

Data mining on Microsoft azure cloud

Prepared by:
Boukayoua Loubna
El boutaheri Najma

Framed by:
Professor Hayat Routaib

OVERVIEW

- 01** INTRODUCTION
- 02** PROBLEM STATEMENT
- 03** PROJECT OVERVIEW
- 04** PROJECT ARCHITECTURE
- 05** PROJECT COMPONENT
- 06** DATA MINING EXPLORATORY
- 07** MACHINE LEARNING MODEL BUILDING
- 08** MODEL DEPLOYMENT
- 09** CONCLUSION

Introduction

Outcomes of Data Mining

Enhanced
Decision-Making

Trend Prediction

Pattern
Recognition

Business
Problem Solving



Problem statement

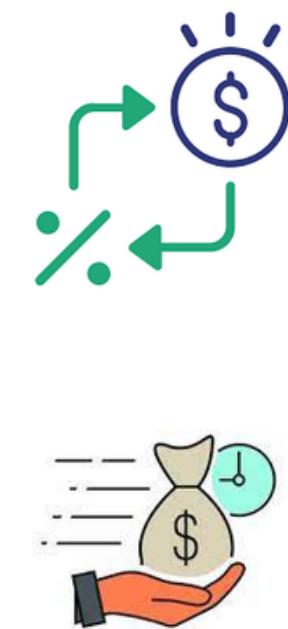
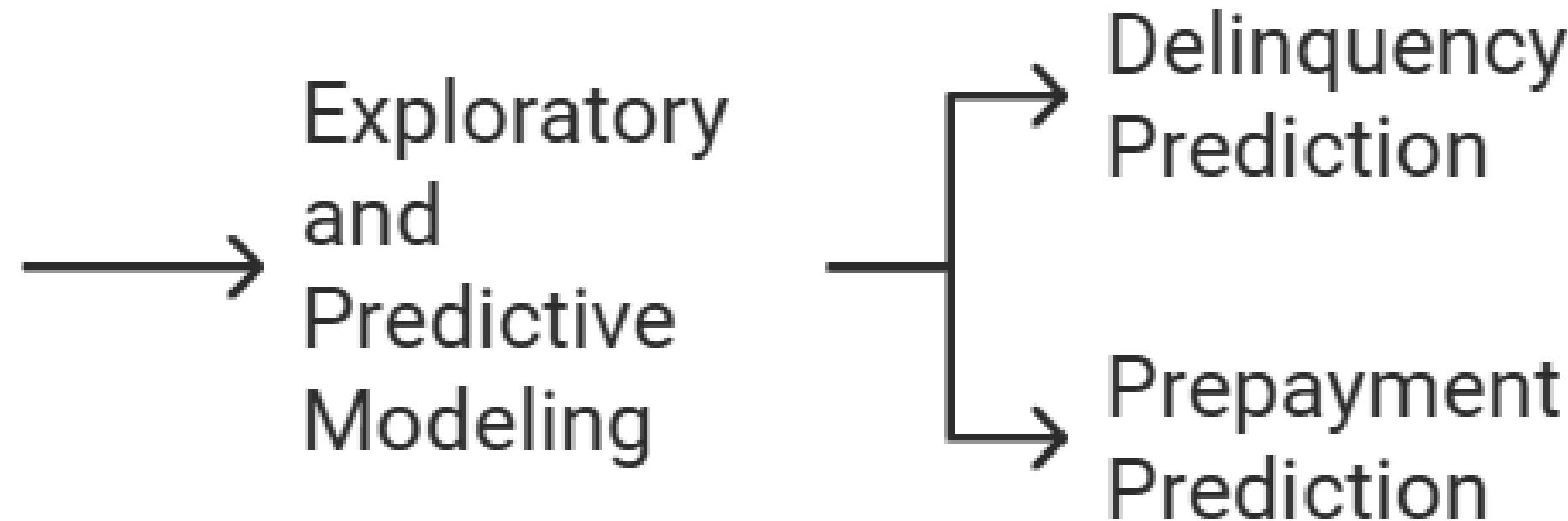
The financial sector faces challenges in managing loan portfolios, particularly in predicting borrower behaviors like delinquency and prepayment.

- **Delinquency:** Disrupts lenders' cash flow.
- **Prepayment:** Impacts revenue forecasts.

