Appendix: Improving Predictive Model Accuracy via Systematic Data Wrangling Pattern Discovery and Application

A Comparison of Data Wrangling Approaches

Table 1 presents a comparative analysis of different data wrangling approaches.

B List of Dataset Characteristics

Table 2 displays the 61 dataset characteristics this study considers.

C Dataset Characteristic for Clusters and Patterns

C.1 Continuous

Table 3 displays the value range associated with each pattern for continuous dataset types.

C.2 Categorical

Table 4 displays the value range associated with each pattern for categorical dataset types.

C.3 Hybrid

Tables 5 and 6 show the value range associated with each pattern for hybrid dataset types.

Table 1: Comparison of Various Data Wrangling Approaches

Approach	Pandas	Wrangler	Learn2Clean	ActiveClean	Auto-sklearn	AutoWeka	PRESISTANT	AI Assistants	VADA
User Role	Manual	Semi-automation	Complete-Automation	Semi-Automation	Complete-Automation	Complete-Automation	Semi-Automation	Semi-Automation	Semi-Automation
UI	Code-based	GUI	Code-based	Code-based	Code-based	GUI	GUI	Various	Web Interface
Scope	Individual operation	Individual operation	Entire recipe	Partial Recipe	Entire recipe	Entire recipe	Entire recipe	Individual operation	Partial Recipe
Input	High	Medium	Low	Medium	Low	Low	Medium	Low	Medium
Decision	N/A	Heuristics	Q-Learning	Rule-based	Bayesian Optimization	Bayesian Optimization	Meta-Learning	Optimi-sation-based ranking algorithm	N/A
Learning	N/A	Heuristics	Reinforce Learn	Stochastic Gradient Descent (SGD)	Meta-Learning	Random Forest	Random Forest, Rank-based	Feedback from user	Rank-based (Heuristics)
Pre-Processing	Data Integration, Data Cleaning, Transform	Data Cleaning, Transform	Normalisation, Feature selection, Imputation, Outlier, Deduplication, Consistency	Data Cleaning	OHE, Imputation Rescaling, Balancing	Feature selection	Data Cleaning, Transformation	Merge, parse, infer column types	Schema Matching
Support	Interactive	Interactive	Automatic	Automatic	Automatic	Automatic	Interactive	Automatic	Interactive
Analysis Impact	No	No	Yes	Yes	Yes	Yes	Yes	No	No
Improve Quality/ Accuracy/ Cost?	N/A	Data Quality	Data Quality, Accuracy	Data Quality, Accuracy	Accuracy	Accuracy, Costs	Accuracy, Costs	N/A	Data Quality
Pros	Customizable	GUI	Determine the optimal sequence of operators.	Focus on cleaning impactful records for statistical analysis.	Focus on hyper-parameter optimization.	Efficient hyperparameter tuning.	Specialised Attention to Pre-processing.	Specialised in individual data wrangling operation.	Accounts for user feedback and priorities.
Cons	Need high skills.	No Impact Analysis.	The pipeline is preset and users can't change it.	The user chooses the pre-processing steps, assuming they're an expert.	Limited data preprocessing support, possible overfitting	No data preprocessing support.	Ranks data wrangling tools by how much they affect the analysis.	Addresses single data wrangling issues, not the entire process.	No impact analysis.

Table 2: List of Dataset Characteristics

No	Characteristic	Type
1..2	Min[Means—Std—Kurtosis—Skewness] of Cont. Att.	Continuous
3..6	Min[Means—Std—Kurtosis—Skewness] of Cont. Att.	Continuous
7..10	Mean[Means—Std—Kurtosis—Skewness] of Cont. Att.	Continuous
11..14	Max[Means—Std—Kurtosis—Skewness] of Cont. Att.	Continuous
15..17	Quartile [1—2—3] of Means of Continuous Attributes	Continuous
18..20	Quartile [1—2—3] of Std of Continuous Attributes	Continuous
21.23	Quartile [1—2—3] of Kurtosis of Cont. Att.	Continuous
24..26	Quartile [1—2—3] of Skewness of Cont. Att.	Continuous
27	Number of Categorical Attributes	Categorical
28	Number of Binary Attributes	Categorical
29	Percentage of Categorical Attributes	Categorical
30	Percentage of Binary Attributes	Categorical
31..33	[Min—Mean—Max] Attribute Entropy	Categorical
34..36	Quartile [1—2—3] Attribute Entropy	Categorical
37..39	[Min—Mean—Max] Mutual Information	Categorical
40..42	Quartile [1—2—3] Mutual Information	Categorical
43	Equivalent Number of Attributes	Categorical
44	Noise to Signal Ratio	Categorical
45..48	[Min—Mean—Max—Std] Attribute Distinct Values	Categorical
49	Number of Instances	Generic
50	Number of Attributes	Generic
51	Dimensionality	Generic
52,53	[Number—Percentage] of Missing Values	Generic
54,55	[Number—Percentage] of Instances with Miss. Vals.	Generic
56	Number of Classes	Generic
57	Class Entropy	Generic
58,59	[Minority—Majority] Class Size	Generic
60,61	[Minority—Majority] Class Percentage	Generic

