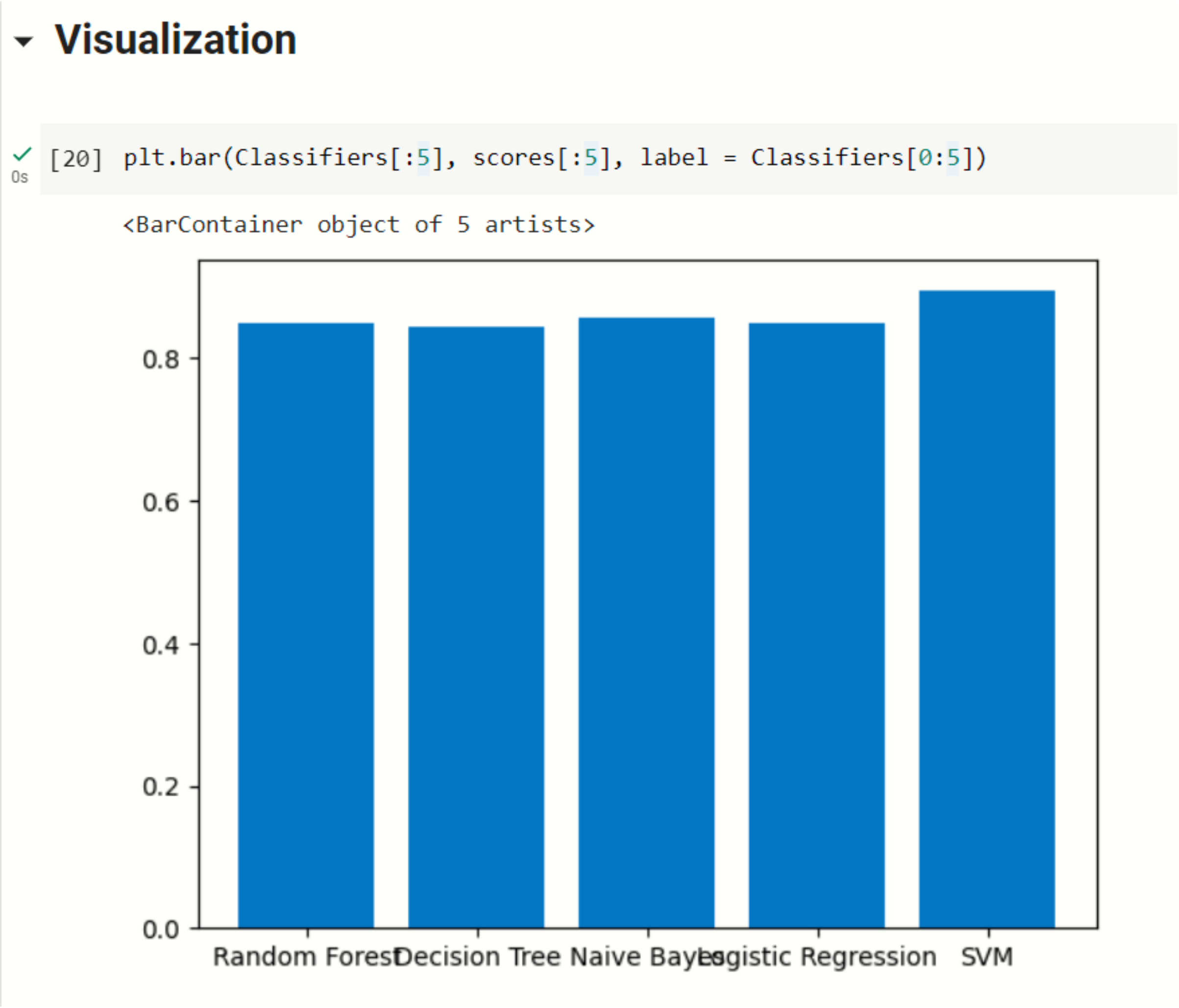


Testing

```
[18] scores = []
     for model in models:
       score = cross_val_score(model, X_test, y_test, scoring = 'accuracy', cv = 10).mean()
       scores.append(score)
```

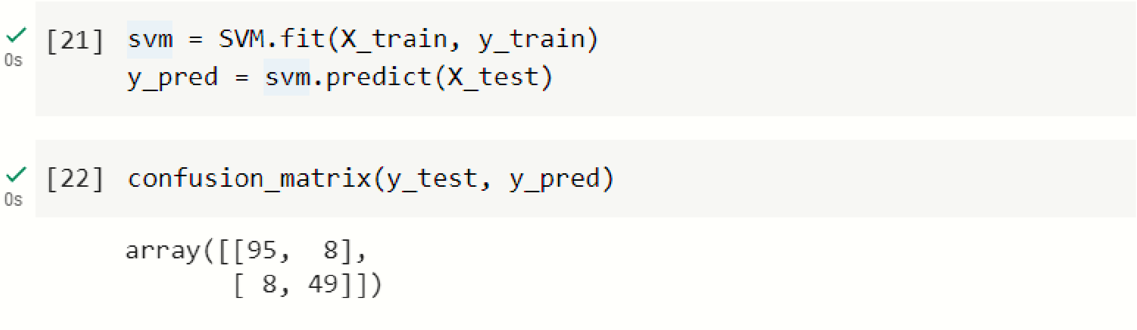
```
[19] results = pd.DataFrame(scores, index = Classifiers, columns = ['score']).sort_values(by = 'score', ascending = False)
     results
```

	score
SVM	0.89375
KNN	0.88125
AdaBoost	0.88125
XGBRF	0.88125
Gradient Boosting	0.87500
LGBM	0.86250
Naive Bayes	0.85625
Random Forest	0.85000
Logistic Regression	0.85000
Decision Tree	0.84375
XGB	0.84375



