Assignment 01:
Machine Learning & NN
(ICT 06135231)

Presented By:

Md. Shahidul Alam

ID: 2023822017

MIT-6th Batch

Tasks on KNN, Naive Bayes and SVM

TASKS:

Each Notebook implement:

- K-Nearest Neighbors (KNN) IMPLEMENTATION
- Naive Bayes (NB) IMPLEMENTATION
- Support Vector Machine (SVM) IMPLEMENTATION

IMPLEMENTATION STEPS:

- Preprocess the dataset (handle missing values, encoding categorical variables, normalization if needed).
- Train each model separately on classification and regression tasks.
- Evaluate the models using appropriate metrics:

1) Classification: Accuracy, Precision, Recall, F1-score, Confusion Matrix.

2) Regression: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R^2 score.

- Visualize Results (decision boundaries for classification, confusion matrices, error distributions for regression).

Three reports (summarizing) for KNN, Naive Bayes, SVM:

1. About Dataset
2. Preprocessing steps
3. Model performance results with explanations
4. Observations on performance changes when modifying parameters

[Three reports repository]

About Dataset

The dataset consists of 5,000 entries with 14 columns, representing customer data from a bank. The key attributes include:

- Demographics: Age, Experience, Family size, Education level.
- Financial Information: Income, Mortgage, Average Credit Card Spending (CCAvg).
- Banking Details: Whether the customer has a Securities Account, Certificate of Deposit (CD) Account, Online Banking access, and a Credit Card.
- Loan Status: A binary column (Personal Loan) indicating whether a customer has taken a personal loan.

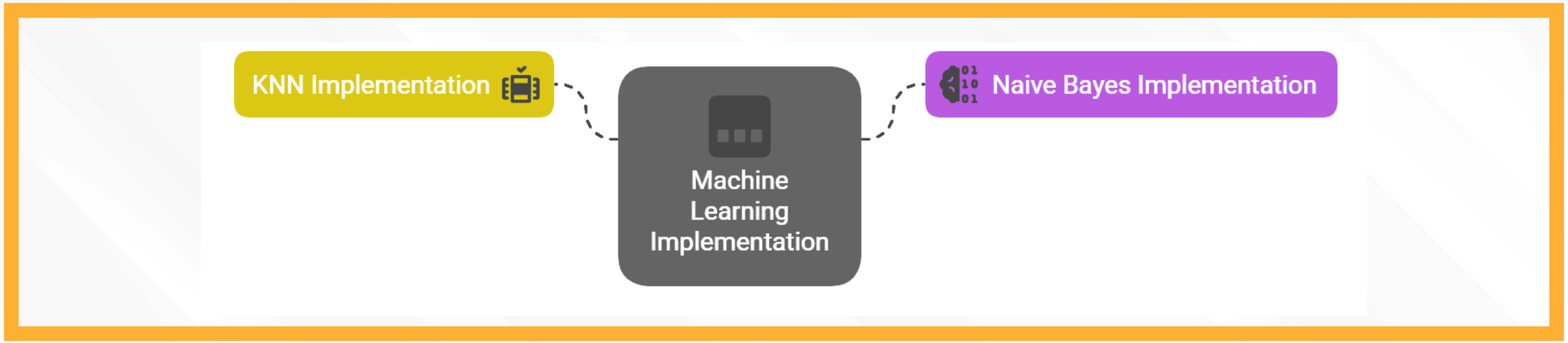
This dataset is useful for analyzing factors that influence a customer's likelihood of taking a loan, making it ideal for predictive modeling and customer segmentation.

Datasets: Bank_Personal_Loan_Modelling.csv

Summary of the Project:

