ABDELMALEK ESSAADI UNIVERSITY
NATIONAL SCHOOL OF APPLIED SCIENCES AL HOCEIMA

Data mining on Microsoft azure cloud

Prepared by:

Boukayoua Loubna

El boutaheri Najma

Framed by:

Professor Hayat Routaib

OVERVIEW

Introduction

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

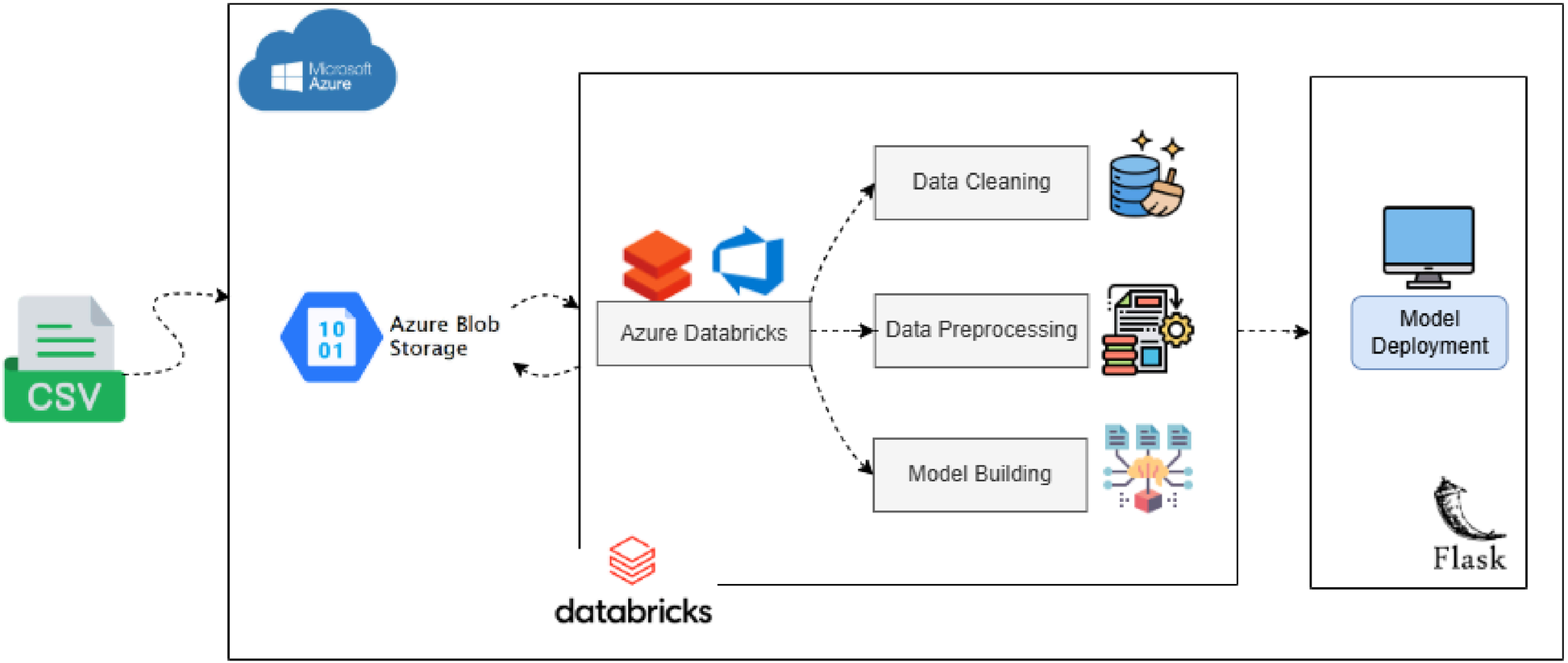
→

Exploratory and Predictive Modeling

→ Delinquency Prediction

→ Prepayment Prediction

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.