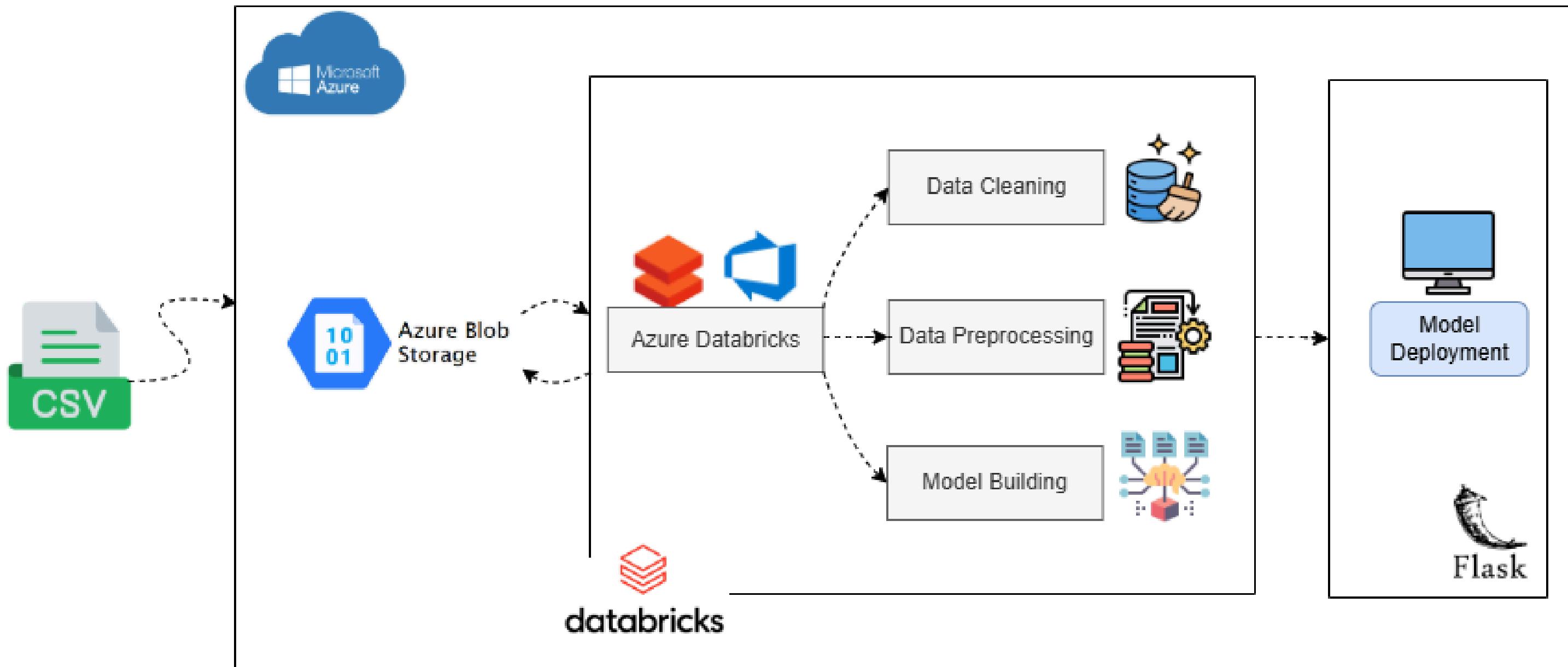


ROLE OF DATA MINING IN FINACIAL SECTOR:

Data Analysis with Azure Cloud



ARCHITECTURE



PROJECT OVERVIEW



This project leverages data mining techniques to build predictive models for delinquency and prepayment behaviors using historical loan data. Key steps include data preprocessing, exploratory data analysis, feature engineering, and deploying machine learning models like logistic regression and gradient boosting.

PROJECT OVERVIEW

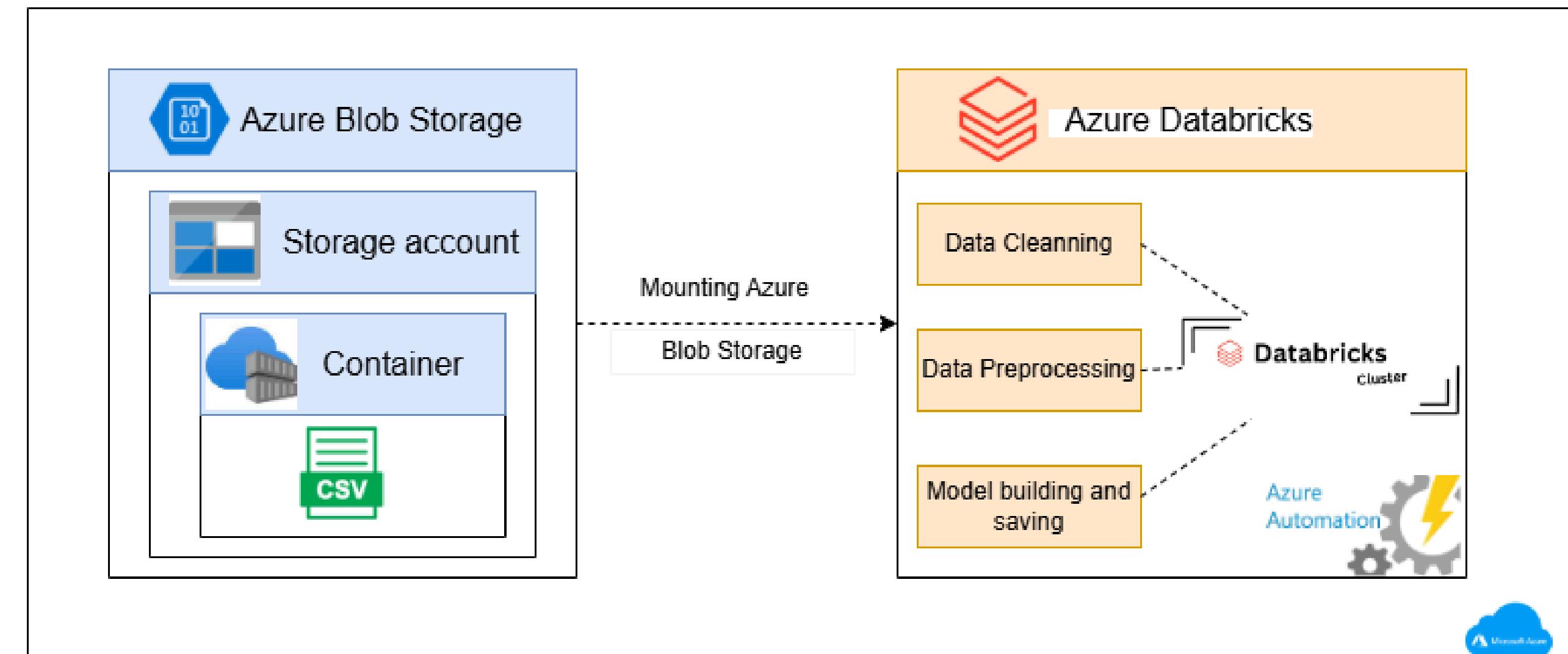


The architecture integrates Microsoft Azure services for scalability and reliability, utilizing Azure Blob Storage for data management, Azure Databricks for processing, and building Machine Learning pipeline for predictions. A Flask-based app enables real-time predictions.

