Preprocessing Datasets for ENCRYPT PPTs

Angelos Papoutsis

CERTH

General View

	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL
1	LASTPAYA	COLLATER	BASE_INTE	CONTRIBU	REPAY_FR	NEXT_INS	ACCOUNT	TOTAL_IN	STATEMEN	REMAIN_I	APPROVE	PENALTY_	TOTAL_BA	TOTAL_OU	OVERDUE_	OVERDUE_	OVERDUE_	PENALTY_	CBS_SDAT	OVERDUE_	LENT_AMO	DF_max_p	DF_min_p	DF_avg_p	DF_std_pa	DF_max_p	DF_min_p	DF_avg_p	DF_std_pa	DF_max_p	DF_min_p	DF_avg_p	DF_std_pa
2	78900	0	725	60	6	0	1	0	1.7E+08	0	3.7E+08	250	1.7E+08	1.8E+08	620000	1217	1.6E+08	1.7E+07	22/11/20	1.6E+08	2.3E+08	2942400	0	1814043	1192569	200000	0	47588.9	48486.8	4120000	0	2817265	1783998
3	5478400	0	1150	60	0	0	1	0	3.3E+07	0	3.5E+08	0	3.3E+07	6.5E+07	3.3E+07	0	0	0	22/11/20	0	3.3E+07	5478400	5478400	5478400		0	0	0		5478400	5478400	5478400	
4	1868000	0	1150	60	0	0	1	0	4588000	0	7.5E+07	0	4580000	9170000	4588000	0	0	0	22/11/20	0	4588000	1868000	1868000	1868000		0	0	0		1868000	1868000	1868000	
5	384400	0	1150	60	0	0	1	0	3.8E+07	0	4E+08	0	3.8E+07	7.7E+07	3.8E+07	0	0	0	22/11/20	0	3.8E+07	710000	384400	414000	98172.1	0	0	0	0	710000	384400	414000	98172.1
6	1580200	0	1150	60	0	0	1	0	1.6E+08	0	1.5E+09	0	1.6E+08	3.2E+08	1.6E+08	0	0	0	22/11/20	0	1.6E+08	1580200	1580200	1580200		0	0	0		1580200	1580200	1580200	
7	3132000	0	1150	60	0	0	1	0	6.4E+07	0	7.5E+07	0	6.4E+07	6.6E+07	2100000	0	6.2E+07	0	22/11/20	6.2E+07	9.4E+07	3132000	3132000	3132000		0	0	0		3132000	3132000	3132000	
8	5162600	0	1150	60	0	0	1	0	2.8E+07	0	1.6E+09	0	2.8E+07	5.6E+07	2.8E+07	0	0	0	22/11/20	0	2.8E+07	5162600	5162600	5162600		0	0	0		5162600	5162600	5162600	
9	677700	0	1150	60	0	0	1	0	1.2E+07	0	3.5E+08	0	1.2E+07	2.4E+07	1.2E+07	0	0	0	22/11/20	0	1.2E+07	1000000	677700	953957	121818	0	0	0	0	1000000	677700	953957	121818
10	200000	0	1150	60	0	0	1	0	3E+07	0	5E+08	0	3E+07	5.9E+07	3E+07	0	0	0	22/11/20	0	3E+07	200000	200000	200000	0	0	0	0	0	200000	200000	200000	0
11	52000	0	1150	60	0	0	1	0	1.6E+07	0	6E+07	0	1.6E+07	3.1E+07	1.6E+07	0	0	0	22/11/20	0	1.6E+07	52000	52000	52000	0	0	0	0	0	52000	52000	52000	0
12	30000	0	1150	60	0	0	1	0	366500	0	7E+07	0	360000	730000	366500	0	0	0	22/11/20	0	366500	1.8E+07	5000	842509	3889196	0	0	0	0	1.8E+07	5000	842509	3889196
13	13500	0	1150	60	0	0	1	0	1.1E+07	0	8E+07	0	1.1E+07	2.3E+07	1.1E+07	0	0	0	22/11/20	0	1.1E+07	80100	7300	20000	21385.5	0	0	0	0	80100	7300	20000	21385.5
14	2.7E+08	0	1150	60	0	0	1	0	4.8E+07	0	5.9E+08	0	4.8E+07	9.7E+07	4.8E+07	0	0	0	22/11/20	0	4.8E+07	2.7E+08	2.7E+08	2.7E+08		0	0	0		2.7E+08	2.7E+08	2.7E+08	
15	446100	0	1150	60	0	0	1	0	5280100	0	7.3E+08	0	5280000	1.1E+07	5280100	0	0	0	22/11/20	0	5280100	446100	446100	446100		0	0	0		446100	446100	446100	
16	9441600	0	1150	60	0	0	1	0	9441600	0	2E+09	0	9440000	1.9E+07	9441600	0	0	0	22/11/20	0	9441600	9441600	372000	4751800	2908964	0	0	0	0	9441600	372000	4751800	2908964
17	1000000	0	1150	60	0	0	1	0	9480200	0	1.5E+08	0	9480000	1.9E+07	9480200	0	0	0	22/11/20	0	9480200	1000000	47300	541007	445469	0	0	0	0	1000000	47300	541007	445469
18	2500000	0	1150	60	0	0	1	0	7.4E+07	0	3E+09	0	8.8E+07	1.6E+08	7.4E+07	0	0	0	22/11/20	0	7.4E+07	2500000	2500000	2500000		0	0	0		2500000	2500000	2500000	
