Data Management Plan (DMP): Fraud Detection in Online Transactions

# 1. Data Summary

This project aims to detect fraudulent online transactions using machine learning. The dataset includes structured data such as transaction amount, timestamp, customer location, payment method, and transaction history. The primary data source is a synthetic dataset created to mimic real-world transaction behavior with a 5% fraud ratio.  
  
The dataset supports supervised learning for binary classification (fraudulent vs. legitimate) and is used to train models like Random Forest and Gradient Boosting. Data outputs include trained models, evaluation metrics (accuracy, precision, recall, F1-score, ROC-AUC), and visualizations such as confusion matrices and ROC curves.

# 2. FAIR data

Findable: Data is hosted in DBRepo and will be published in Zenodo with a DOI. All files are named consistently with metadata and README documentation.  
  
Accessible: All files are openly accessible via Zenodo and GitHub under appropriate licenses.  
  
Interoperable: Data is stored in CSV format; code is in Jupyter notebooks with standard libraries (pandas, sklearn). Metadata uses CodeMeta and FAIR4ML standards.  
  
Reusable: Data and code are licensed under CC BY 4.0 and MIT respectively. Full metadata and documentation are provided to ensure reproducibility.

# 3. Allocation of Resources

Data storage and backup are managed through DBRepo and TUWRD. Versioning and public sharing are facilitated via GitHub and Zenodo. The primary resources are the author's institutional access to these platforms.

# 4. Data Security

Data is synthetic and poses no personal data risks. Access to DBRepo and TUWRD is authenticated via TU Wien credentials. Public versions are reviewed to ensure no sensitive data is included.

# 5. Ethical Aspects

The experiment uses synthetic data only. No personal or sensitive information is processed, hence no ethical concerns arise.

# 6. Other Issues

The project complies with institutional guidelines on data sharing and research reproducibility. Outputs will be maintained post-project on Zenodo and GitHub. ORCID and DOIs are used for researcher and artifact identification.