Vishwakarma Institute of Information Technology, Pune

**(An Autonomous Institute Maharashtra)**

**A Report**

**on**



**CSE (AI) Department**

**Vishwakarma Institute of Information Technology**

**Academic Year: 2023-24**

**“Assignment 3: Predictive Modelling in a Cosmetics Shop”**

**Submitted by**

|  |  |  |
| --- | --- | --- |
| **Sr No** | **Name** | **Email** |
| **1** | **Omkar Bhaskar Rode** | omkar.22211052@viit.ac.in |

**Under Guidance of**

Dr. Anuradha Yenkikar

Problem Statement: -

**Apply appropriate ML algorithm on a dataset collected in a cosmetics shop showing details of customers to predict customer response for special offers.**

Packages / Libraries used: -

• **NumPy**: Utilized for numerical computations, array manipulation, and handling multi-dimensional arrays. NumPy provides efficient implementations of mathematical functions and operations, making it essential for various machine learning tasks, such as data preprocessing, feature engineering, and model training.

• **Pandas**: Primarily used for data manipulation and analysis, including reading data from various formats like CSV and Excel files, indexing, selecting, sorting, and describing attributes of data. Pandas facilitates tasks such as checking data types, handling missing values, and transforming data structures, making it a fundamental tool for data preprocessing and exploration in machine learning projects.

• **Matplotlib**: Employed for creating static, interactive, and publication-quality visualizations in Python. Matplotlib provides a wide range of plotting functions and customization options, enabling the creation of various types of plots, including line plots, scatter plots, histograms, and bar plots. Matplotlib is highly customizable, allowing users to control every aspect of their plots, from colours and markers to labels and axes.

• **Seaborn**: Utilized for statistical data visualization and enhancing the visual appeal of plots. Seaborn complements Matplotlib with additional high-level functions for creating informative and visually appealing visualizations, such as scatter plots with regression lines, box plots with violin plots, and pair plots with correlation matrices. Seaborn simplifies the process of creating complex statistical plots, making it a valuable tool for exploring relationships between variables and visualizing patterns in the data.

• **scikit-learn**: A comprehensive machine learning library that provides tools for data preprocessing, model selection, model evaluation, and predictive modelling. The **train\_test\_split()** function is used to split the dataset into training and testing sets. The **StandardScaler()** class is employed for feature scaling, ensuring that all features have the same scale. The **cross\_val\_score()** function facilitates cross-validation for estimating the performance of machine learning models. The **SimpleImputer()** class is used for handling missing values. The **accuracy\_score()**, **classification\_report()**, and **confusion\_matrix()** functions are employed for evaluating the performance of classification models. Additionally, classification algorithms such as Logistic Regression, Decision Tree Classifier, Random Forest Classifier, and Support Vector Classifier (SVC) are imported from scikit-learn for model training and prediction.

Theory: -

The primary objective is to develop and evaluate machine learning models for classification tasks using Python and relevant libraries. This project involves leveraging tools such as NumPy, Pandas, Matplotlib, Seaborn, and scikit-learn to preprocess data, build predictive models, and evaluate their performance. Data preprocessing tasks include splitting the dataset into training and testing sets, scaling numerical features, handling missing values, and encoding categorical variables. Classification algorithms such as Logistic Regression, Decision Tree Classifier, Random Forest Classifier, and Support Vector Classifier (SVC) are trained on the training data and evaluated using cross-validation and performance metrics such as accuracy score, classification report, and confusion matrix. Data visualization techniques may also be employed to visualize model performance and interpret the results. By combining data preprocessing, model building, evaluation, and visualization techniques, the project aims to develop robust classification models, derive actionable insights, and facilitate informed decision-making across various domains and industries.

Support Vector Machines: -

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane:



Random Forest Algorithm: -

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.



Methodology: -

* **Data Loading and Preprocessing**:
* The dataset is loaded into a DataFrame, and the features (X) and target variable (y) are separated. The features include 'Price' and 'Rank', while the target variable is labeled 'Label'.
* Missing values in the features are imputed using the mean strategy, ensuring that no data is lost during preprocessing.
* **Data Splitting**:
* The dataset is split into training and testing sets using the **train\_test\_split()** function from scikit-learn. This is crucial for evaluating the performance of the machine learning model on unseen data.
* **Feature Scaling**:
* Standardization is applied to the numerical features using the **StandardScaler()** from scikit-learn. This ensures that all features have the same scale, which is important for certain machine learning algorithms.
* **Model Training**:
* A logistic regression model is instantiated and trained on the training data using the **LogisticRegression()** function from scikit-learn. Logistic regression is a commonly used classification algorithm that predicts the probability of a binary outcome.
* **Model Evaluation**:
* The trained model is used to make predictions on the testing data, and the accuracy of the predictions is computed using the **accuracy\_score()** function from scikit-learn.
* The confusion matrix and classification report are generated using the **confusion\_matrix()** and **classification\_report()** functions, respectively. These metrics provide insights into the performance of the model in terms of classification accuracy, precision, recall, and F1-score.
* **Visualization**:
* A heatmap of the confusion matrix is created using Seaborn's **heatmap()** function. This visualization helps in understanding the distribution of correct and incorrect predictions made by the model.
* **Documentation and Reporting**:
* Finally, the findings from model evaluation are documented and reported. This includes summarizing key insights, describing the methodology used, and presenting the results visually using plots and charts. Clear and concise communication of findings is essential for stakeholders to understand and make informed decisions based on the analysis.

Application: -

* **E-commerce Pricing Strategy**: The features 'Price' and 'Rank' could represent product pricing and ranking data, respectively. By training a logistic regression model on this data, e-commerce companies can predict the likelihood of a product being purchased based on its price and ranking. This information can inform pricing strategies and promotional efforts to maximize sales and revenue.
* **Customer Churn Prediction**: Similar classification techniques can be applied in industries like telecommunications or subscription-based services to predict customer churn. By analyzing customer behavior and characteristics (such as usage patterns and account status), companies can identify customers at risk of churn and take proactive measures to retain them.
* **Credit Risk Assessment**: In the financial sector, logistic regression models can be used to assess credit risk by predicting the likelihood of loan default based on borrower characteristics and credit history. Lenders can use these predictions to make informed decisions about loan approvals and set appropriate interest rates.
* **Medical Diagnosis**: Logistic regression models can assist in medical diagnosis tasks by predicting the likelihood of a patient having a certain condition based on their symptoms and medical history. These models can aid healthcare professionals in making accurate diagnoses and developing personalized treatment plans.
* **Employee Attrition Prediction**: Human resources departments can use logistic regression models to predict employee attrition based on factors such as job satisfaction, performance ratings, and tenure. By identifying employees at risk of leaving, organizations can implement retention strategies to improve employee satisfaction and reduce turnover.

Diagrams: -



Conclusion: -

In conclusion, the implemented methodology showcases the application of logistic regression and classification techniques in predictive modeling tasks. By leveraging machine learning algorithms and appropriate preprocessing techniques, valuable insights can be derived from data, leading to informed decision-making across various domains. The evaluation metrics such as accuracy score, confusion matrix, and classification report provide a comprehensive understanding of the model's performance, enabling stakeholders to assess its effectiveness in real-world scenarios. Furthermore, the visualization of results enhances interpretability and facilitates communication of findings to stakeholders. Overall, the successful implementation of machine learning techniques underscores their significance in solving complex problems and driving actionable insights from data.