Dataset Link:

The dataset used in the project can be found on Kaggle:

https://www.kaggle.com/datasets/hendraherviawan/germancreditdata

Description of Project:

The code you provided is a Jupyter notebook that predicts credit default risk for a dataset of German credit applicants. The goal of the project is to develop a machine learning model that can accurately predict whether a credit applicant is likely to default on their loan.

The notebook uses a dataset of credit applicants that includes information such as age, sex, income, credit history, and loan amount. The dataset is preprocessed and cleaned before being used to train and test several machine learning models, including logistic regression, decision trees, and random forests.

The notebook walks through several steps of the machine learning process, including data exploration, data preprocessing, feature selection, model training and evaluation, and hyperparameter tuning. The goal is to develop a model that can accurately predict credit default risk for new credit applicants.

A brief explanation of the outputs of the Credit risk modeling Project

1. Data exploration and visualization: The code performs data exploration and visualization to gain insights into the characteristics of the German credit dataset, including the distribution of credit risk, the distribution of credit amounts, the distribution of credit durations, and the correlation between different features.

2. Data preprocessing: The code preprocesses the data by converting categorical variables into numerical variables, scaling the features, and splitting the data into training and testing sets.

3. Model training and evaluation: The code trains different machine learning models, including Logistic Regression, K-Nearest Neighbors, Decision Tree, Random Forest, and Gradient Boosting, and evaluates their performance using different metrics such as accuracy, precision, recall, F1-score, and AUC-ROC curve.

4. Model selection and tuning: The code selects the best-performing model and performs hyperparameter tuning to improve its performance.

5. Model deployment: The code deploys the selected model on new data to make predictions about credit default risk.

6. Conclusion and future work: The code concludes with a summary of the findings and suggests future work to improve the performance of the model.

Description of Output:

The output of the notebook is a set of insights and recommendations for predicting credit default risk in German credit applicants. The notebook provides visualizations and insights into the data, as well as code for training and evaluating several machine learning models.

In addition to the models themselves, the notebook also provides evaluation metrics for each model, such as accuracy, precision, and recall, as well as visualizations of the model performance over time. The notebook also provides feature importance rankings, which can help identify which factors are most important in predicting credit default risk.

Based on the analysis, the notebook provides several recommendations for improving the accuracy of credit default risk predictions, such as using different machine learning algorithms, improving feature selection methods, and gathering more data to better train the model.

Instructions on How to Run the Code/Project/File:

1. To run the code in the notebook, you will need to have Jupyter Notebook installed on your computer. Once you have installed the Jupyter Notebook, you can download the notebook from the Kaggle website and open it in the Jupyter Notebook.
2. Before running the code, you will need to make sure that you have downloaded the necessary data files and saved them in the correct directory. The notebook provides instructions on how to download the data files and where to save them.
3. Once you have downloaded the data files and opened the notebook in Jupyter Notebook, you can run each cell of the notebook by clicking on the cell and then clicking the "Run" button in the toolbar or by using the keyboard shortcut "Shift + Enter".
4. It is recommended that you run the code cells in order, as some cells depend on the output of earlier cells.

Note: Make sure to update the file paths in the code cells to match the location of the downloaded dataset and kernel files on your local machine.