PROJECT COMPONENTS: BLOB STORAGE & DATABRICKS

Why Blob Storage:

- Scalability
- Cost-effective
- Flexibility
- Integration



PROJECT COMPONENTS: DATABRICKS JOBS

The screenshot shows the Databricks Jobs interface. At the top, there's a navigation bar with 'Runs' and 'Tasks'. Below it, a dependency graph shows 'run1' (Notebook, Cluster) pointing to 'run2' (Notebook, Cluster). The 'Tasks' tab is selected, displaying the configuration for 'run2':

- Task name***: run2
- Type***: Notebook
- Source***: Workspace
- Path***: ...a.boukayoua@etu.uae.ac.ma/ExploratoryDataAnalysisForLoanExportData(2)
- Cluster***: Cluster (DBR 15.4 LTS · Photon · Spark 3.5.0 · Scala 2.12)
- Depends on**: run1
- Run if dependencies**: All succeeded

On the right, the 'Job details' panel shows:

- Job ID**: 632965477140308
- Creator**: LOUBNA BOUKAYOUA
- Run as**: LOUBNA BOUKAYOUA
- Tags**: Add tag
- Description**: Add description
- Git**: Not configured. Add Git settings.
- Schedule**: None. Add trigger.
- Compute**: Cluster

PROJECT COMPONENTS: DATABRICKS JOBS



PROJECT COMPONENTS

Azure Blob Storage is used to store raw data files for the data mining workflow.

DATA STORAGE

Azure Databricks facilitates:

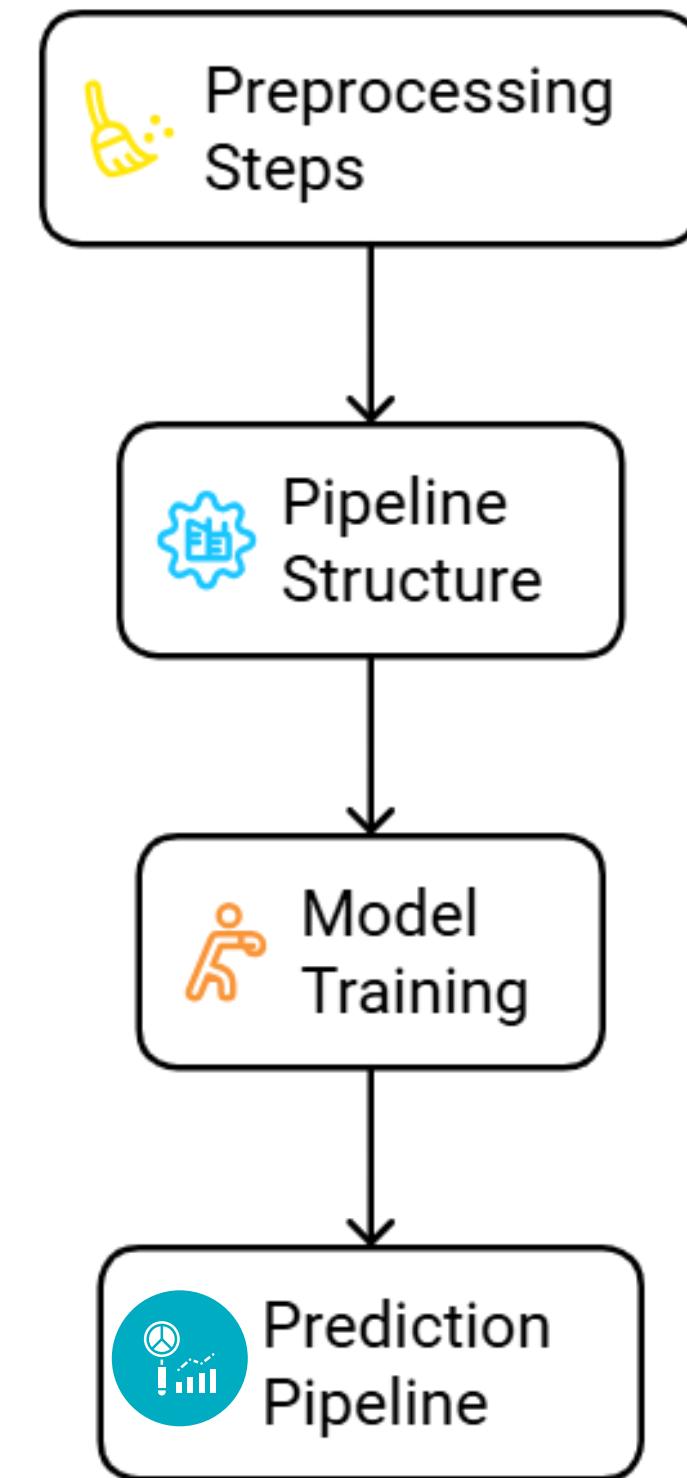
- Data cleaning.
- Preprocessing.
- Model Building: Training and validating machine learning models.

DATA PROCESSING AND MODEL DEVELOPMENT

Deploying the trained model to a live environment using a virtual machine provided by Microsoft Azure. Flask will be used for hosting the model as an API.

