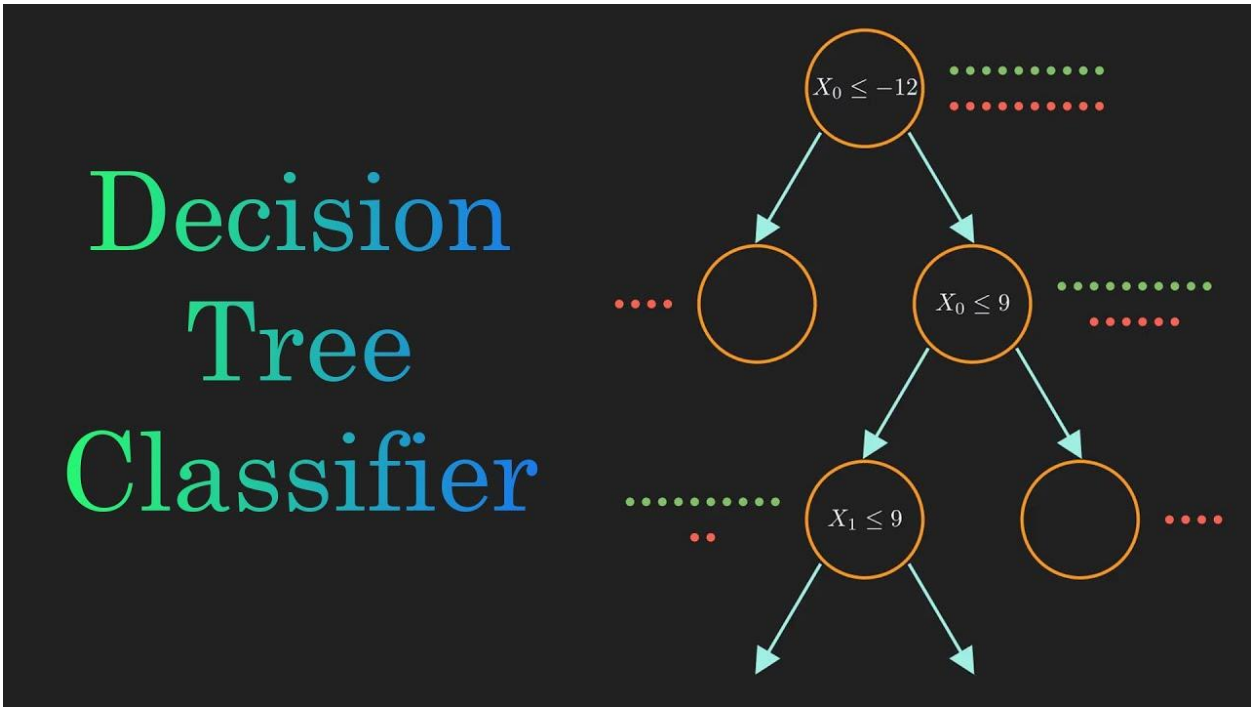
Classification on the basis of ‘job_security’ as target variable

Name – Sujay Kumar

Roll – 22052514

Cs 33 DMDW



Report

1. Dataset Overview and Feature Descriptions

The dataset, `covid_impact_on_work.csv`, analyzes various aspects of employees' work conditions during the COVID-19 pandemic. It contains the following features:

- **Sector**: Industry sector of the employee (e.g., Retail, IT, Education).
- **Increased_Work_Hours**: Indicates if work hours have increased due to the pandemic.
- **Work_From_Home**: Indicates if the employee is working from home.
- **Hours_Worked_Per_Day**: Average hours worked per day.
- **Meetings_Per_Day**: Average number of meetings per day.
- **Productivity_Change**: Change in productivity level.
- **Health_Issue**: Indicates if the employee reported health issues.
- **Job_Security** (Target): Indicates if the employee's job security was affected.
- **Commuting_Changes**: Indicates changes in commuting routines.
- **Technology_Adaptation**: Reflects adaptation to new technology demands.
- **Team_Collaboration_Challenges**: Reports challenges in team collaboration.

Dropped Columns:

- **Stress_Level**: Reported stress levels during the pandemic.
- **Childcare_Responsibilities**: Indicates if the employee has childcare responsibilities.
- **Salary_Changes**: Indicates any changes in salary.

- **Affected_by_Covid**: Indicates if the employee was directly affected by COVID-19.

2. Data Transformation and Cleaning

- **Dropping Unnecessary Columns**: Columns like `Stress_Level`, `Childcare_Responsibilities`, `Salary_Changes`, and `Affected_by_Covid` were dropped as they were deemed irrelevant to our predictive model for `Job_Security`.
- **Extraction of Integer Values**: Some columns, such as `Meetings_Per_Day` and `Hours_Worked_Per_Day`, contained floating-point values that needed simplification. The `extract_first_number` function was created to convert these values to integer representations by taking the integer part only.
- **Encoding Categorical Data**: `Sector`, a categorical column, was encoded into numerical values using `LabelEncoder` to ensure compatibility with the machine learning model.

3. Feature Selection

- After data transformation, the features selected for the model were:
  - `Increased_Work_Hours`
  - `Work_From_Home`
  - `Hours_Worked_Per_Day`
  - `Meetings_Per_Day`
  - `Productivity_Change`
  - `Health_Issue`
  - `Commuting_Changes`
  - `Technology_Adaptation`
  - `Team_Collaboration_Challenges`
  - `Sector_encoded` (encoded `Sector` column)

- The target variable chosen for the classification task was `Job_Security`.

4. Model Details

- **Decision Tree Classifier**:
  - An initial Decision Tree model was created with a default configuration.
  - Performance evaluation revealed an accuracy of 54% with some bias in classification towards the majority class.
- **Hyperparameter Tuning with Grid Search**:
  - A parameter grid for the Decision Tree was defined, exploring options for `max_depth`, `min_samples_split`, `min_samples_leaf`, `max_features`, and `criterion`.
  - Grid search with cross-validation (10 folds) was used to identify the best combination of parameters, resulting in an optimized model with parameters:
    - `criterion`: `entropy`
    - `max_depth`: `20`
    - `max_features`: `log2`
    - `min_samples_leaf`: `6`
    - `min_samples_split`: `10`
  - The optimized Decision Tree yielded an accuracy of 55.55%, showing limited improvement but a more balanced classification report.

- **Random Forest Classifier**:
  - Random Forest was used as an ensemble method to improve accuracy and generalization.

- Parameter tuning for `n_estimators`, `max_depth`, `min_samples_split`, and `min_samples_leaf` was conducted using GridSearchCV.
- The best parameters for Random Forest were:
  - `max_depth`: None
  - `min_samples_leaf`: `1`
  - `min_samples_split`: `2`
  - `n_estimators`: `200`
- The final Random Forest model achieved an accuracy of 56.65% with a better recall for both classes, indicating improved model performance in capturing the distribution of both secure and insecure jobs.

5. Evaluation Results

- The **Random Forest Classifier** outperformed the Decision Tree in both accuracy and balance between precision and recall.
- Confusion matrices and classification reports showed that while the Random Forest model provided improved accuracy, there is room for further optimization, potentially through additional feature engineering or ensemble methods.

This analysis highlighted feature selection, transformations, and model tuning steps in the modeling process, providing a comprehensive understanding of employee job security factors during COVID-19.