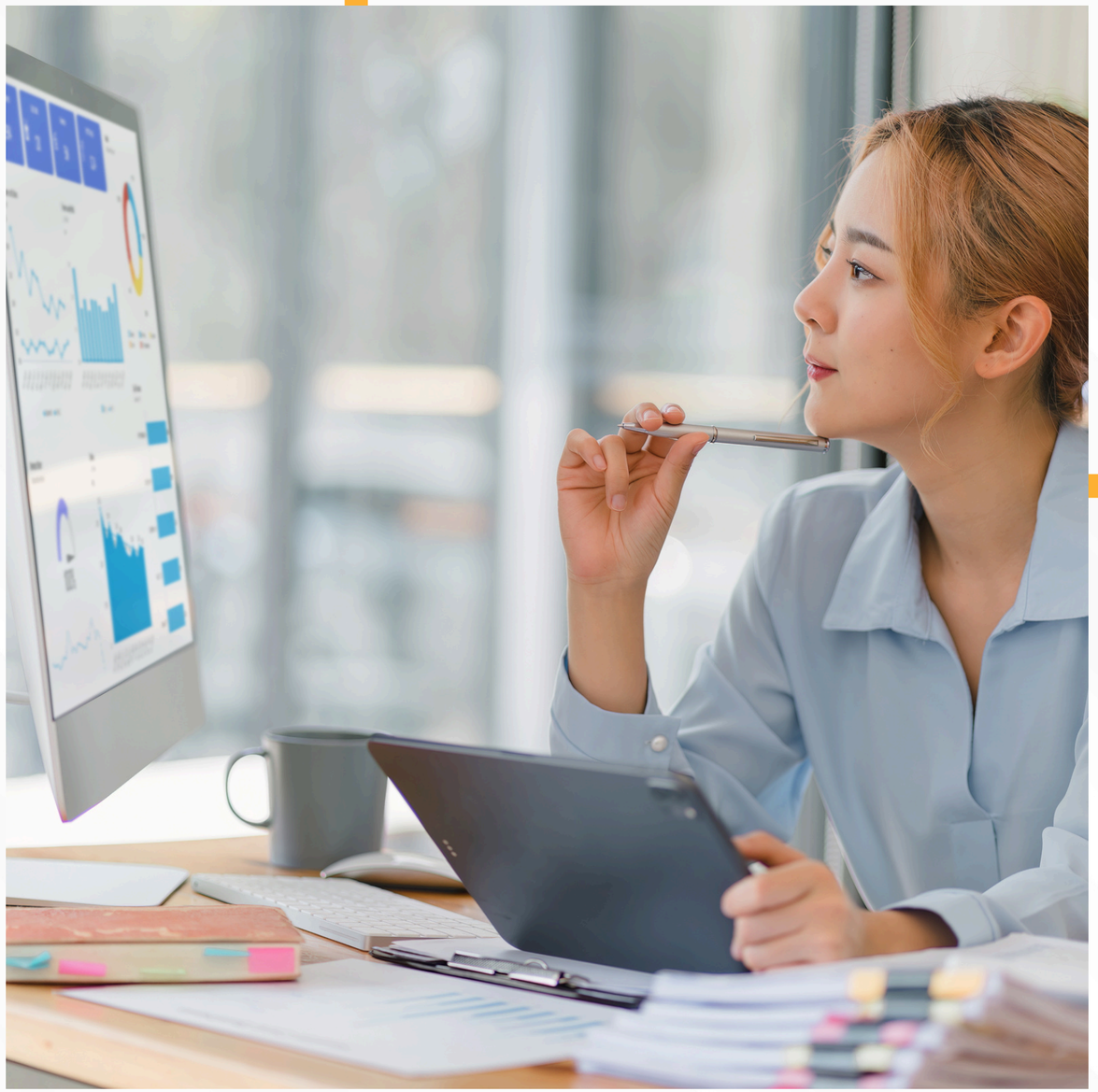
This project focuses on implementing and evaluating two fundamental machine learning algorithms, K-Nearest Neighbors (KNN) and Naive Bayes. The dataset used is the Bank Personal Loan Modelling dataset, which contains information about customers and whether they accepted a personal loan offer. The goal is to classify whether a customer will accept a loan and evaluate the performance of the models using metrics such as accuracy, precision, recall, F1-score and confusion matrices. Additionally, the project explores regression tasks with MSE, RMSE and R^2 for continuous target values.

Project repository



Data Preprocessing

- The dataset was preprocessed to ensure it was suitable for machine learning. Missing values were handled by filling them with the mean of the respective columns. Categorical variables (If any) were encoded using label encoding, and the data was normalized using StandardScaler to ensure all features were on the same scale. The dataset was then split into training and testing sets (80% training, 20% testing) to evaluate the models.



