Data Preprocessing and Feature Engineering in Machine Learning

**Objective:**

This assignment aims to equip you with practical skills in data preprocessing, feature engineering, and feature selection techniques, which are crucial for building efficient machine learning models. You will work with a provided dataset to apply various techniques such as scaling, encoding, and feature selection methods including isolation forest and PPS score analysis.

Dataset:

Given "Adult" dataset, which predicts whether income exceeds $50K/yr based on census data.

Tasks:

**1. Data Exploration and Preprocessing:**

* Load the dataset and conduct basic data exploration (summary statistics, missing values, data types).
* Handle missing values as per the best practices (imputation, removal, etc.).
* Apply scaling techniques to numerical features:
  + Standard Scaling
  + Min-Max Scaling
* Discuss the scenarios where each scaling technique is preferred and why.

**2. Encoding Techniques:**

* Apply One-Hot Encoding to categorical variables with less than 5 categories.
* Use Label Encoding for categorical variables with more than 5 categories.
* Discuss the pros and cons of One-Hot Encoding and Label Encoding.

**3. Feature Engineering:**

* Create at least 2 new features that could be beneficial for the model. Explain the rationale behind your choices.
* Apply a transformation (e.g., log transformation) to at least one skewed numerical feature and justify your choice.

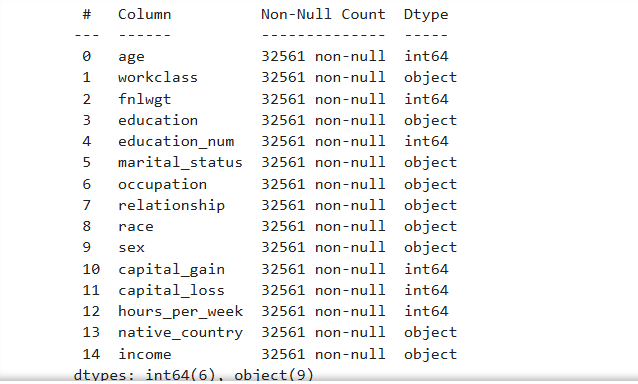
**4. Feature Selection:**

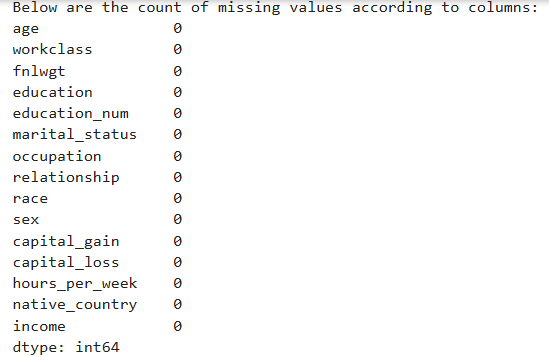
* Use the Isolation Forest algorithm to identify and remove outliers. Discuss how outliers can affect model performance.
* Apply the PPS (Predictive Power Score) to find and discuss the relationships between features. Compare its findings with the correlation matrix.

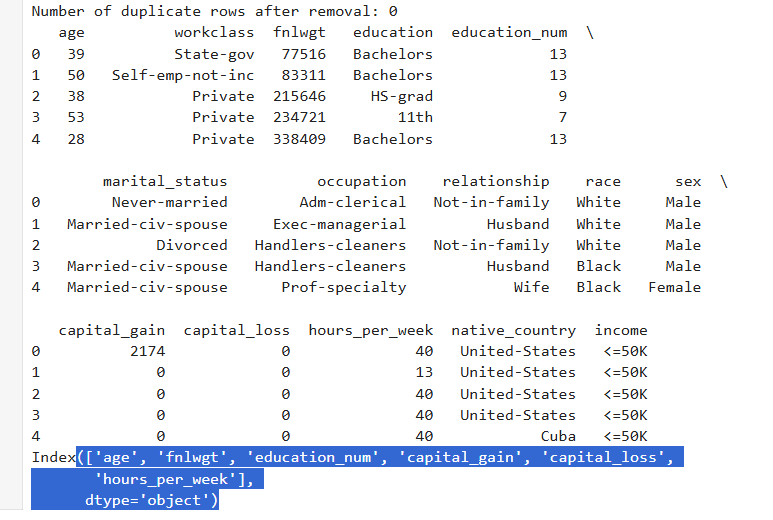
**Observations:**

**1. Data Exploration and Preprocessing:**

* Load the dataset and conduct basic data exploration (summary statistics, missing values, data types).
* Handle missing values as per the best practices (imputation, removal, etc.).
* Apply scaling techniques to numerical features:
  + Standard Scaling
  + Min-Max Scaling





* Discuss the scenarios where each scaling technique is preferred and why.