19	1500000	0	1150	60	0	0	1	0	2.6E+09	0	2.5E+09	0	2.6E+09	2.7E+09	2.9E+07	3022	2.5E+09	0	22/11/20	2.5E+09	7E+09	1500000	1000000	1491803	64018.4	0	0	0	0	1500000	1000000	1491803	64018.4
20	350200	0	1150	60	0	0	1	0	6.7E+08	0	6.5E+08	0	6.7E+08	6.9E+08	2.4E+07	0	6.4E+08	0	22/11/20	6.4E+08	1.5E+09	350200	350200	350200		0	0	0		350200	350200	350200	
21	459000	0	1150	60	0	0	1	0	8.9E+07	0	3E+09	0	9E+07	1.8E+08	8.9E+07	0	0	0	22/11/20	0	8.9E+07	459000	459000	459000		0	0	0		459000	459000	459000	
22	1000000	0	1150	60	0	0	1	0	2.4E+08	0	1.5E+08	0	2.4E+08	3.1E+08	3700000	6240	1.8E+08	0	22/11/20	1.8E+08	3.7E+08	1000000	136200	427668	402430	0	0	0	0	1000000	136200	427668	402430
23	1000000	0	1150	60	0	0	1	0	2138100	0	1E+08	0	1.1E+07	1.4E+07	2138100	0	0	0	22/11/20	0	2138100	1000000	1000000	1000000		0	0	0		1000000	1000000	1000000	
24	2000000	0	1150	60	0	0	1	0	8.5E+08	0	1E+09	0	8.5E+08	8.6E+08	1.1E+07	0	8.3E+08	0	22/11/20	8.3E+08	1.7E+09	2000000	2000000	2000000	0	0	0	0	0	2000000	2000000	2000000	0
25	6607500	0	1150	60	0	0	1	0	8.8E+08	0	7.9E+08	0	8.8E+08	9E+08	2.4E+07	0	8.5E+08	0	22/11/20	8.5E+08	1.7E+09	6607500	6607500	6607500		0	0	0		6607500	6607500	6607500	
26	2101000	0	1150	60	0	0	1	0	9081800	0	7.9E+08	0	2.2E+07	3.1E+07	9081800	0	0	0	22/11/20	0	9081800	2101000	2101000	2101000		0	0	0		2101000	2101000	2101000	
27	100	0	1150	60	0	0	1	0	5.3E+08	0	5E+08	0	5.3E+08	5.5E+08	2.4E+07	0	5.1E+08	0	22/11/20	5.1E+08	1.6E+09	1200	100	480	443.847	0	0	0	0	1200	100	480	443.847
28	3549000	0	1150	60	0	0	1	0	3959000	0	5E+08	0	3950000	7910000	3959000	0	0	0	22/11/20	0	3959000	3549000	-410000	582417	962076	0	0	0	0	3549000	-410000	582417	962076
29	308100	0	1150	60	0	0	1	0	1.1E+07	0	1.1E+09	0	1.1E+07	2.2E+07	1.1E+07	0	0	0	22/11/20	0	1.1E+07	5000000	282800	1181148	1384503	0	0	0	0	5000000	282800	1181148	1384503
30	3.1E+07	0	1150	60	0	0	1	0	1.4E+07	0	7E+08	0	1.6E+07	3E+07	1.4E+07	0	0	0	22/11/20	0	1.4E+07	3.1E+07	3.1E+07	3.1E+07		0	0	0		3.1E+07	3.1E+07	3.1E+07	
31	5940300	0	1150	60	0	0	1	0	3.1E+09	0	3.2E+09	0	3.1E+09	3.2E+09	6.8E+07	0	3.1E+09	0	22/11/20	3.1E+09	6E+09	3385600	2554700	2970150	423074	0	0	0	0	3385600	2554700	2970150	423074
32	351500	0	400	12	6	0	1	0	1.8E+08	0	8.5E+08	250	9.1E+08	9.6E+08	3187000	3899	1.4E+08	0	22/11/20	1.4E+08	6.4E+08	1500000	0	130959	313616	375000	0	45359.1	105401	1500000	0	483367	466701
33	1291500	0	1150	60	0	0	1	0	1.2E+07	0	7.9E+08	0	1.3E+07	2.6E+07	1.2E+07	0	0	0	22/11/20	0	1.2E+07	1291500	1291500	1291500		0	0	0		1291500	1291500	1291500	
34	3000000	0	1150	60	0	0	1	0	1.7E+08	0	2.6E+09	0	1.7E+08	3.4E+08	1.7E+08	0	0	0	22/11/20	0	1.7E+08	3000000	3000000	3000000		0	0	0		3000000	3000000	3000000	
35	293400	0	1150	60	0	0	1	0	1500900	0	1.6E+08	0	1500000	3000000	1500900	0	0	0	22/11/20	0	1500900	1000000	293400	695900	363432	0	0	0	0	1000000	293400	695900	363432
36	2498000	0	0	0	0	0	1	0	0	0	5.5E+08	0	8.5E+07	8.5E+07	0	0	0	0	22/11/20	0	5.6E+07	2500000	0	138833	580544	0	0	0	0	2500000	0	138833	580544
37	1148000	0	1150	60	0	0	1	0	1.1E+07	0	5.8E+08	0	1.1E+07	2.2E+07	1.1E+07	0	0	0	22/11/20	0	1.1E+07	6000000	800000	2274000	2489411	0	0	0	0	6000000	800000	2274000	2489411
38	700000	0	1150	60	0	0	1	0	1.5E+07	0	2.1E+08	0	1.5E+07	2.9E+07	1.5E+07	0	0	0	22/11/20	0	1.5E+07	1450000	700000	1052941	385872	0	0	0	0	1450000	700000	1052941	385872