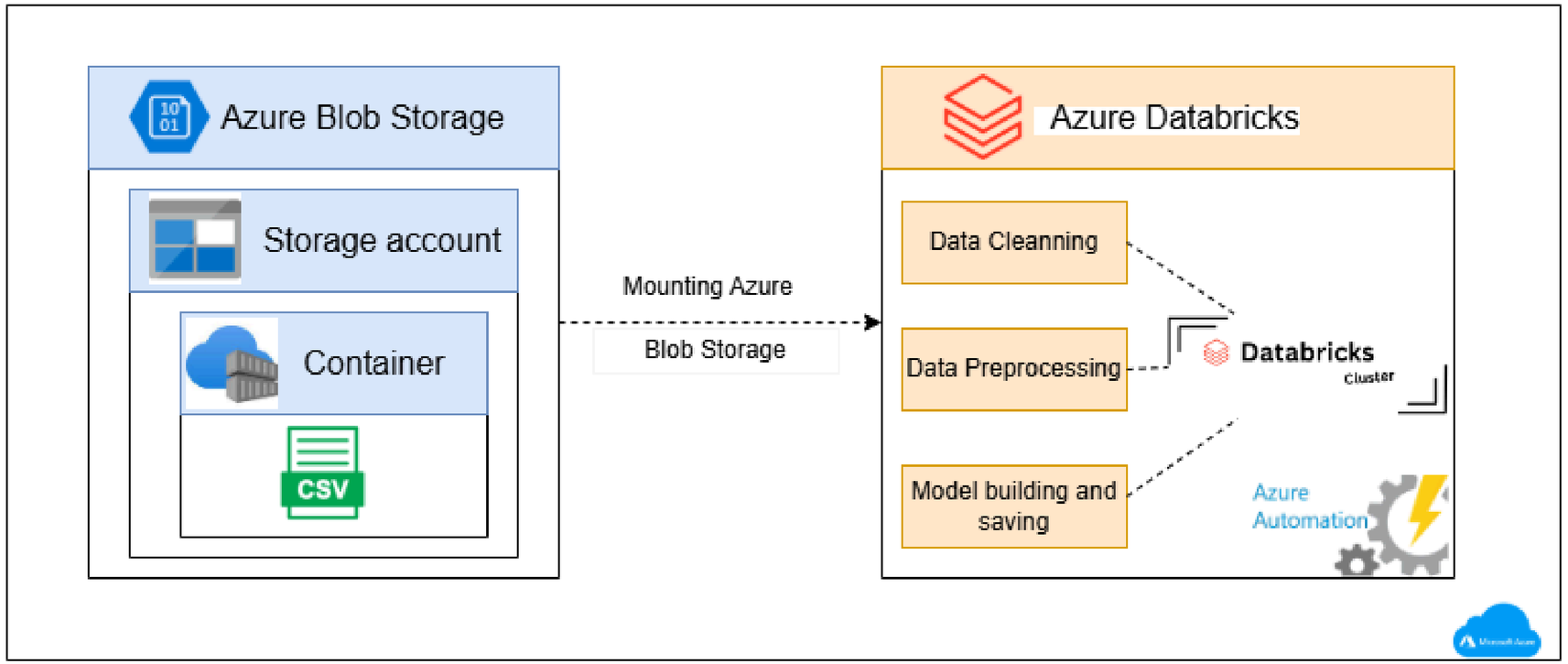
PROJECT OVERVIEW

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 pipline for predictions
A Flask-based app enables real-time predictions.

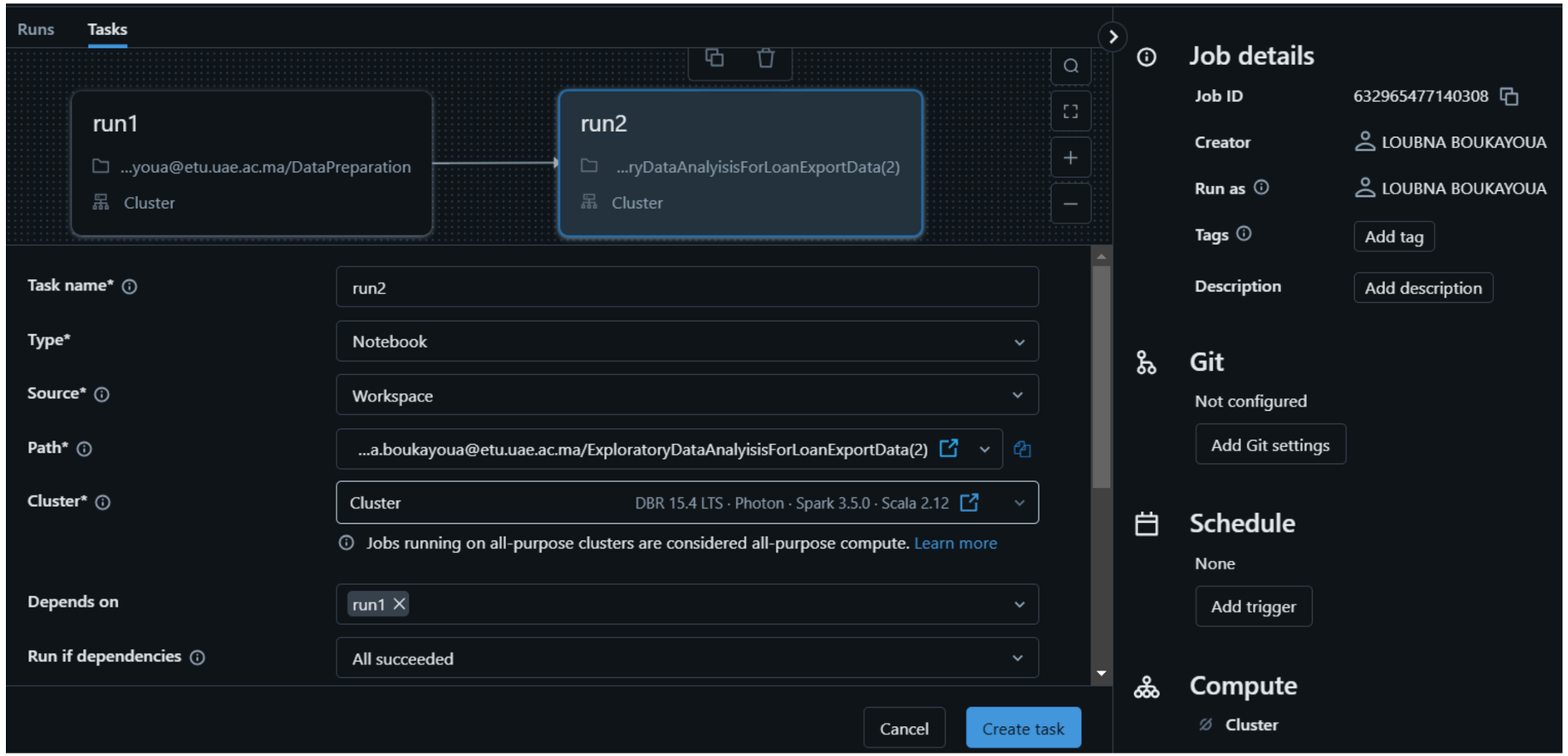
PROJECT COMPONENTS: BLOB STORAGE & DATABRICKS

