**Random Forest Model**

**Methodology for Random Forest Model**

**1) Importing Libraries as follows-**

Pandas, Seaborn, Time, sklearn.preprocessing for LabelEncoder ,Random forest classifier,train\_test\_split,matplotlib, r2\_score, sklearn.metrics for confusion\_matrix, classification\_report, accuracy\_score

**2) Reading data**

Read data from Company\_data and fraud\_check dataset by defining **pd.read\_data()** function and access those data by calling **read\_data()** function

**3) Info of data**

**4)** **Convert sales to Sales\_Category target Categorical variables and Taxable\_Income to Risk Target Categorical variable**

**5) Encode the Sales\_Category and Risk Target variables**

**6) One-hot encoding for categorical variables**

**7) Check the data types to confirm all categorical variables are encoded**

**8) Exploratory Data Analysis**

Cleaning data, ploting graphs andshowing correlation with data by using functions like

Plt.figure()

Sns.displot()

Plt.show()

df.corr()

# Histogram for Sales

# Boxplot for Advertising

# Countplot for Sales Category vs. Shelf Location

# Scatter plot for Price vs. Sales

# Countplot for Urban vs. Sales Category

# Calculate the correlation matrix for company\_data and fraud\_check dataset

# Plot the heatmap for company\_data and fraud\_check dataset

## Countplot for distribution of risk

# Histplot for Distribution of city\_population

# Scattarplot for work\_Experience VS risk

**9) Define features and target variable X and y**

**10) Data splitting**

define X**\_train, X\_test,y\_train, y\_test** variables by using **train\_test\_split()** function

**11)**  **Define random forest model**

**12) Fitting the X\_train y\_train dataset in model**

**13) Check model score**

**14) Predictions for model**

**15) Check Accuracy score**

**16)** **Check Confusion matrix**

**17)** **Feature importance for plotting data**

**Challenges faced while doing Random forest classifier**

**1. Overfitting**

* **Description**: Although Random Forests are less prone to overfitting than single decision trees, they can still overfit, especially if the number of trees is very large or if the trees are deep.

### 2. ****Data Imbalance****

* **Description**: If the target classes are imbalanced (e.g., many more "Good" sales than "Risky"), the model may be biased towards the majority class.

### 3. ****Feature Importance Interpretation****

* **Description**: While Random Forest provides feature importance scores, interpreting these scores can be challenging, especially with correlated features.

### 4. ****Handling Categorical Variables****

* **Description**: Random Forests require numerical input, which means categorical variables must be appropriately encoded.

### 5. ****Model Complexity****

* **Description**: Random Forest models can be complex and difficult to visualize or explain compared to simpler models like logistic regression.

### 6. ****Feature Correlation****

* **Description**: Highly correlated features can lead to misleading feature importance scores and affect model performance.