encrypt

Characteristics of the Dataset

```
RangeIndex: 4177 entries, 0 to 4176
Data columns (total 39 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   CASE_ID                         4177 non-null   int64
 1   INTEREST                        4177 non-null   int64
 2   LENTAMOUNT                      4177 non-null   float64
 3   LENTINITAMOUNT                  4177 non-null   float64
 4   LASTPAYDATE                     4176 non-null   object
 5   LASTPAYAMOUNT                   4176 non-null   float64
 6   COLLATERAL_AMOUNT               4177 non-null   int64
 7   BASE_INTEREST_RATE              4177 non-null   int64
 8   CONTRIBUTION_INTEREST_RATE      4177 non-null   int64
 9   REPAY_FREQUENCY                 4177 non-null   int64
 10  NEXT_INSTALLMENT_AMOUNT         4177 non-null   int64
 11  ACCOUNTSTATUS                   4177 non-null   object
 12  TOTAL_INST                      4177 non-null   int64
 13  STATEMENT_OVERDUE_AMOUNT        4177 non-null   int64
 14  REMAIN_INST                     4177 non-null   int64
 15  APPROVED_CAPITAL                4177 non-null   int64
 16  PENALTY_INTEREST_RATE           4177 non-null   int64
 17  TOTAL_BALANCE_AMOUNT_ACY        4177 non-null   int64
 18  TOTAL_OUTSTANDING_AMOUNT_ACY    4177 non-null   int64
 19  OVERDUE_EXPENSES_AMOUNT_ACY     4177 non-null   int64
 20  OVERDUE_INTEREST_AMOUNT_ACY     4177 non-null   int64
 21  OVERDUE_CAPITAL_AMOUNT_ACY      4177 non-null   int64
 22  PENALTY_INTEREST_AMT            4177 non-null   int64
 23  CBS_SDATE                       4177 non-null   object
 24  OVERDUE_CAPITAL                 4177 non-null   int64
 25  LENT_AMOUNT_BCY                 4177 non-null   int64
 26  DF_max_pay_amount_ever          4177 non-null   int64
 27  DF_min_pay_amount_ever          4177 non-null   int64
 28  DF_avg_pay_amount_ever          4177 non-null   float64
 29  DF_std_pay_amount_ever          3585 non-null   float64
 30  DF_max_pay_expenses_ever        4177 non-null   int64
 31  DF_min_pay_expenses_ever        4177 non-null   int64
 32  DF_avg_pay_expenses_ever        4177 non-null   float64
 33  DF_std_pay_expenses_ever        3585 non-null   float64
 34  DF_max_pay_amount_bcy_ever      4177 non-null   int64
 35  DF_min_pay_amount_bcy_ever      4177 non-null   int64
 36  DF_avg_pay_amount_bcy_ever      4177 non-null   float64
 37  DF_std_pay_amount_bcy_ever      3585 non-null   float64
 38  DEFAULT                         4177 non-null   int64
dtypes: float64(9), int64(27), object(3)
```