Why Blob Storage:

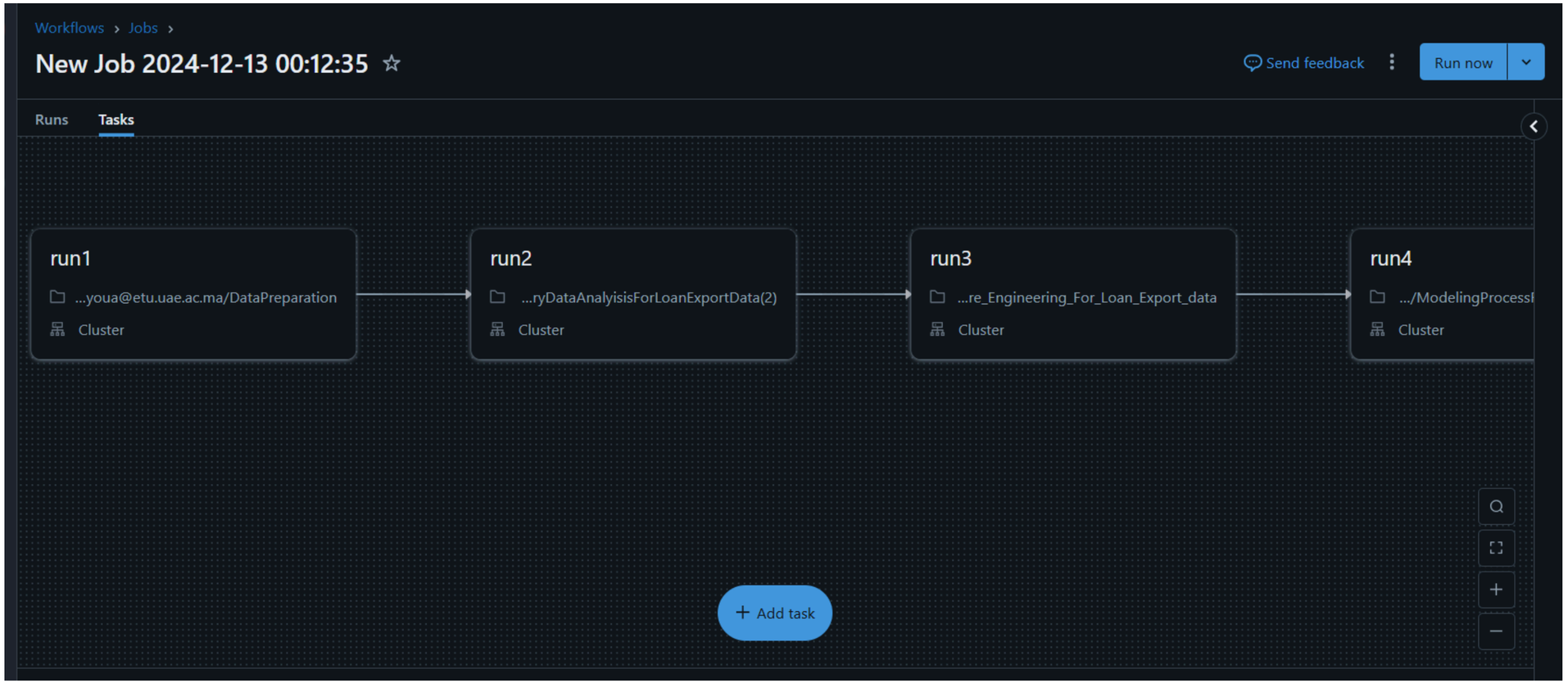
- Scalability
- Cost-effective
- Flexibility
- Integration