# When to Use Standard Scaling vs Min-Max Scaling

# Standard Scaling (Z-score normalization) is preferred when data has a normal distribution and outliers are present.

# It centers the data to have mean=0 and standard deviation=1.

# Min-Max Scaling is useful when data has a non-Gaussian distribution and needs to be scaled to a fixed range (e.g., 0 to 1).

# It is commonly used for algorithms that require bounded input features, like neural networks.

|  |  |
| --- | --- |
| **Scaling Technique** | **When to Use** |
| **Standard Scaling (Z-score)** | When features follow a normal distribution and machine learning models (e.g., linear regression, logistic regression, SVM) assume standardization. |
| **Min-Max Scaling** | When features have varying scales but do not follow a normal distribution (e.g., decision trees, neural networks, KNN, etc.). It's also useful when feature values need to be between 0 and 1. |

**Insights from the Shapiro-Wilk Normality Test Results**

The **Shapiro-Wilk test checks if a feature follows a normal distribution**. It provides:

* **Statistic**: Measures how close the data is to normality (closer to 1 means more normal).
* **p-value**: If **p < 0.05**, the data is **not normally distributed** (rejects the null hypothesis of normality).

**For given data set Min-Max Scaling is prefreable**

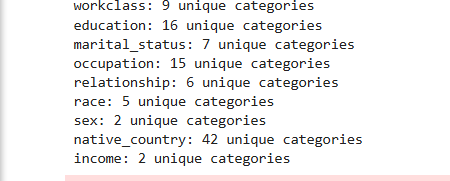
|  |  |  |  |
| --- | --- | --- | --- |
| **Feature** | **Statistic** | **p-value** | **Interpretation** |
| **age** | 0.9668 | 0 | Not normally distributed |
| **fnlwgt** | 0.9223 | 0 | Not normally distributed |
| **education\_num** | 0.9266 | 0 | Not normally distributed |
| **capital\_gain** | 0.1228 | 0 | Highly skewed, not normal |
| **capital\_loss** | 0.2184 | 0 | Highly skewed, not normal |
| **hours\_per\_week** | 0.8851 | 0 | Not normally distributed |

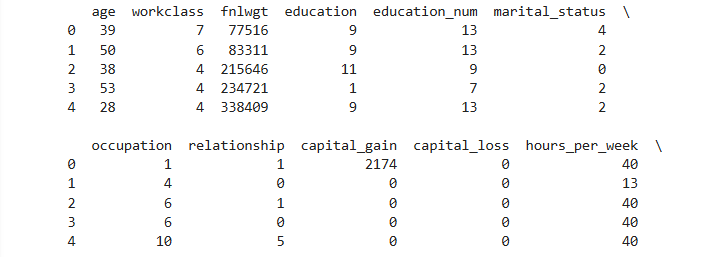
**2. Encoding Techniques:**

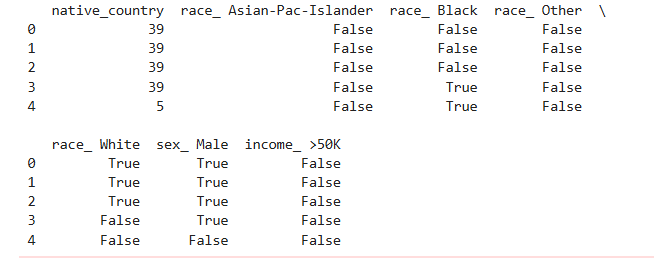
* Apply One-Hot Encoding to categorical variables with less than 5 categories.
* Use Label Encoding for categorical variables with more than 5 categories.
* Discuss the pros and cons of One-Hot Encoding and Label Encoding.

**Apply Encoding Based on Category Count**

**One-Hot Encoding** (If categories ≤ 5)  
 **Label Encoding** (If categories > 5)







**3. Feature Engineering:**

* Create at least 2 new features that could be beneficial for the model. Explain the rationale behind your choices.
* Apply a transformation (e.g., log transformation) to at least one skewed numerical feature and justify your choice.