MODEL DEPLOYMENT

MACHINE LEARNING MODEL BUILDING STEPS



ABOUT THE DATASET

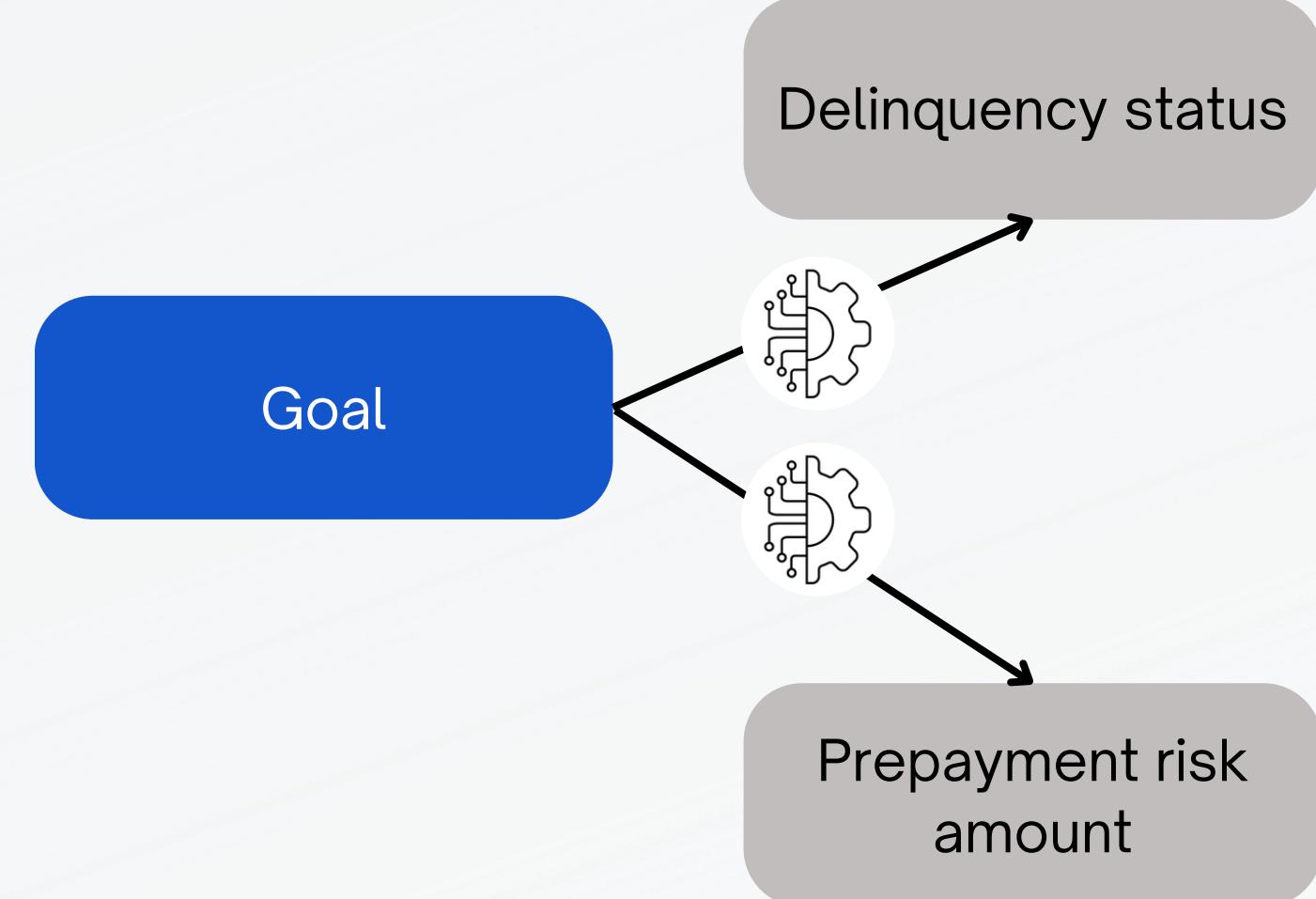
DATA
COLLECTION

COLUMN DETAILS

PROPERTY
INFORMATION

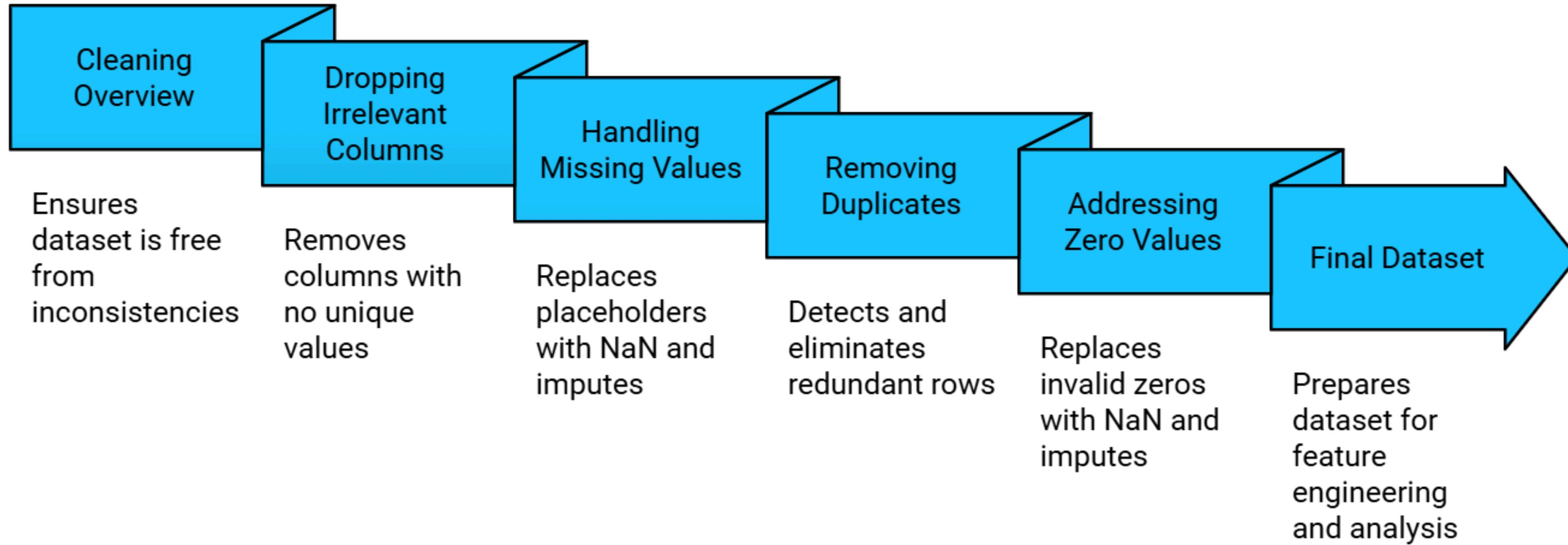