PROJECT COMPONENTS: DATABRICKS JOBS



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:

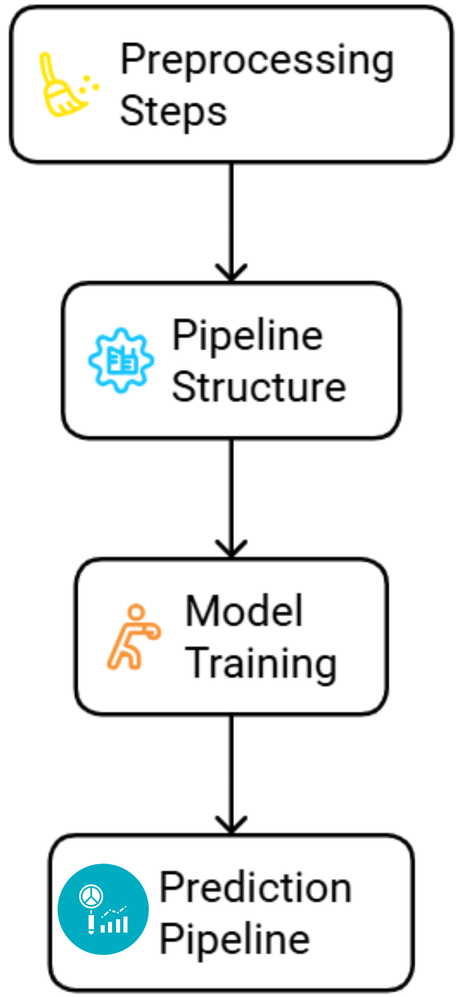
- Data cleaning.
- Preprocessing.
- Model Building: Training and validating machine learning models.