**Feature 1: "is\_high\_earner" (Binary Feature Based on Capital Gain & Capital Loss)**

**Rationale**:

**Capital Gain & Capital Loss** indicate investments or assets, which often correlate with higher income.

Instead of using them separately, we can create a **binary feature:**

* 1 if **capital\_gain** or **capital\_loss** is greater than zero (indicating financial investment activity).
* 0 otherwise.

**Feature 2: "work\_hours\_category" (Categorizing** hours\_per\_week**)**

**Rationale**:

**Work hours’** influence income level, but instead of treating it as a continuous variable, we can categorize it:

* **"Part-Time"** (≤ 30 hours)
* **"Full-Time"** (31-50 hours)
* **"Overtime"** (> 50 hours)

**Summary of Feature Engineering**

* **New Features:**

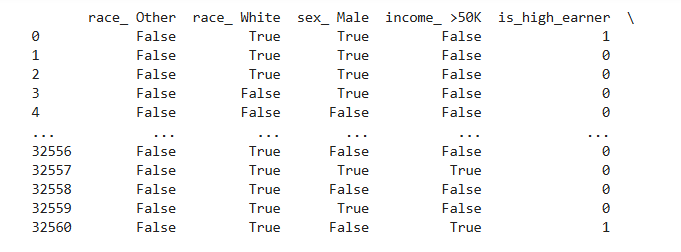
is\_high\_earner (Binary: Based on capital\_gain and capital\_loss)

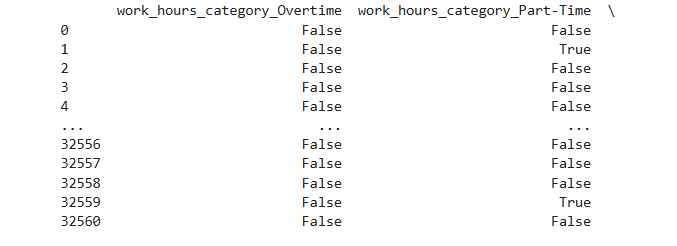
work\_hours\_category (Categorized from hours\_per\_week → One-Hot Encoded)

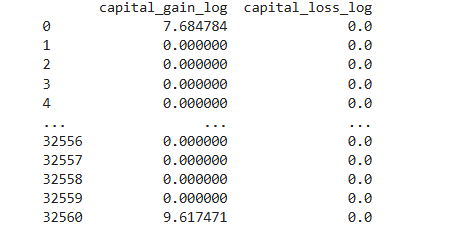
* **Applied Log Transformation to Reduce Skewness:**

capital\_gain\_log

capital\_loss\_log





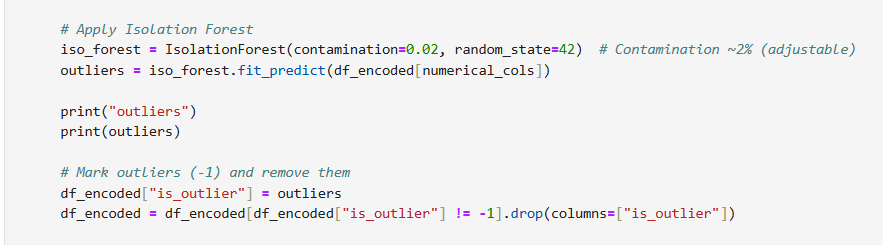


**4. Feature Selection:**

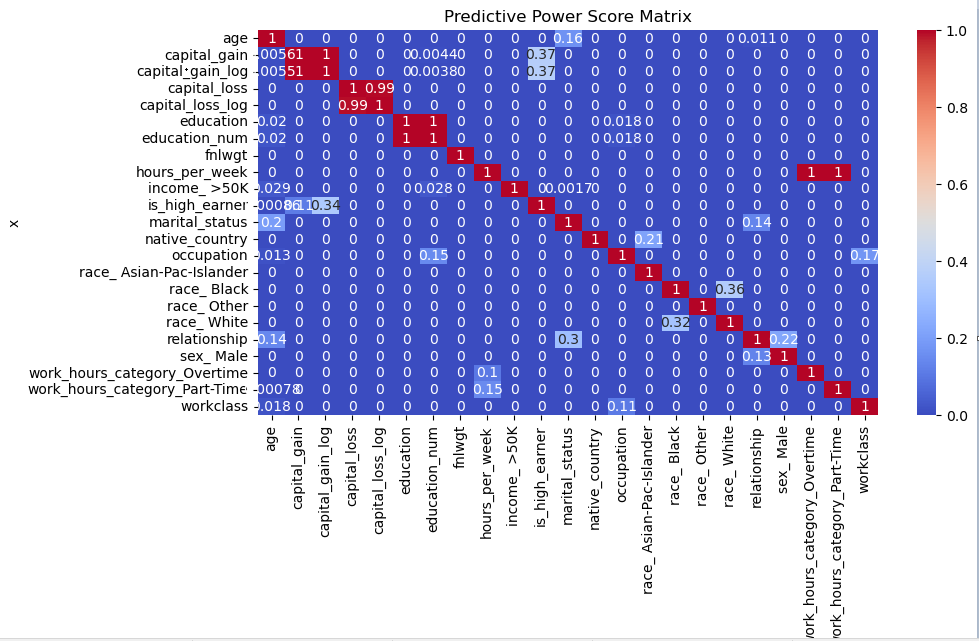
* Use the Isolation Forest algorithm to identify and remove outliers. Discuss how outliers can affect model performance.