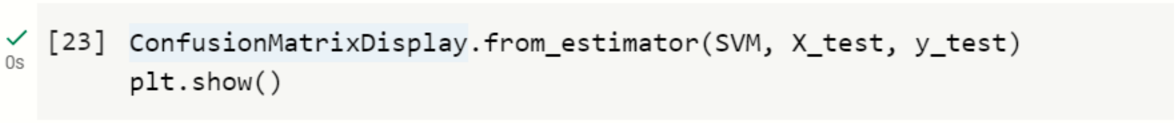
Visualization

```
[20] plt.bar(Classifiers[:5], scores[:5], label = Classifiers[0:5])

<BarContainer object of 5 artists>
```

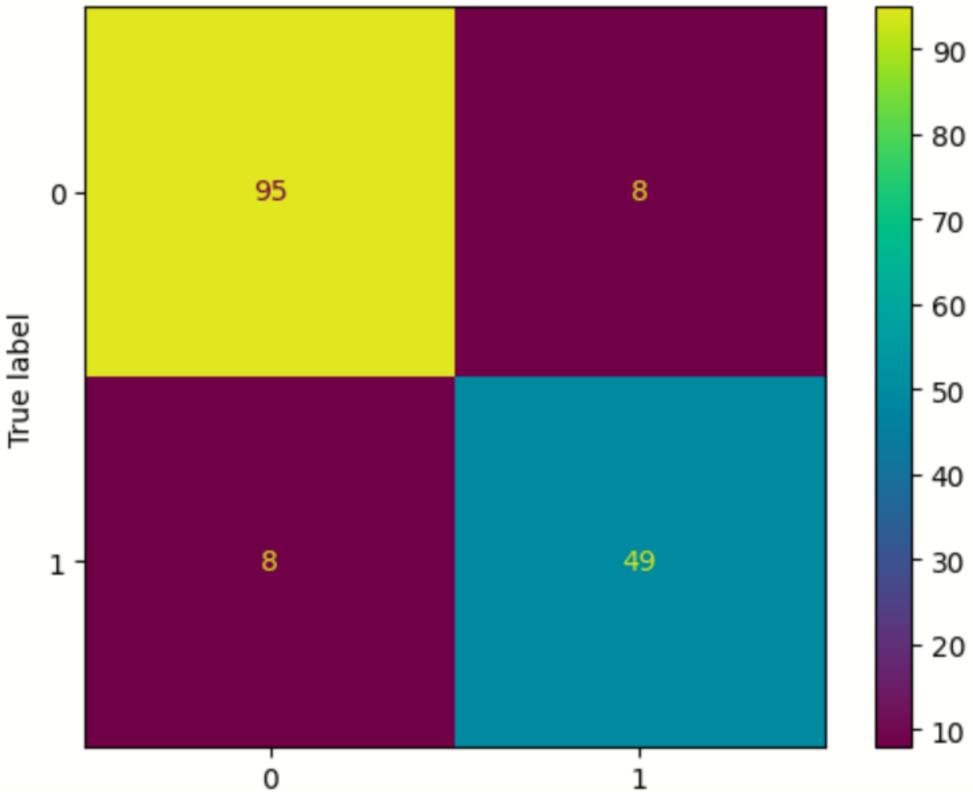


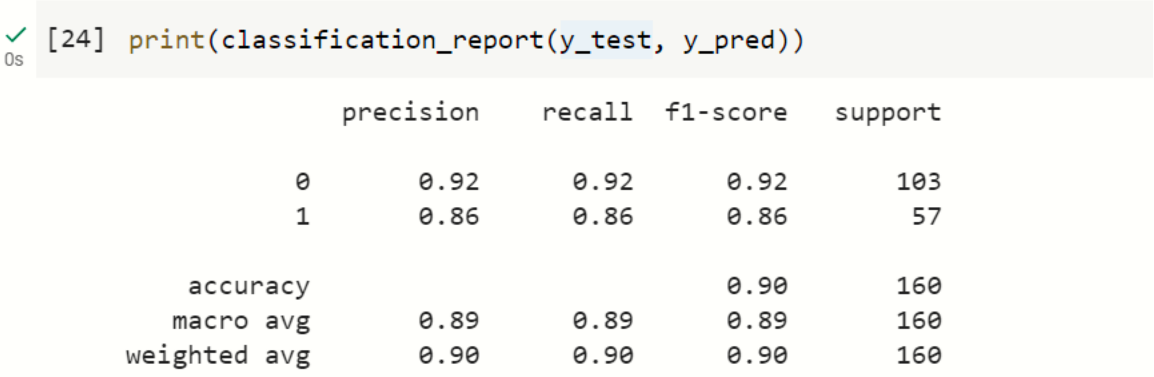
```
[21] svm = SVM.fit(X_train, y_train)
     y_pred = svm.predict(X_test)
```

```
[22] confusion_matrix(y_test, y_pred)

array([[95,  8],
       [ 8, 49]])
```



```
[23] ConfusionMatrixDisplay.from_estimator(SVM, X_test, y_test)
     plt.show()
```





```
[24] print(classification_report(y_test, y_pred))

              precision    recall  f1-score   support

           0       0.92      0.92      0.92       103
           1       0.86      0.86      0.86        57

    accuracy                           0.90       160
   macro avg       0.89      0.89      0.89       160
weighted avg       0.90      0.90      0.90       160
```