**Project Title: Fraud detection by using Random forest classifier**

**Project Duration:** 1 Days  
**Date Completed:** April 20, 2025  
**Author:** *sagar khese*

**1. project overview :**

This project focus on developing a machine learning model to classify individual ‘risky’ and

‘good’ based on taxable income and demographic features by using random forest model

**2. Dataset Overview**

* **Source file:** fraud check.csv
* **Shape:** (600, 6) columns
* **Features:** underaged , marital status , city population , work exp , urban
* **Target Variable:** taxable income(risky = 30000 , good >30000)

**3. Solution Architecture**

**Raw data > data processing > Feature encoding (**underaged , marital status , city population , work exp , urban) **> Model Training (RF) > Model Evaluation (accuracy score , distribution on feature ,correlation map) > Model saving**

**4. Methodology**

| **Step** | **Reason** |
| --- | --- |
| **Data Collection** | Loading csv file from local library |
| **Data Cleaning** | Checked for null values, selected relevant columns and convert tax income into categorical variable |
| **Exploratory Data Analysis (EDA)** | Visualized categorical feature distributions and correlation heatmaps , bar graph and pariplot |
| **Feature Engineering** | Used LabelEncoder for converting categorical columns into numeric format |
| **Modeling** | Random forest with entropy criterion with split data into x and y |
| **Evaluation** | Accuracy score for training data |
| **Deployment Prep** | Saved the model using Joblib for future use |

**5. Time Taken**

| **Task** | **Time Spent** |
| --- | --- |
| Data Cleaning & EDA | 1 hour |
| Model Building | 2 hour |
| Visualization | 30 minutes |
| Documentation | 1 hour |
| **Total Time** | **4 hours** |

**6. Challenges Faced**

* Size of dataset was small, which could lead to overfitting.
* Deciding how to categorize the risk threshold (Taxable Income cutoff)
* Handling categorical variables required careful encoding

**7. Complexity**

* **Complexity:** medium
* **Skills Required:** Python, Pandas, Seaborn, Scikit-learn, Data Visualization

**8. Recommendations**

* Collect more data for better generalization.