encrypt

Checking for NaN

Removing columns that are entirely NaN (missing values) is essential for machine learning preprocessing.

- **These columns:**
  - Contain no useful data
  - Offer no predictive value
  - Result in no information gain

- **Empty columns may cause:**
  - Errors during model training
  - Unpredictable behavior in the model

- **Practical benefits of removing these columns:**
  - Keeps the dataset clean and efficient
  - Reduces memory usage
  - Speeds up data processing
  - Simplifies data inspection and visualization

- **Overall impact:**
  - Improves data quality
  - Ensures the model is trained on a clean, consistent dataset
  - Leads to better performance and more reliable results

Checking for Duplicates

- **Removes duplicate rows, ensuring:**
  - Each data point is unique
  - No data is overrepresented, which could skew model learning
  - Reduces the risk of overfitting

- **Resets row indices after removing duplicates:**
  - Maintains a clean, continuous sequence
  - Improves consistency and interpretability

- **Benefits for machine learning:**
  - Enhances the model's ability to generalize
  - Ensures the dataset used for training and evaluation is accurate and representative

Converting Special Characters to Numerical

Converting ratio-like string values (e.g., '3/5') into decimal numbers is a valuable preprocessing step in machine learning.

- **This transformation:**
  - Converts human-readable ratios into numeric features
  - Makes the data model-ready for machine learning algorithms
- **Benefits:**
  - Ensures cleaner, consistent columns
  - Improves model interpretability and performance
  - Eliminates noise from mixed data types
- **Handles data quality issues by filtering out:**
  - Invalid entries (e.g., 'n/a', 'abc/3')
  - Division by zero cases (e.g., '4/0')
- **Post-conversion advantages:**
  - Enables scaling, correlation analysis, and clustering
  - Simplifies downstream processing

Dealing with Missing Values

- Handling Missing Values in a DataFrame is crucial for machine learning to maintain data integrity.

- Filling Strategy:
  - Categorical columns → Missing values are replaced with the mode (most frequent value) to ensure category consistency.
  - Integer columns → Gaps are filled with the mean, rounded to preserve the integer type.
  - Floating-point columns → Missing values are replaced with the mean for numerical continuity.

- Importance:
  - Prevents errors during model training by ensuring a complete dataset.
  - Preserves the original structure and distribution of data.
  - Enhances model performance by retaining useful information while avoiding bias from improper imputation.

Handling Various Types of Data

Converting Strings to Numeric Values for Machine Learning

- **Why Convert Strings to Numbers?**
  - Machine learning models require numerical inputs—text-based data must be transformed into numbers for effective training
  - Mixed data types (e.g., numbers stored as strings) can lead to errors or inefficient processing

- **Key Steps in the Conversion Process:**

1. Remove Non-Informative Identifiers → Columns like IDs that don't contribute to model learning are dropped to prevent overfitting

2. Extract Numeric Values from Date Fields → Convert datetime columns into separate year, month, and day features for better temporal analysis

3. Transform Numeric Strings into Proper Formats:
   - Strings that represent numbers (e.g., "1000") are converted into integers
   - Special formats like hyphenated numbers (e.g., "1-200") are cleaned and transformed into floats

4. Encoding Categorical Features:
   - High-cardinality categorical columns are binary encoded to reduce dimensionality
   - Low-cardinality categorical columns are one-hot encoded for structured numerical representation

Feature Scaling (1/2)

Feature scaling is the process of standardizing the range of features in your dataset. It transforms the values so they fall within a specific range (typically 0-1 for normalization or mean 0 and standard deviation 1 for standardization).