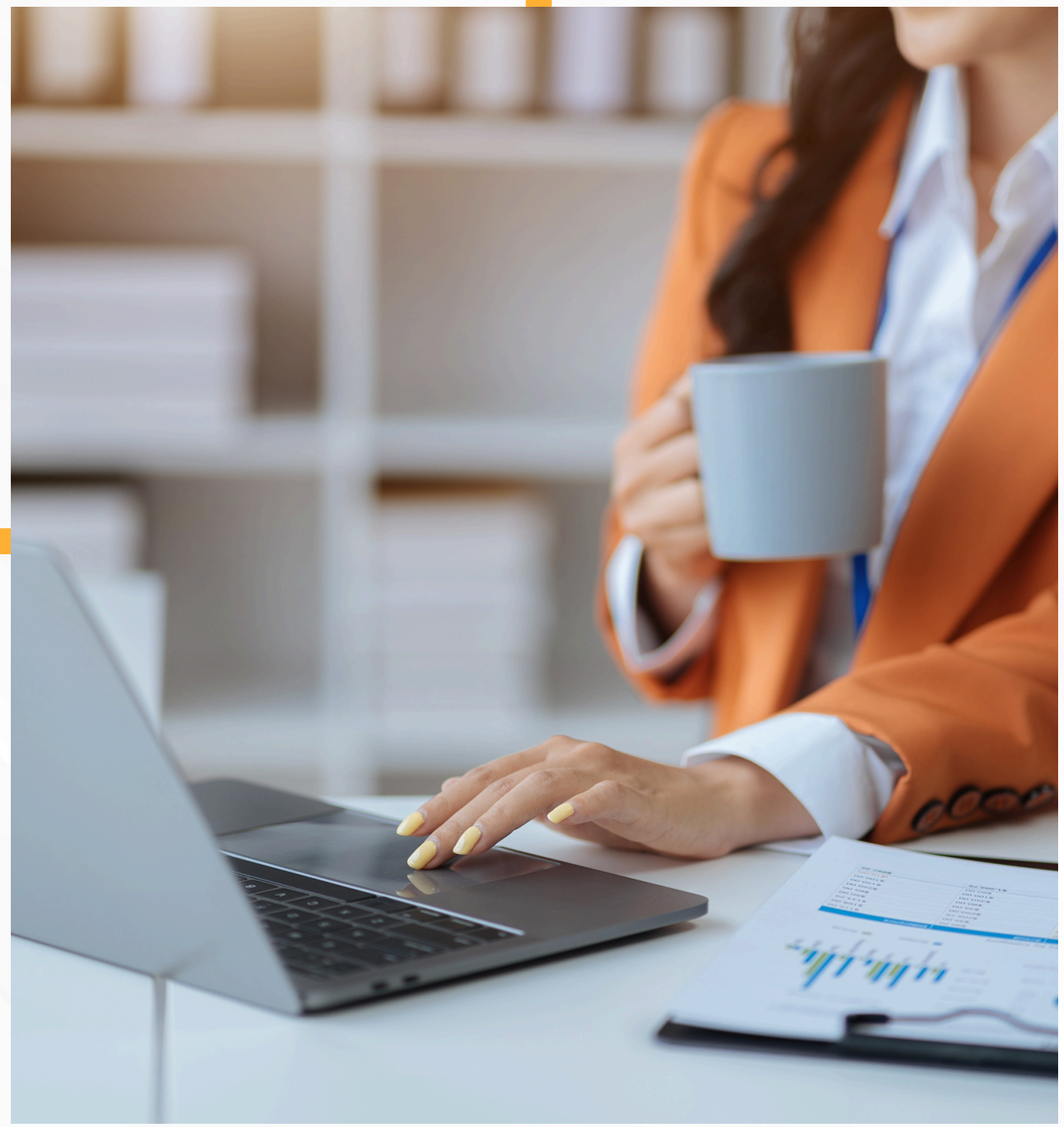
KNN Implementation

- After drop 2 attributes, the dataset consists of 5000 records with 12 customer-related attributes, used for predicting personal loan approval (classification) and income levels (regression). Preprocessing included handling missing values, dropping irrelevant columns, encoding categorical variables, and normalizing numerical features. The KNN classifier achieved high accuracy, with a well-balanced confusion matrix and strong precision, recall, and F1-score. In regression, the model recorded an MSE of 1090.83, RMSE of 33.03, and an R^2 score of 0.4856, indicating moderate variance explanation. Performance is influenced by the number of neighbors, distance metric, and feature scaling. Proper hyperparameter tuning and dataset balancing can further enhance results while preventing overfitting or underfitting.

Project repository for KNN

Naive Bayes Implementation

- The Naïve Bayes classifier (GaussianNB) achieved high accuracy, with balanced precision, recall, and F1-score, and a well-distributed confusion matrix. The regression model showed outstanding performance, with an R^2 score of 1.0, indicating perfect variance explanation. Performance tuning through var_smoothing for classification and alpha_1, alpha_2 for regression can further optimize results while balancing overfitting and underfitting risks.