LOAN
CHARACTERISTICS

BORROWER BEHAVIOR

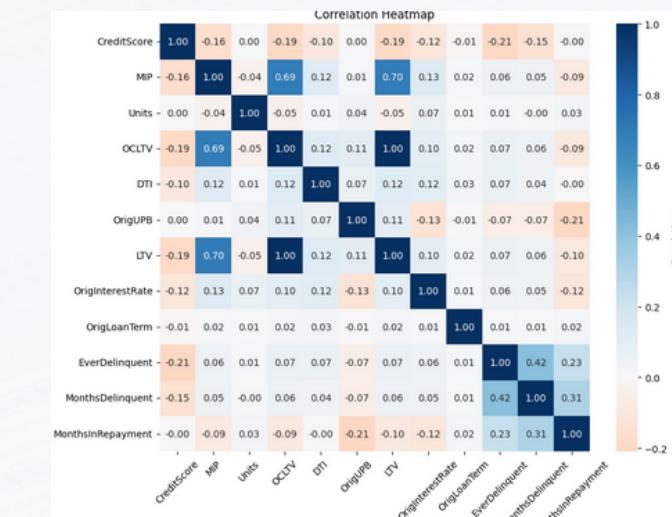
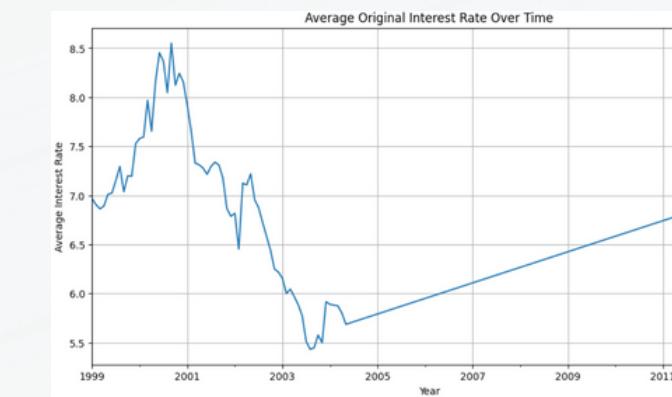
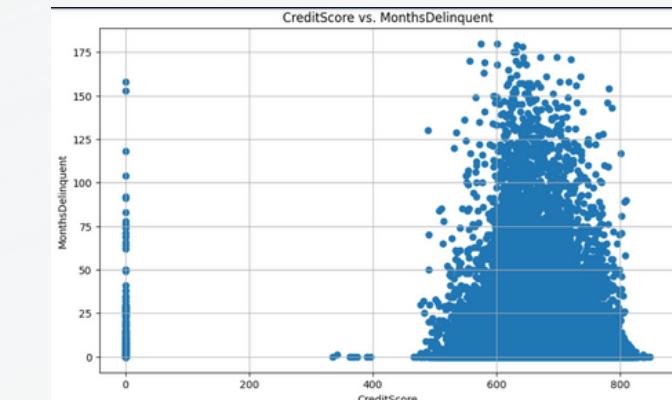
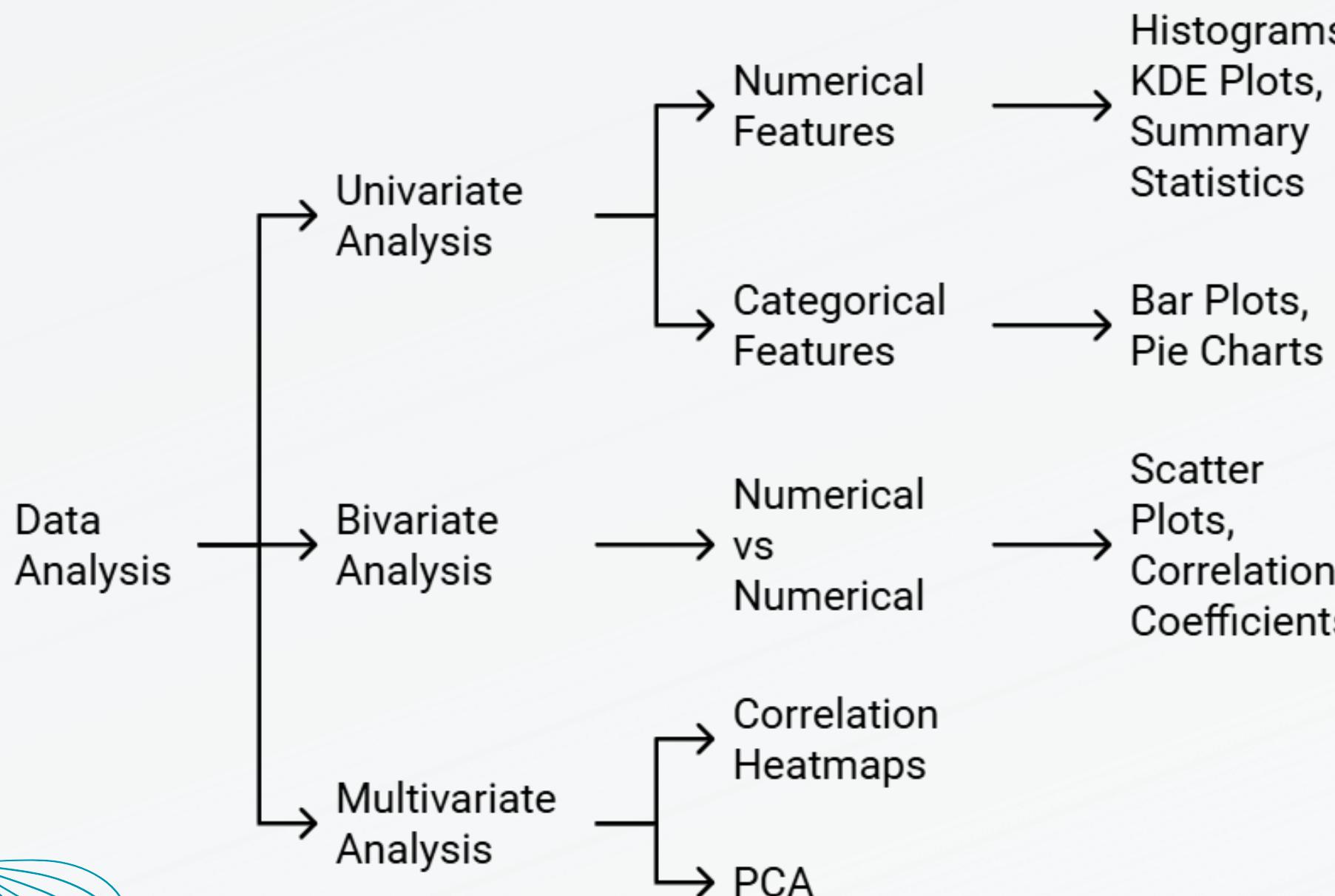