Why it's important for machine learning:

- **Prevents dominance of large-scale features:** Without scaling, features with larger numerical ranges (like income in thousands) would have more influence than features with smaller ranges (like age in years).
- **Improves convergence speed:** For gradient-based algorithms like neural networks, unscaled features can cause the optimization process to oscillate and take longer to find optimal weights.
- **Essential for distance-based algorithms:** Methods like K-nearest neighbors, K-means clustering, and Support Vector Machines rely on calculating distances between data points, which becomes meaningless when features have vastly different scales.
- **Algorithm stability:** Many algorithms are numerically unstable without scaling, leading to overflow/underflow issues.

Feature Scaling (2/2)

Within feature scaling, there are several specific methods:

Normalization (Min-Max Scaling):

Scales features to a fixed range, typically [0,1]

Standardization (Z-score Normalization):

Scales features to have mean=0 and standard deviation=1

Robust Scaling:

Uses the median and interquartile range instead of mean/std

Less affected by outliers

REPAY_FR	NEXT_INS	TOTAL_IN	STATEMEN	REMAIN_I	APPROVE	PENALTY_	TOTAL_BA	TOTAL_OU	OVERDUE_	OVERDUE_	OVERDUE_	PENALTY_	CBS_SDAT	OVERDUE_	LENT_AM	DF_max_p	DF_min_p	DF_avg_p	DF_std_p	DF_max_p	DF_min_p	DF_avg_p	DF_std_p	DF_max_p	DF_min_p	DF_avg_p	DF_std_p	DEFAULT	ACCOUNTSTATU	ACCOUNTSTATUS_EGG_CB
6	0	0	1.7E+08	0	3.7E+08	250	1.7E+08	1.8E+08	620000	1217	1.6E+08	1.7E+07	2	1.6E+08	2.3E+08	2942400	0	1814043	1192569	200000	0	47588.9	48486.8	4120000	0	2817265	1783998	1	1	0
0	0	0	3.3E+07	0	3.5E+08	0	3.3E+07	6.5E+07	3.3E+07	0	0	0	2	0	3.3E+07	5478400	5478400	5478400	1.7E+07	0	0	0	13956.6	5478400	5478400	5478400	1.8E+07	1	1	0
0	0	0	4588000	0	7.5E+07	0	4580000	9170000	4588000	0	0	0	2	0	4588000	1868000	1868000	1868000	1.7E+07	0	0	0	13956.6	1868000	1868000	1868000	1.8E+07	1	1	0
0	0	0	3.8E+07	0	4E+08	0	3.8E+07	7.7E+07	3.8E+07	0	0	0	2	0	3.8E+07	710000	384400	414000	98172.1	0	0	0	0	710000	384400	414000	98172.1	1	1	0
0	0	0	1.6E+08	0	1.5E+09	0	1.6E+08	3.2E+08	1.6E+08	0	0	0	2	0	1.6E+08	1580200	1580200	1580200	1.7E+07	0	0	0	13956.6	1580200	1580200	1580200	1.8E+07	1	1	0