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

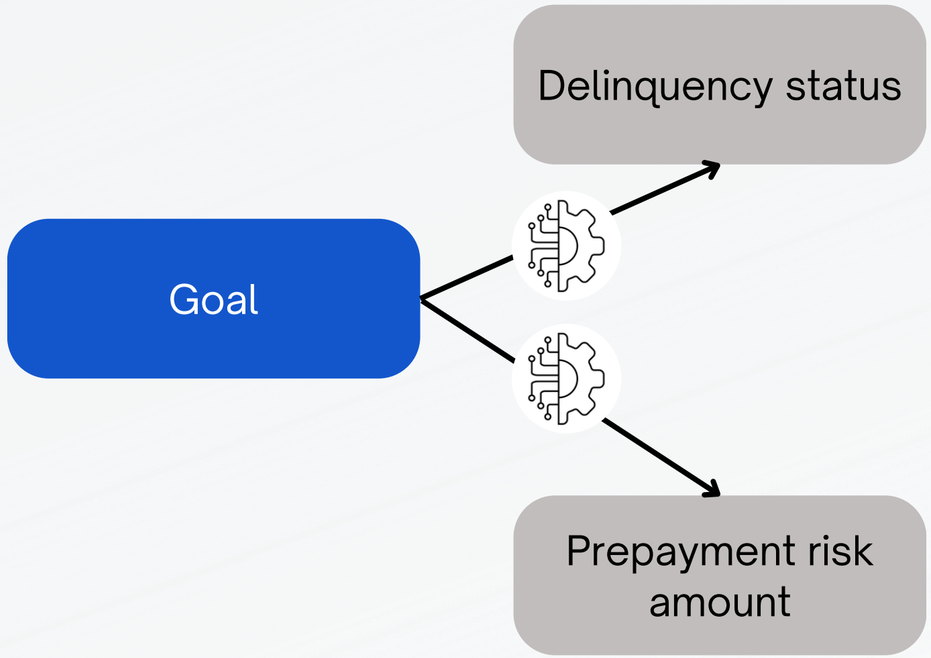
DATA COLLECTION

COLUMN DETAILS

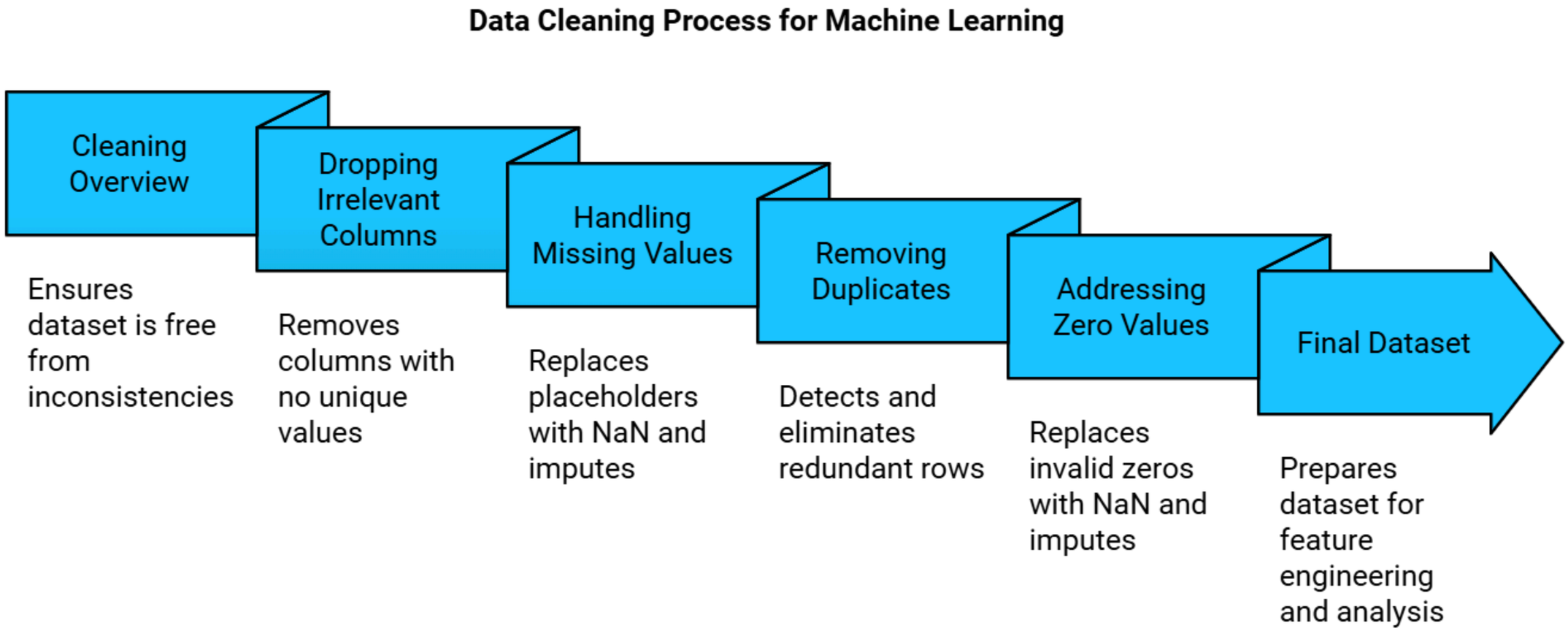
PROPERTY INFORMATION

LOAN CHARACTERISTICS

BORROWER BEHAVIOR

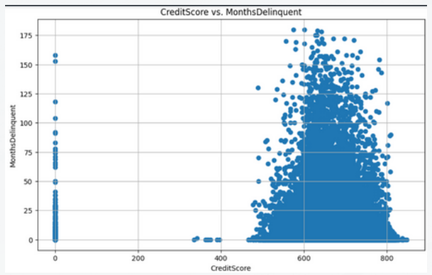


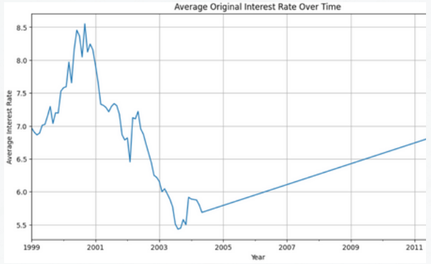
DATA PREPROCESSING

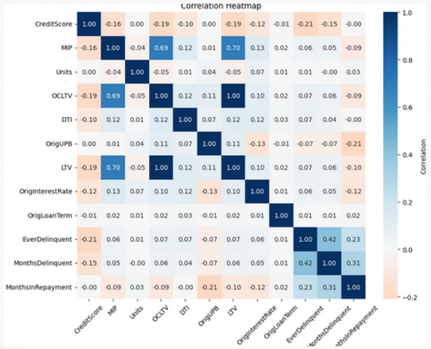


EXPLORATORY DATA ANALYSIS (EDA)

- Data Analysis
  - Univariate Analysis
    - Numerical Features → Histograms, KDE Plots, Summary Statistics
    - Categorical Features → Bar Plots, Pie Charts
  - Bivariate Analysis
    - Numerical vs Numerical → Scatter Plots, Correlation Coefficients
  - Multivariate Analysis
    - Correlation Heatmaps
    - PCA