DATA PREPROCESSING

Data Cleaning Process for Machine Learning



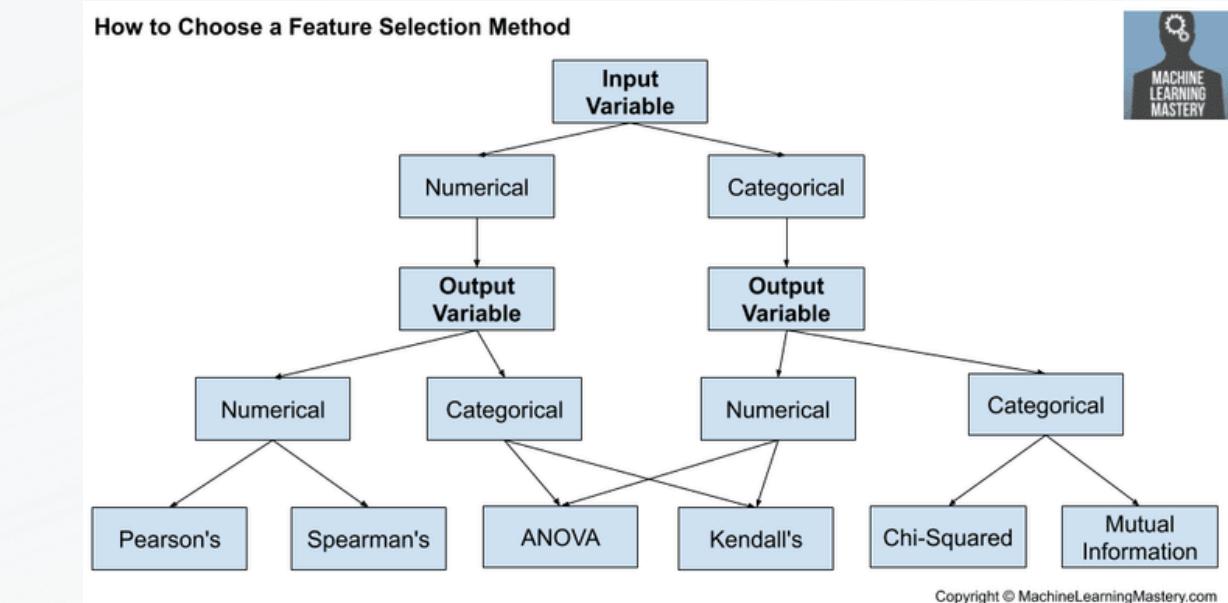
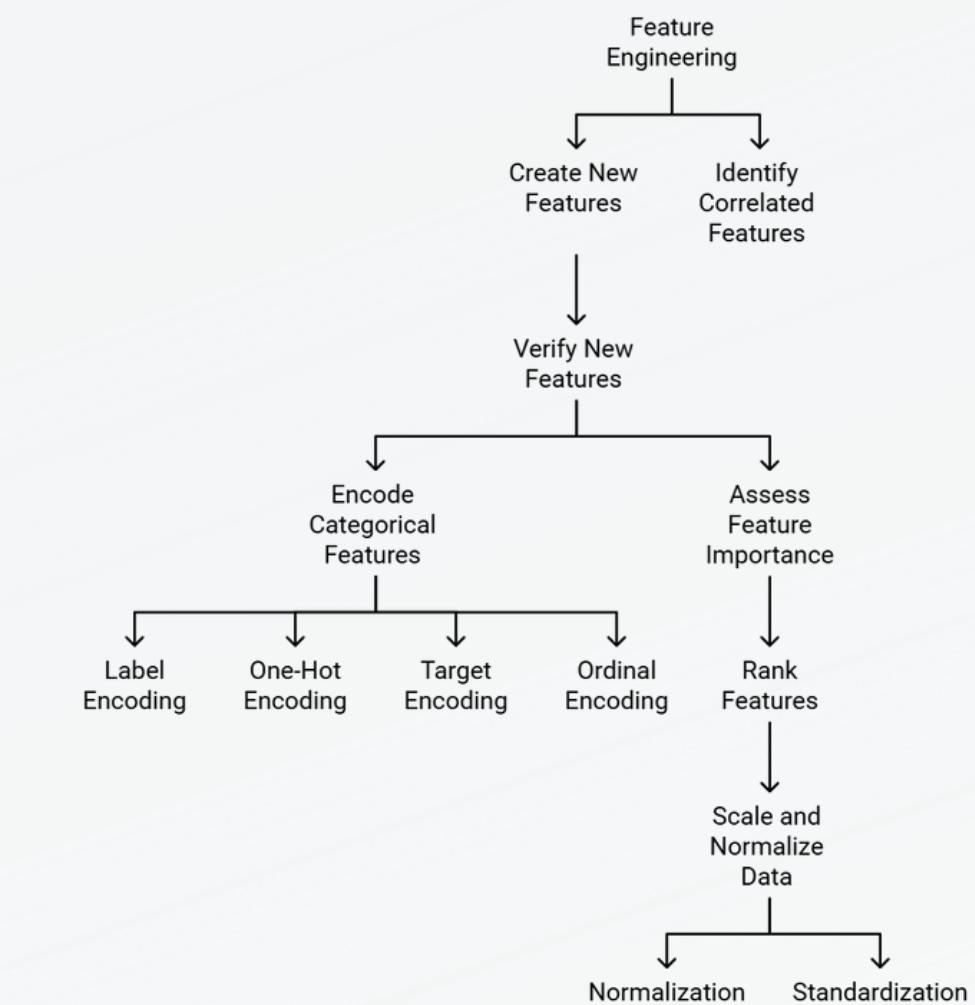
EXPLORATORY DATA ANALYSIS (EDA)



