**Fraud Detection Model**

**Report**

**1. Project Overview**

This project focuses on developing a machine learning model to detect potentially fraudulent individuals by classifying them as "Risky" or "Good" based on their taxable income and demographic details.

**2. Objective**

To create a classification model that flags individuals with **taxable income ≤ 30,000** as “Risky”, aiding financial institutions in early fraud detection and credit decision-making.

3. **Solution Architecture**

+------------------+

| Raw Data |

+--------+---------+

|

+----------v------------+

| Data Preprocessing |

| - Label Risk column |

| - Encode categorical |

+----------+------------+

|

+----------v------------+

| Exploratory Data |

| Analysis (EDA) |

| - Graphs using Seaborn|

+----------+------------+

|

+----------v------------+

| Random Forest Model |

| - Train/Test split |

| - Fit & Evaluate |

+----------+------------+

|

+----------v------------+

| Model Evaluation |

| - Accuracy, ConfMat |

| - Feature Importance |

+-----------------------+

**4. Methodology**

| **Step** | **Description** |
| --- | --- |
| Data Collection | Imported dataset containing demographic and income information |
| Data Preparation | Created binary target label Risk, encoded categorical features |
| EDA | Analyzed distributions and feature relationships using Seaborn |
| Model Building | Used Random Forest Classifier with hyperparameters tuning |
| Evaluation | Assessed model using accuracy, classification report, and feature insights |
| Deployment Readiness | Model and preprocessing pipeline ready for integration with real-time input |

**5. Technologies Used**

* Language: Python
* Libraries: Pandas, Seaborn, Scikit-learn, Matplotlib
* IDE: Jupyter Notebook
* Environment: Local machine

**6. Time Taken**

| **Task** | **Estimated Time** |
| --- | --- |
| Data Understanding | 2 hours |
| EDA and Visualization | 3 hours |
| Data Preprocessing | 1 hour |
| Model Development | 2 hours |
| Evaluation and Tuning | 1 hour |
| Documentation & Reporting | 1 hour |
| **Total Time** | **10 hours** |

**7. Challenges Faced**

* Imbalanced class distribution (more "Good" than "Risky")
* Encoding categorical variables while preserving interpretability
* Visualizing complex relationships in mixed-type data
* Avoiding model overfitting on limited data

**8. Project Complexity**

* **Complexity Level**: Moderate
* **Reason**: Although Random Forest simplifies modeling, meaningful feature engineering and careful interpretation are crucial for fraud contexts.

**9. Business Impact**

* Enables institutions to proactively filter out high-risk individuals
* Automates early fraud detection without relying on manual profiling
* Can be integrated into financial approval systems

**10. Conclusion**

The model provides a reliable classification system to identify high-risk individuals. With feature importance insights and EDA visualizations, the approach is transparent and scalable across similar financial risk domains.