0	0	0	6.4E+07	0	7.5E+07	0	6.4E+07	6.6E+07	2100000	0	6.2E+07	0	2	6.2E+07	9.4E+07	3132000	3132000	3132000	1.7E+07	0	0	0	13956.6	3132000	3132000	3132000	1.8E+07	1	1	0
0	0	0	2.8E+07	0	1.6E+09	0	2.8E+07	5.6E+07	2.8E+07	0	0	0	2	0	2.8E+07	5162600	5162600	5162600	1.7E+07	0	0	0	13956.6	5162600	5162600	5162600	1.8E+07	1	1	0
0	0	0	1.2E+07	0	3.5E+08	0	1.2E+07	2.4E+07	1.2E+07	0	0	0	2	0	1.2E+07	1000000	677700	953957	121818	0	0	0	0	1000000	677700	953957	121818	1	1	0
0	0	0	3E+07	0	5E+08	0	3E+07	5.9E+07	3E+07	0	0	0	2	0	3E+07	200000	200000	200000	0	0	0	0	0	200000	200000	200000	0	1	1	0
0	0	0	1.6E+07	0	6E+07	0	1.6E+07	3.1E+07	1.6E+07	0	0	0	2	0	1.6E+07	52000	52000	52000	0	0	0	0	0	52000	52000	52000	0	1	1	0
0	0	0	366500	0	7E+07	0	360000	730000	366500	0	0	0	2	0	366500	1.8E+07	5000	842509	3889196	0	0	0	0	1.8E+07	5000	842509	3889196	1	1	0
0	0	0	1.1E+07	0	8E+07	0	1.1E+07	2.3E+07	1.1E+07	0	0	0	2	0	1.1E+07	80100	7300	20000	21385.5	0	0	0	0	80100	7300	20000	21385.5	1	1	0
0	0	0	4.8E+07	0	5.9E+08	0	4.8E+07	9.7E+07	4.8E+07	0	0	0	2	0	4.8E+07	2.7E+08	2.7E+08	2.7E+08	1.7E+07	0	0	0	13956.6	2.7E+08	2.7E+08	2.7E+08	1.8E+07	1	1	0
0	0	0	5280100	0	7.3E+08	0	5280000	1.1E+07	5280100	0	0	0	2	0	5280100	446100	446100	446100	1.7E+07	0	0	0	13956.6	446100	446100	446100	1.8E+07	1	1	0
0	0	0	9441600	0	2E+09	0	9440000	1.9E+07	9441600	0	0	0	2	0	9441600	9441600	372000	4751800	2908964	0	0	0	0	9441600	372000	4751800	2908964	1	1	0
0	0	0	9480200	0	1.5E+08	0	9480000	1.9E+07	9480200	0	0	0	2	0	9480200	1000000	47300	541007	445469	0	0	0	0	1000000	47300	541007	445469	1	1	0
0	0	0	7.4E+07	0	3E+09	0	8.8E+07	1.6E+08	7.4E+07	0	0	0	2	0	7.4E+07	2500000	2500000	2500000	1.7E+07	0	0	0	13956.6	2500000	2500000	2500000	1.8E+07	1	1	0
0	0	0	2.6E+09	0	2.5E+09	0	2.6E+09	2.7E+09	2.9E+07	3022	2.5E+09	0	2	2.5E+09	7E+09	1500000	1000000	1491803	64018.4	0	0	0	0	1500000	1000000	1491803	64018.4	1	1	0
0	0	0	6.7E+08	0	6.5E+08	0	6.7E+08	6.9E+08	2.4E+07	0	6.4E+08	0	2	6.4E+08	1.5E+09	350200	350200	350200	1.7E+07	0	0	0	13956.6	350200	350200	350200	1.8E+07	1	1	0
0	0	0	8.9E+07	0	3E+09	0	9E+07	1.8E+08	8.9E+07	0	0	0	2	0	8.9E+07	459000	459000	459000	1.7E+07	0	0	0	13956.6	459000	459000	459000	1.8E+07	1	1	0
0	0	0	2.4E+08	0	1.5E+08	0	2.4E+08	3.1E+08	3700000	6240	1.8E+08	0	2	1.8E+08	3.7E+08	1000000	136200	427668	402430	0	0	0	0	1000000	136200	427668	402430	1	1	0
