**Risk Prediction Analysis**

**Introduction**

This document outlines the end-to-end process of analyzing a dataset related to taxable income. The objective is to predict taxable income categories ('Risky' or 'Good') based on various features. The dataset includes columns such as Undergrad, Marital Status, Taxable Income, City Population, Work Experience, and Urban.

**Methodology**

**1. Data Preprocessing:**

**1.1 Importing Necessary Packages and Extracting Data**

The project started by importing essential Python packages and extracting the dataset.

**1.2 Categorizing Taxable Income**

Taxable Income was categorized into 'Risky' and 'Good' based on a condition: if taxable income is less than or equal to 30000, it was labeled as 'Risky'; otherwise, it was labeled as 'Good'. This becomes the target variable 'y'.

**1.3 Feature Engineering**

The dataset was split into features (X) and the target variable (y). Mapping was applied to 'Undergrad' and 'Urban' columns, and one-hot encoding (get\_dummies) was used for 'Marital Status' and 'Taxable Income' columns.

**1.4 Handling Missing Values**

No missing values were found in the dataset, ensuring data integrity for analysis.

**2. Modeling:**

**2.1 Random Forest Classifier**

The Random Forest Classifier was chosen for modeling. The features ['Undergrad', 'Work Experience', 'Urban', 'Divorced', 'Married', 'Single'] were selected as X variables, and 'Risky' was set as the y variable.

**2.2 Data Scaling**

Data scaling was performed using Standard Scaler to standardize the feature values.

**2.3 Model Fitting and Prediction**

The dataset was fitted into the Random Forest Classifier, and predictions were made. The model achieved an accuracy of 87.50%.

**Conclusion**

The project successfully analyzed the dataset to predict taxable income categories using a Random Forest Classifier. Through careful preprocessing and feature engineering, the model achieved an accuracy of 87.50%.