PREPARING DATA FOR MODELING

Thechniques used:

- Binning/Discretization
- Feature Importance
- Univariate Statistical Tests
- Dimensionality Reduction



DATA MODELING



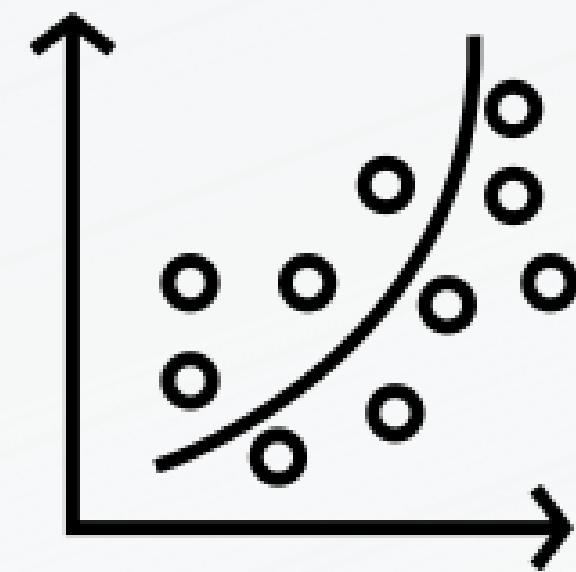
Classification Models:

- Logistic Regression
- Naive Bayes
- Decision Tree



Regression Models:

- Linear Regression
- Ridge Regression



MODEL EVALUATION OVERVIEW

Algorithm	CV_Roc_Auc_Score	Training_accuracy
Logistic Regression	1.000000	1.000000
Decision Tree	1.000000	1.000000
GaussianNB	1.000000	0.988566
BernoulliNB	0.845195	0.904684

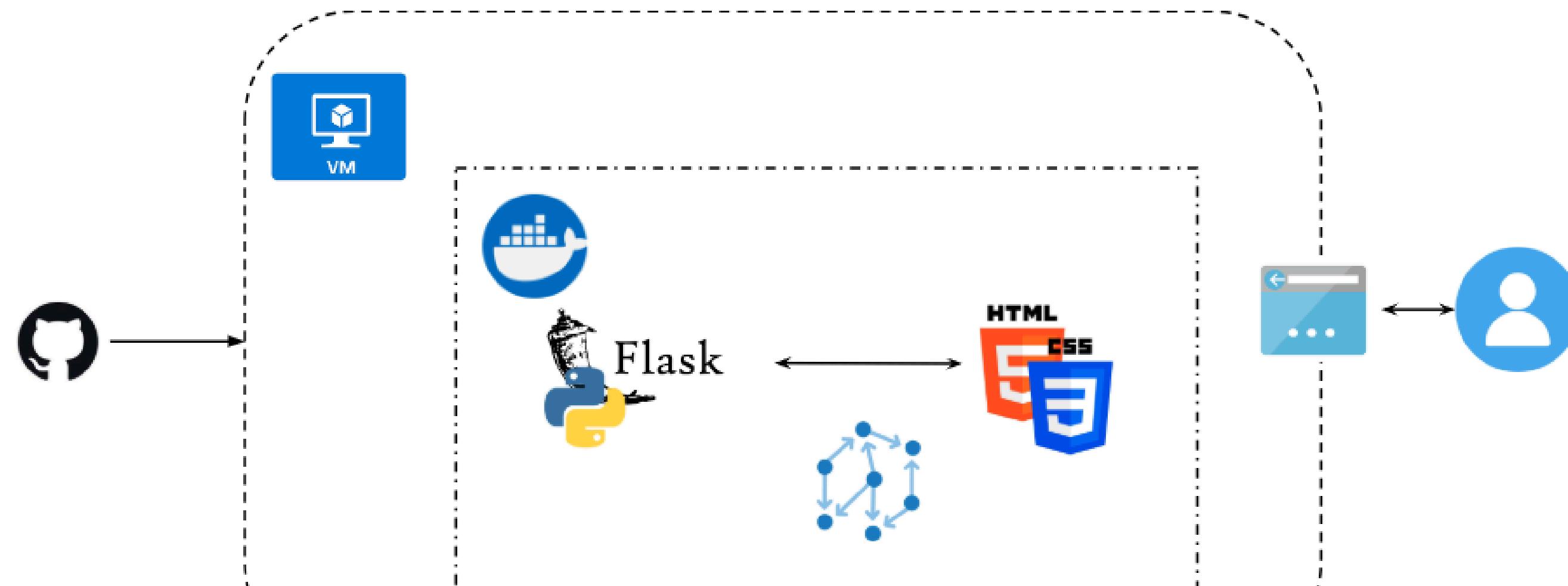
Algorithm	Testing_accuracy	Roc_Auc_score	f1_score
Logistic Regression	1.000000	1.000000	1.000000
Decision Tree	1.000000	1.000000	1.000000
GaussianNB	0.988425	1.000000	0.971297
BernoulliNB	0.904568	0.848416	0.717914

After evaluating the models on accuracy and F1-score, we found that GaussianNB outperformed the other models due to its ability to handle imbalanced datasets and its robustness to overfitting.