0	0	0	2138100	0	1E+08	0	1.1E+07	1.4E+07	2138100	0	0	0	2	0	2138100	1000000	1000000	1000000	1.7E+07	0	0	0	13956.6	1000000	1000000	1000000	1.8E+07	1	1	0
0	0	0	8.5E+08	0	1E+09	0	8.5E+08	8.6E+08	1.1E+07	0	8.3E+08	0	2	8.3E+08	1.7E+09	2000000	2000000	2000000	0	0	0	0	0	2000000	2000000	2000000	0	1	1	0
0	0	0	8.8E+08	0	7.9E+08	0	8.8E+08	9E+08	2.4E+07	0	8.5E+08	0	2	8.5E+08	1.7E+09	6607500	6607500	6607500	1.7E+07	0	0	0	13956.6	6607500	6607500	6607500	1.8E+07	1	1	0
0	0	0	9081800	0	7.9E+08	0	2.2E+07	3.1E+07	9081800	0	0	0	2	0	9081800	2101000	2101000	2101000	1.7E+07	0	0	0	13956.6	2101000	2101000	2101000	1.8E+07	1	1	0
0	0	0	5.3E+08	0	5E+08	0	5.3E+08	5.5E+08	2.4E+07	0	5.1E+08	0	2	5.1E+08	1.6E+09	1200	100	480	443.847	0	0	0	0	1200	100	480	443.847	1	1	0
0	0	0	3959000	0	5E+08	0	3950000	7910000	3959000	0	0	0	2	0	3959000	3549000	-410000	582417	962076	0	0	0	0	3549000	-410000	582417	962076	1	1	0
0	0	0	1.1E+07	0	1.1E+09	0	1.1E+07	2.2E+07	1.1E+07	0	0	0	2	0	1.1E+07	5000000	282800	1181148	1384503	0	0	0	0	5000000	282800	1181148	1384503	1	1	0
0	0	0	1.4E+07	0	7E+08	0	1.6E+07	3E+07	1.4E+07	0	0	0	2	0	1.4E+07	3.1E+07	3.1E+07	3.1E+07	1.7E+07	0	0	0	13956.6	3.1E+07	3.1E+07	3.1E+07	1.8E+07	1	1	0
0	0	0	3.1E+09	0	3.2E+09	0	3.1E+09	3.2E+09	6.8E+07	0	3.1E+09	0	2	3.1E+09	6E+09	3385600	2554700	2970150	423074	0	0	0	0	3385600	2554700	2970150	423074	1	1	0
6	0	0	1.8E+08	0	8.5E+08	250	9.1E+08	9.6E+08	3187000	3899	1.4E+08	0	2	1.4E+08	6.4E+08	1500000	0	130959	313616	375000	0	45359.1	105401	1500000	0	483367	466701	1	1	0
0	0	0	1.2E+07	0	7.9E+08	0	1.3E+07	2.6E+07	1.2E+07	0	0	0	2	0	1.2E+07	1291500	1291500	1291500	1.7E+07	0	0	0	13956.6	1291500	1291500	1291500	1.8E+07	1	1	0
0	0	0	1.7E+08	0	2.6E+09	0	1.7E+08	3.4E+08	1.7E+08	0	0	0	2	0	1.7E+08	3000000	3000000	3000000	1.7E+07	0	0	0	13956.6	3000000	3000000	3000000	1.8E+07	1	1	0
0	0	0	1500900	0	1.6E+08	0	1500000	3000000	1500900	0	0	0	2	0	1500900	1000000	293400	695900	363432	0	0	0	0	1000000	293400	695900	363432	1	1	0
0	0	0	0	0	5.5E+08	0	8.5E+07	8.5E+07	0	0	0	0	2	0	5.6E+07	2500000	0	138833	580544	0	0	0	0	2500000	0	138833	580544	1	1	0
0	0	0	1.1E+07	0	5.8E+08	0	1.1E+07	2.2E+07	1.1E+07	0	0	0	2	0	1.1E+07	6000000	800000	2274000	2489411	0	0	0	0	6000000	800000	2274000	2489411	1	1	0
0	0	0	1.5E+07	0	2.1E+08	0	1.5E+07	2.9E+07	1.5E+07	0	0	0	2	0	1.5E+07	1450000	700000	1052941	385872	0	0	0	0	1450000	700000	1052941	385872	1	1	0