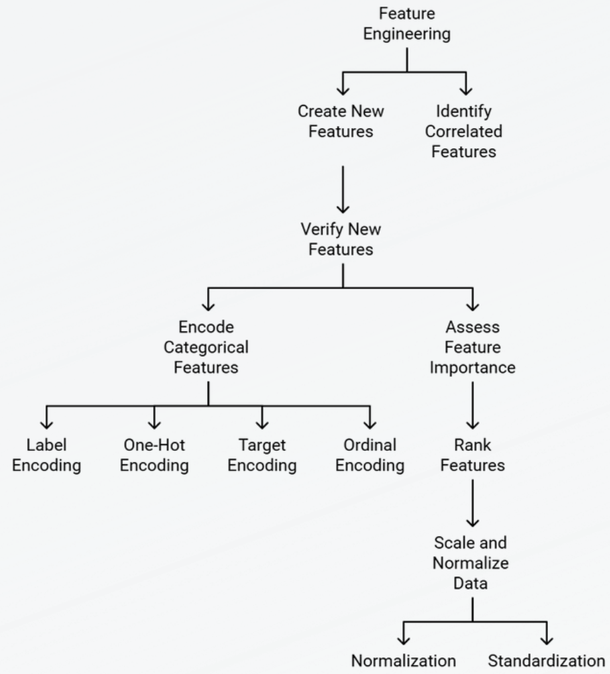


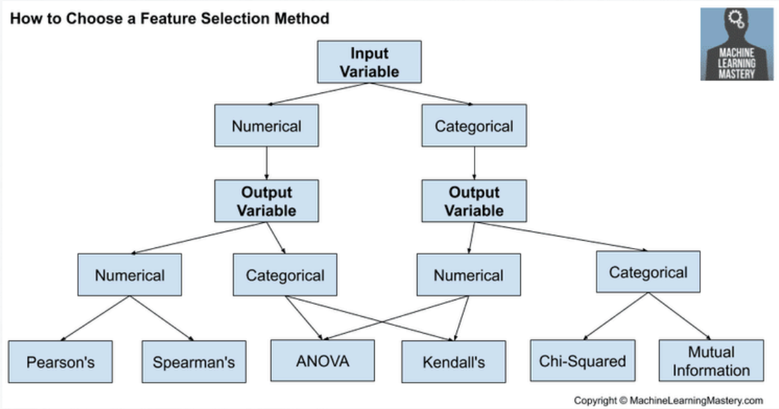


PREPARING DATA FOR MODELING

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

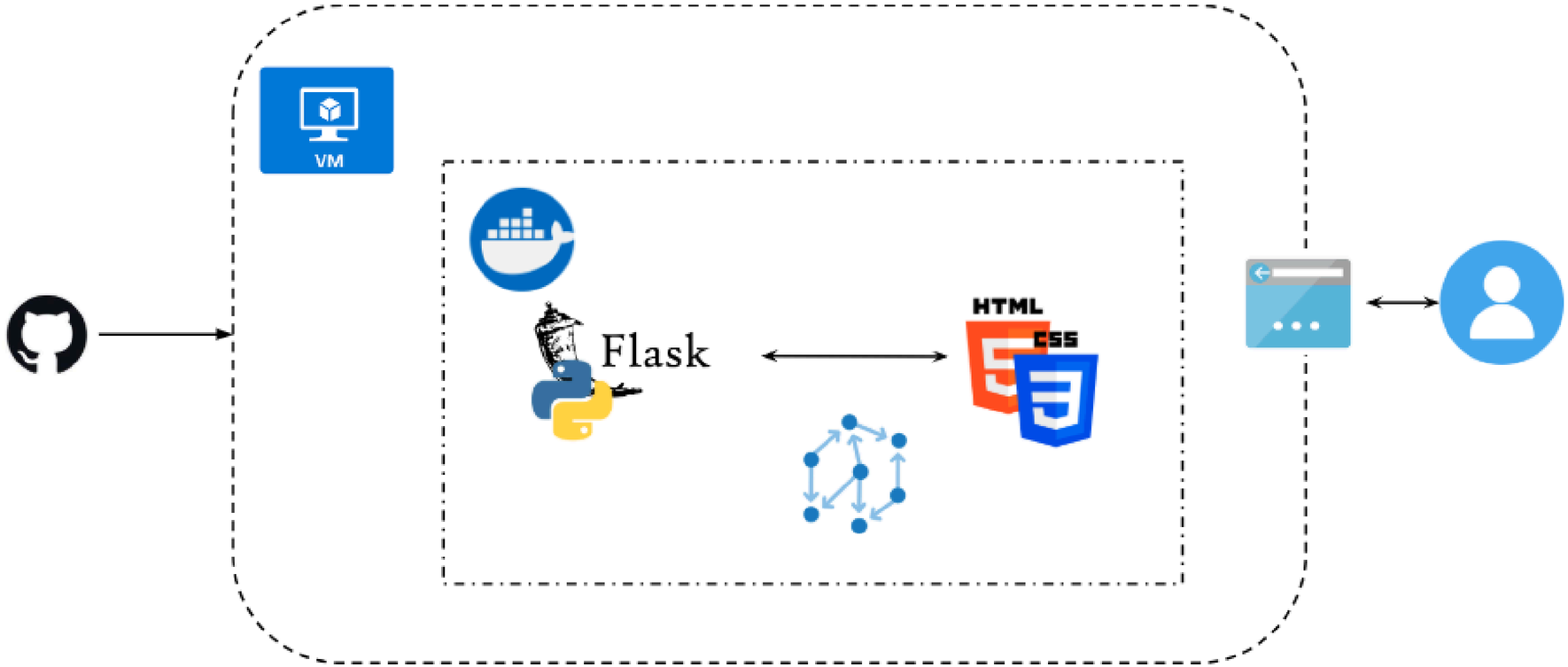
Algorithm	Testing_accuracy	Roc_Auc_score	f1_score
Logistic Regression	1.000000	1.000000	1.000000
Decision Tree	1.000000	1.000000	1.000000
GaussianNB	0.988425	1.000000	0.971297
BernoulliNB	0.904568	0.848416	0.717914

