Anoma Data Project Report

1. Introduction

1.1 Project Overview

Anoma Data is a machine learning project aimed at detecting anomalies in a dataset. The primary goal is to build a model that can accurately classify data points as normal or anomalous, helping businesses detect irregular patterns that may indicate fraud, errors, or unusual behavior.

1.2 Objective

- Perform **Exploratory Data Analysis (EDA)** to understand the dataset.
- Clean and preprocess the data for model training.
- Train an optimal **machine learning model** for anomaly detection.
- Evaluate the model's performance using appropriate metrics.

---

2. Exploratory Data Analysis (EDA)

2.1 Dataset Overview

- The dataset consists of **X rows** and **Y columns**.
- The target column is **"y.1"**, which determines whether a data point is anomalous.
- Data types include **float, integer, and categorical variables**.

2.2 Data Cleaning and Preprocessing

- **Handling Missing Values**: Missing values were handled by **imputation** or **removal**.
- **Outlier Detection**: Boxplots and statistical methods were used to detect and handle outliers.
- **Feature Scaling**: Standardization or normalization was applied where necessary.
- **Encoding Categorical Variables**: One-hot encoding or label encoding was used for categorical data.

---

3. Model Development

3.1 Splitting the Data

- **Training Set Size**: 80% of the data.
- **Test Set Size**: 20% of the data.

3.2 Model Selection

The following models were evaluated:

- **Logistic Regression**
- **Random Forest Classifier**
- **Support Vector Machine (SVM)**
- **XGBoost**

After evaluation, **[Best Model]** was selected as the final model.

3.3 Model Training

- The model was trained using the **training dataset**.
- Hyperparameter tuning was performed using **GridSearchCV or RandomizedSearchCV**.

---

4. Model Evaluation

4.1 Evaluation Metrics

- **Accuracy**: X%
- **Precision**: X%
- **Recall**: X%
- **F1 Score**: X%
- **ROC-AUC Score**: X%

These metrics indicate that the model performs **well/needs improvement**.

---

5. Conclusion & Future Work

5.1 Conclusion

- The anomaly detection model successfully identifies anomalous patterns in the dataset.
- The chosen model achieves an accuracy of **X%**, which meets/exceeds the project requirements.

5.2 Future Work

- Improve feature engineering techniques.
- Experiment with **deep learning models** such as **Autoencoders or LSTMs**.
- Deploy the model into a real-time monitoring system.

---

6. References

- [Include relevant research papers, articles, and documentation used]

This report provides a detailed summary of the **Anoma Data Project**, including data analysis, model development, and results.

Prepared by: [Debarghya Roy Choudhury]
 Date: [31/01/2025]