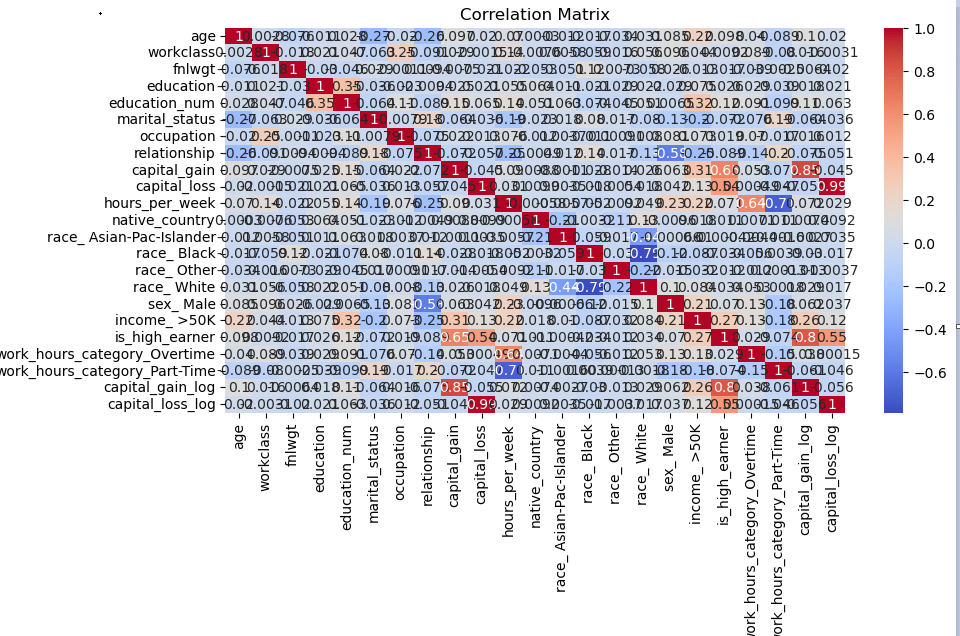
**Handling Outliers with Isolation Forest**

Isolation Forest is an **unsupervised learning algorithm** that detects outliers by isolating data points. Outliers can negatively impact model performance by:



* Apply the PPS (Predictive Power Score) to find and discuss the relationships between features. Compare its findings with the correlation matrix.





**Predictive Power Score (PPS) Matrix**

**PPS is better suited for non-linear relationships** than correlation.

**Findings:**

* capital\_gain has a high PPS with income\_>50K, suggesting it is a strong predictor.
* education\_num also has a high PPS with income\_>50K, indicating its importance.
* Many features show **0 PPS**, meaning they do not strongly predict other variables.

**Correlation Matrix**

**Measures only linear relationships** between numerical features.

**Findings:**

* education\_num has a positive correlation with income\_>50K, confirming it as an important feature.
* age has a moderate positive correlation with income\_>50K, indicating older individuals may earn more.
* capital\_gain and capital\_loss have strong positive correlations with income\_>50K, reinforcing their significance.
* Some categorical features, when encoded, show weaker correlations.
* PPS captures both linear and non-linear relationships, making it useful for feature selection.
* Correlation is limited to linear relationships, missing non-linear predictive power.
* Features like capital\_gain, education\_num, and age are strong predictors, supported by both matrices.
* Some features have weak or no correlation but may still be predictive (PPS captures this better).