Remove Correlated Features

•Why Remove Highly Correlated Features?

•When features are highly correlated, they contain overlapping information, which can:

- •Confuse models, leading to unstable coefficients (especially in linear models)
- •Inflate feature importance, making some variables seem more influential than they are
- •Increase overfitting risk, reducing the model's ability to generalize to new data

•Key Steps in Identifying and Removing Redundant Features:

- Compute Feature Correlations → Use correlation matrices to detect variables that are highly correlated (e.g., correlation > 0.9)
- Identify Redundant Features → If two features have a strong correlation, drop one to simplify the dataset
- Preserve Important Features → Ensure that the most informative and independent features remain

Benefits of This Approach:

- •Improves Model Interpretability → A cleaner dataset leads to easier analysis and better insights
- •Enhances Model Stability → Reducing feature redundancy results in more reliable coefficients
- •Prevents Overfitting → Eliminates unnecessary complexity that could hurt generalization
- •Speeds Up Training → Fewer features reduce computational costs and improve efficiency

Ground Preprocessing Techniques for Machine Learning (1/2)

1. Encoding Categorical Labels

- **Label Encoding**: Transforms categorical labels into numerical values (0, 1, 2, etc.)
- **When to use**: Appropriate for target variables in classification where no ordinal relationship is implied
- **Considerations**: Retains the ability to easily convert back to original classes

2. Handling Imbalanced Classes

- **Undersampling**: Reduces majority class samples to balance with minority classes
  - Random undersampling

- **Oversampling**: Increases minority class samples
  - Random oversampling
  - Synthetic sample generation (SMOTE, ADASYN)

Ground Preprocessing Techniques for Machine Learning (2/2)

- **Algorithm-level approaches**:
  - Class weighting (assign higher importance to minority classes)

- **Alternative evaluation metrics**:
  - Precision, Recall, F1-score

3. Train/Test/Validation Split Strategies

- **Stratified splitting**: Maintains class distribution across all splits
- 70-80% training, 10-15% validation, 10-15% testing

4. **Mutual Information Feature Selection**

- **Concept**: Measures dependency between features and target variable
- **Application**: Ranks features by information gain, enables selection of most informative features

Order of the Preprocessing Steps (1/2)

- **Missing value handling and general preprocessing**
  - Comes first because missing values can disrupt all subsequent analyses
  - Early handling prevents these gaps from cascading problems through the pipeline
  - Missing data must be addressed before any meaningful analysis can occur

- **Categorical encoding**
  - Follows immediately after cleaning because most algorithms require numerical input
  - Must precede scaling since categorical variables need to be converted to numbers first
  - Enables correlation analysis and other numerical techniques to work with all variables

- **Outlier treatment**
  - Positioned after basic cleaning but before scaling to prevent outliers from skewing scaling parameters
  - Outliers can drastically affect means and standard deviations used in scaling
  - Better to handle extreme values before they influence downstream transformations

Order of the Preprocessing Steps (2/2)

- **Feature scaling**
  - Follows Categorical enconding and nsures all features (original and transformed) are on comparable scales
  - Creates uniformity across features with different original units and ranges
  - Prevents features with larger magnitudes from dominating distance-based or gradient-based algorithms

- **Feature selection with correlation analysis**
  - Performed after scaling so that correlations are computed on properly scaled features
  - Correlation coefficients are more meaningful when features are on the same scale
  - Helps remove redundant information after all features are properly prepared

Data Leakage

Preventing Data Leakage

- **Definition**: When information from outside the training dataset is used to create the model

- **Common causes**:
  - **Preprocessing on the entire dataset**: Always fit preprocessors on training data only
  - **Train-test contamination**: Information bleeding between train and test sets

- **Prevention techniques**:
  - ✓ **Pipeline approach**: Encapsulate all preprocessing within cross-validation folds
  - ✓ **Data splitting first**: Split data before any analysis or preprocessing
  - ✓ **Holdout validation**: Keep a truly untouched test set for final evaluation

Essential Preprocessing Steps for PPTs

No.	Data Transformation Techniques Applied	Justification
1.	Handing of Missing values	Missing values pose challenges when manipulating a dataset. In our case, these challenges arise when applying the various PPTs.
2.	Deleting of Duplicates values	Duplicate values enlarge the dataset, thereby increasing the computational complexity of HE.
3.	Cleaning a dataset from not useful characters (e.g., "-")	Special characters are treated as strings, whereas Homomorphic Encryption (HE) can only be applied to numerical data. The presence of such characters necessitates the transformation of more columns, resulting in additional features (i.e., more dimensions). This, in turn, increases the computational complexity of HE.
4.	Replacing some values with some others. For example, ratio-like string values (e.g., 14/3) to decimal values	Same reason as described above.
5.	Performing data normalization	Normalization can enhance computational efficiency and accuracy, particularly for operations that are sensitive to the scale of data. Additionally, may also provide added privacy benefits
6.	Performing feature selection (i,e. Removing correlations)	The correlation between features can expose private information about the end-users in the context of DP. This necessitates the injection of additional noise to protect the end-user's data
7.	Converting Categorical Features to numerical	Machine Learning algorithms and PPT work on number rather on string

Thank you!

Angelos Papoutsis, apapoutsis@iti.gr

Stay in touch

https://encrypt-project.eu/

encrypt-project

@encrypt_project

encrypt