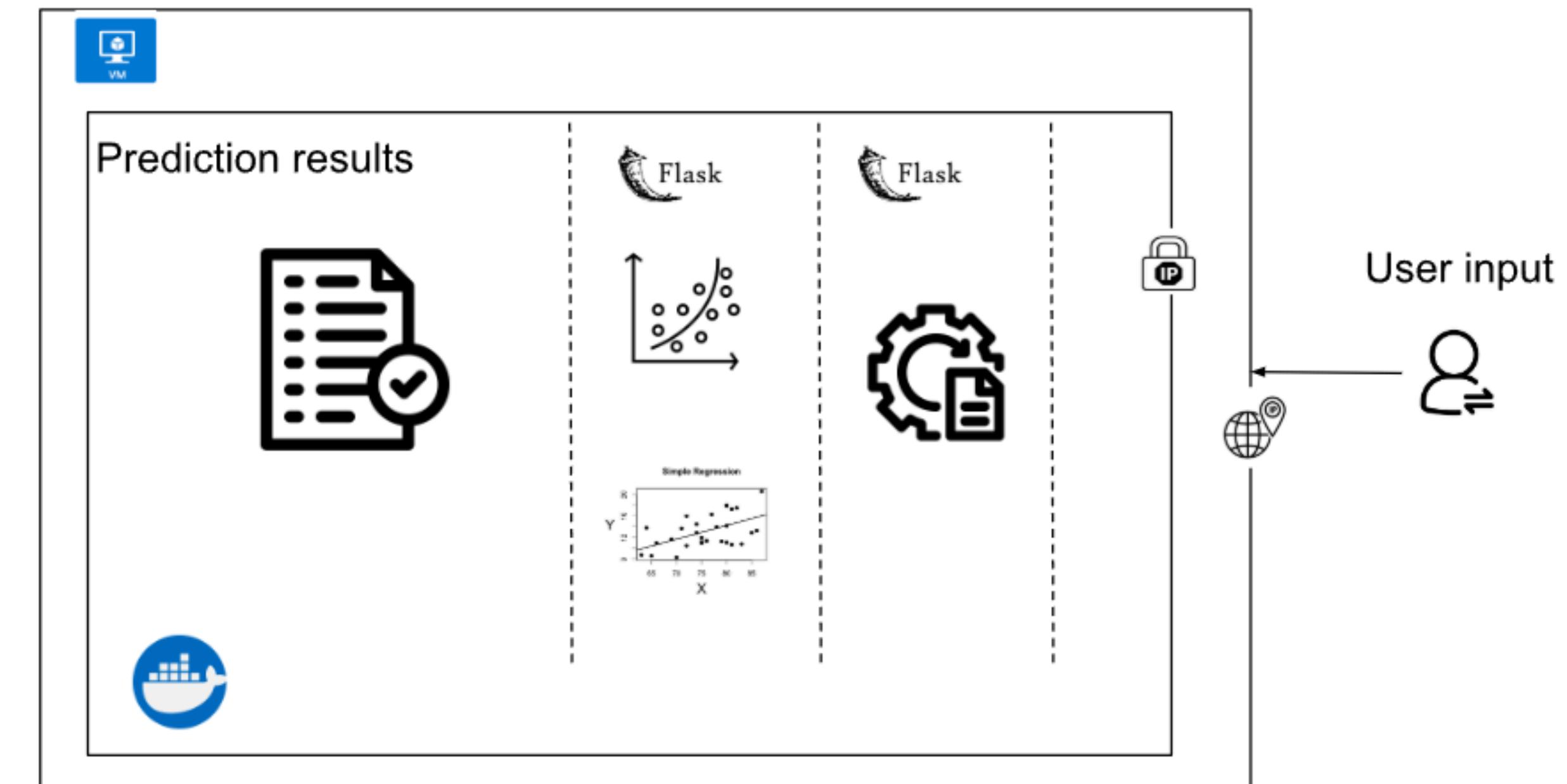
Data problem fixed:

- Fixing Data Leakage
- Preventing Overfitting
- Handling Outliers

Deployment architecture



Classification and Regression Pipeline Workflow



Technical demonstration

The screenshot shows the Microsoft Azure portal interface. The top navigation bar includes tabs for 'loandatastorage - Microsoft Azure', 'Compute - Databricks', and 'Mortgage Prediction Form'. The main address bar displays the URL: portal.azure.com/#@uae.ac.ma/resource/subscriptions/7f158fe1-b15d-4507-a993-04d7905e5587/resourceGroups/data-mining-project/pro.... The user's email, 'loubna.boukayoua@etu...', and name, 'UNIVERSITÉ ABDELMALEK ESSA', are visible in the top right corner.

The left sidebar shows a navigation tree with the following items:

- Home >
- loandatastorage Storage account
- Overview
- Activity log
- Tags
- Diagnose and solve problems
- Access Control (IAM)
- Data migration
- Events
- Storage browser
- Storage Mover
- Partner solutions
- > Data storage
- > Security + networking
- > Data management

The main content area displays the 'loandatastorage' storage account details under the 'Properties' tab. The 'Essentials' section includes:

- Resource group ([move](#)) [data-mining-project](#)
- Location [eastus](#)
- Subscription ([move](#)) [Azure for Students](#)
- Subscription ID [7f158fe1-b15d-4507-a993-04d7905e5587](#)
- Disk state [Available](#)
- Tags ([edit](#)) [Add tags](#)

On the right side, detailed account information is listed:

- Performance [Standard](#)
- Replication [Locally-redundant storage \(LRS\)](#)
- Account kind [StorageV2 \(general purpose v2\)](#)
- Provisioning state [Succeeded](#)
- Created [28/11/2024 18:04:42](#)

Below the main content, there are tabs for 'Properties' (selected), 'Monitoring', 'Capabilities (7)', 'Recommendations (3)', 'Tutorials', and 'Tools + SDKs'. At the bottom, there are sections for 'Blob service' and 'Security'.

Conclusion

References

1. **Scikit-learn Documentation.** (n.d.). Retrieved from:
<https://scikit-learn.org>.
1. **Microsoft Azure Documentation.** (n.d.). Retrieved from
<https://azure.microsoft.com>
2. **Postman API Testing Tool.** (n.d.). Retrieved from
<https://www.postman.com>
3. Géron, A. (2019). **Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow.** O'Reilly Media.
4. <https://www.techtarget.com/searchbusinessanalytics/definition/data-mining>



**Thanks for
your attention**