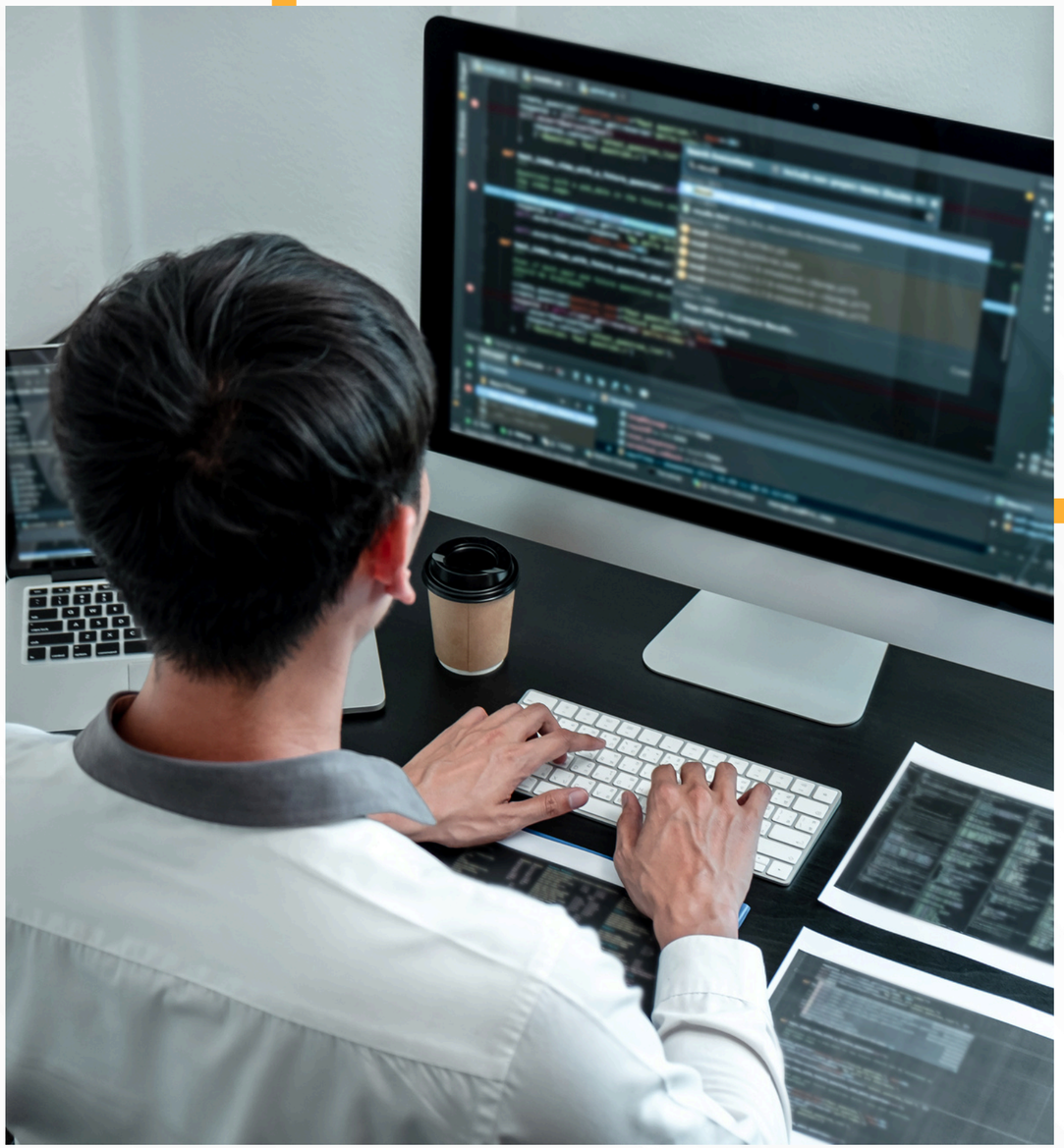


[Project repository for Naive Bayes]

Support Vector Machine (SVM)

- The dataset contains customer attributes from a banking institution, used to predict personal loan approval and income levels. Preprocessing involved handling missing values, cleaning column names, dropping unnecessary columns, encoding categorical features, and applying normalization. An SVM with a linear kernel achieved high accuracy in classification, with balanced precision, recall, and F1-score. The regression model (SVR) demonstrated excellent performance with an MSE of 0.0033, RMSE of 0.057, and an R^2 score of 0.9999. Performance can be optimized by tuning kernel type, regularization parameter (C), and gamma, ensuring proper feature scaling for improved results.

[Project repository for SVM]



Comparison and Results

- The comparison of KNN, Naïve Bayes, and SVM models highlights their strengths and weaknesses in classification and regression tasks. In classification, SVM achieved the highest accuracy with strong precision, recall, and F1-score, making it the most effective in predicting personal loan approval. Naïve Bayes also performed well, with balanced metrics and efficient computations, but it assumes feature independence, which may impact complex data. KNN showed good accuracy but is sensitive to parameter tuning and computationally expensive for large datasets. In regression, SVM had the best performance with an R^2 score of 0.9999, while KNN struggled with an R^2 of 0.4856. Overall, SVM proved the most robust, while Naïve Bayes offered efficiency and KNN required careful optimization.

Thank You

That is summary of the project