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

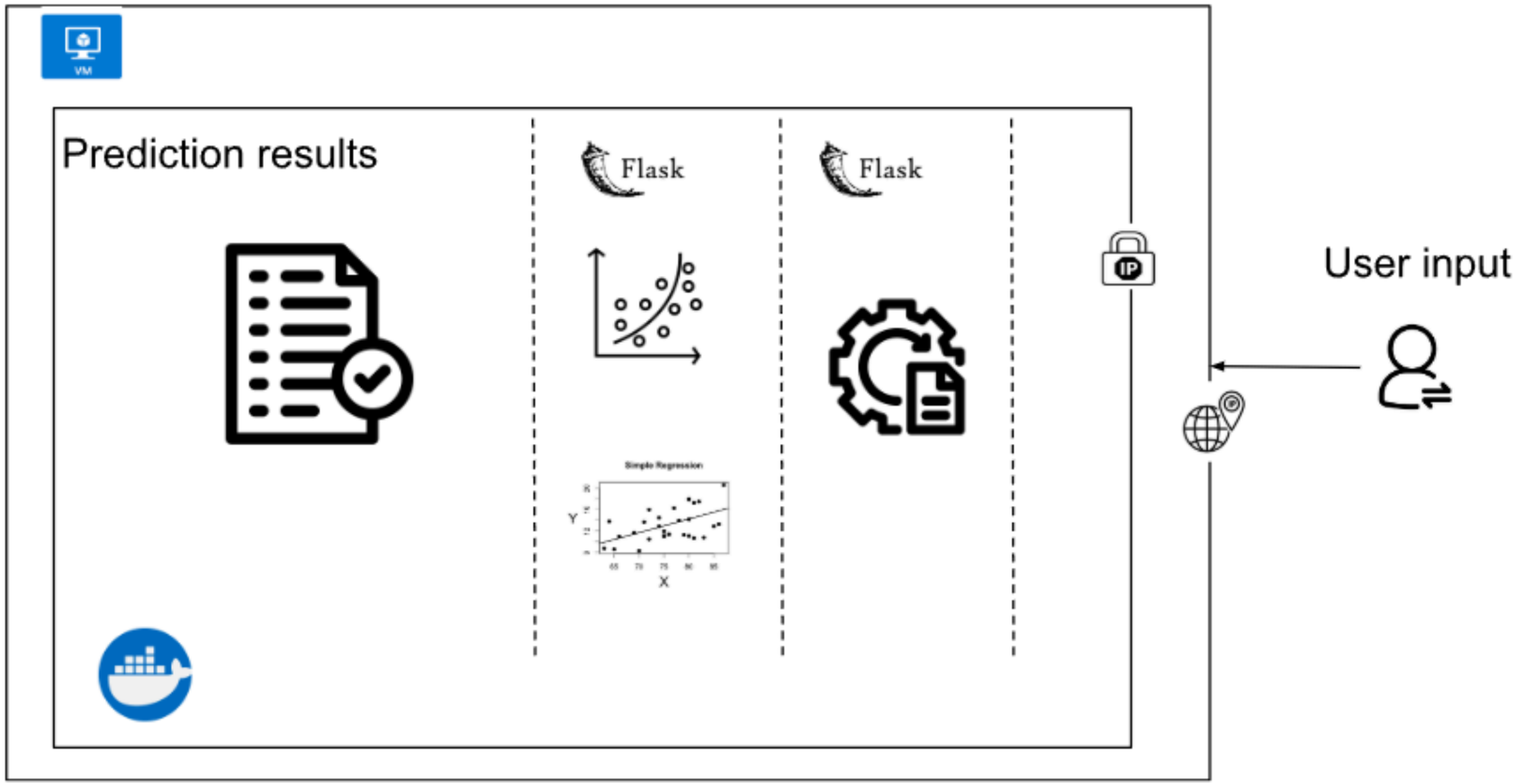
Data problem fixed:

- Fixing Data Leakage
- Preventing Overfitting
- Handling Outliers

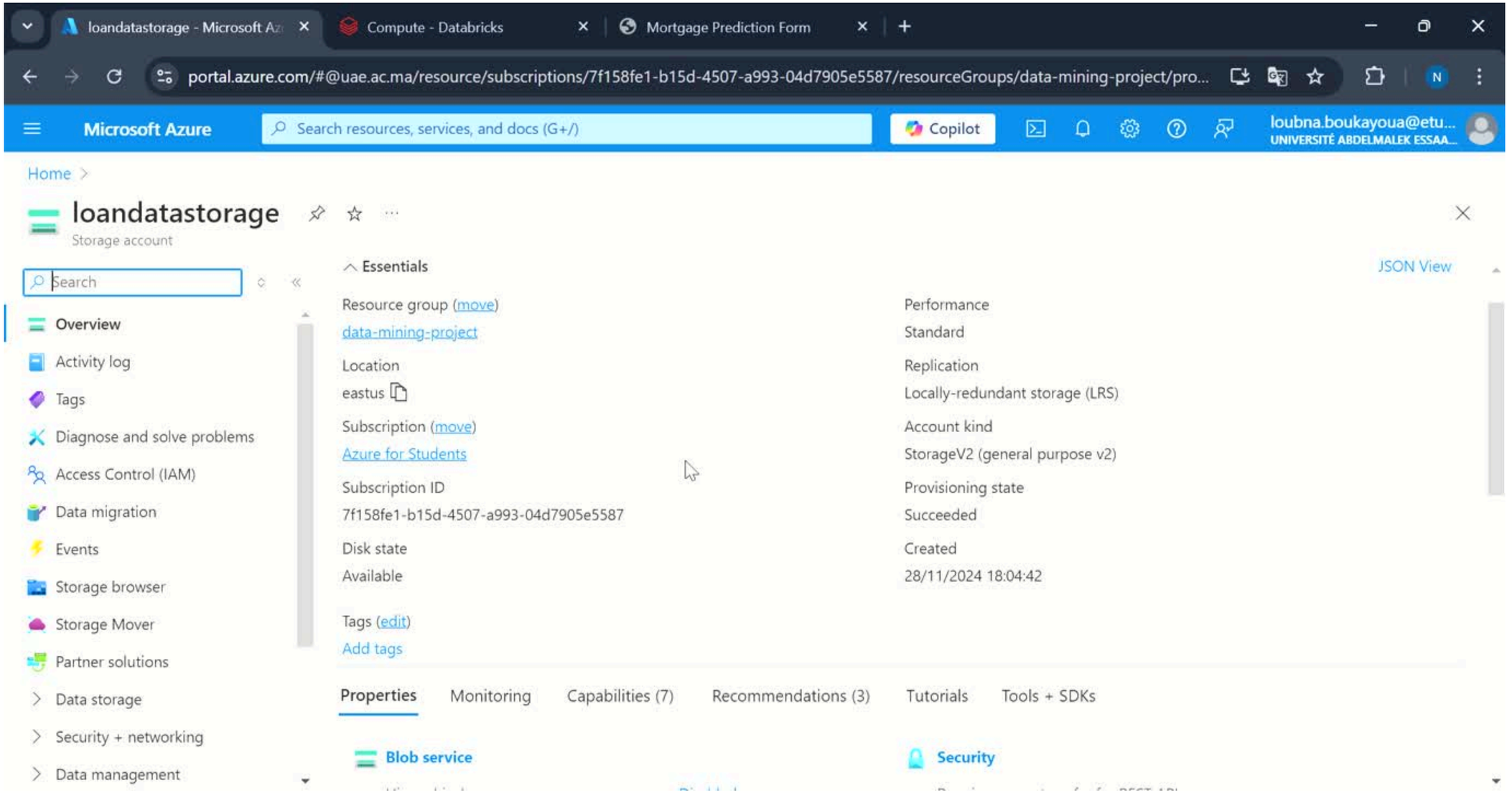
Deployment architecture



Classification and Regression Pipeline Workflow



Technical demonstration



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