Table 3: Range of Values for Each Characteristic in Continuous Dataset

Pattern ID / Characteristics	**Cont 0**		**Cont 1**		**Cont 2**		**Cont 3/Cont 3 (MV)**	
	Min	**Max**	**Min**	**Max**	**Min**	**Max**	**Min**	**Max**
Class Entropy	0.425	0.692	0.15	0.838	0.111	1	0.420191	1
Dimensionality	0.035	0.206	0.029	0.15	0.001	0.474	0.003644	0.348754
Number of Features	38	40	7	73	3	1777	3	98
Number of Instances	194	1077	209	2534	3751	7400	50	1372
Max Means of Numeric Atts	30965.46	49208.27	10164.2	18542.99	0	631.887	0.004162	1378.676
Min Means of Numeric Atts	0.075	0.085	-10.511	4.699	-0.022	72.358	-52.936	24.77778
Quartile 1 Means of Numeric Atts	0.954	2.602	1.578	14.876	-0.006	74.663	-30.279	35.84604
Max Skewness of Numeric Atts	6.869	28.541	4.027	18.367	0.052	64.397	-0.226883	26.84598
Min Skewness of Numeric Atts	-1.297	0.038	-2.666	2.141	-31.461	1.592	-2.532392	0.74945
Percentage of Instances With Missing Values	0	0	0	0	0	0	0	55.55556
Min Kurtosis of Numeric Atts	-1.212	-0.573	-1.037	5.902	-2.001	5.257	-2.039553	1.524249
Min Std Dev of Numeric Atts	0.04	0.058	0.074	6.816	0	11.659	0	7.558468
Quartile 1 Std Dev of Numeric Atts	0.847	2.929	1.166	21.202	0.045	12.346	0.009841	16.01825

Table 4: Range of Values for Each Characteristic in Categorical Dataset

Pattern ID / Characteristics	**Cat 0**		**Cat 1**	
	Min	**Max**	**Min**	**Max**
Class Entropy	0.808	1	0.319	1
Equivalent Number Of Atts	4.073	22.564	2.65	68.602
Max Attribute Entropy	1	3.03	0.996	11.258
Mean Attribute Entropy	0.761	1.409	0.59	6.906
Min Attribute Entropy	0	0.929	0.004	3.003
Dimensionality	0.003	3.3	0	0.286
Number Of Features	7	33	4	37
Max Mutual Information	0.176	1	0.025	0.902
Min Mutual Information	0	0.026	0	0.046
Quartile 1 Mutual Information	0.004	0.109	0	0.079
Percentage Of Instances With Missing Values	30.527	70	0	17.716
Min Nominal Att Distinct Values	1	2	2	2

Table 5: Range of Values for Each Characteristic in Hybrid Dataset (1/2)

Pattern ID / Characteristics	Hb 0		Hb 1		Hb 2		Hb 3		Hb 4		Hb 5	
	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max
Class Entropy	0.454	0.984	0.476	0.974	0.332	1	0.881	0.998	0.943	0.995	0.373	1
Equivalent Number of Atts	13.055	392.122	4.106	124.064	14.432	64.929	1.441	43.593	7.574	9.338	4.919	108.944
Max Attribute Entropy	1.368	9.723	1.128	4.585	0.763	1.521	2.667	8.11	1.683	1.739	0.908	6.57
Min Attribute Entropy	0	2.807	0.11	4.459	0	0.613	0.228	1.592	0.391	0.606	0.027	3.527
Dimensionality	0	0.003	0.02	0.127	0.008	0.125	0.014	0.022	0.046	0.048	0.013	0.298
Number of Features	9	119	9	26	30	300	8	21	14	14	4	32
Number of Instances	31406	45312	205	528	2407	3772	398	1161	294	303	57	368
Max Means of Numeric Atts	0.501	1978.971	10734.18	13207.13	0.722	110.47	2970.425	3271.258	246.264	250.849	31.189	150.248
Min Means of Numeric Atts	0.003	6.995	0.136	3.255	-0.561	0.995	1.155	15.568	0	1.04	0.217	60.436
Max Mutual Information	0.003	0.277	0.007	0.485	0.06	0.233	0.095	0.954	0.208	0.275	0.008	0.559
Min Mutual Information	0	0.003	0.001	0.114	0	0.025	0.001	0.37	0	0.001	0	0.042
Quartile 1 Mutual Information	0.002	0.014	0.007	0.114	0	0.034	0.005	0.37	0.01	0.033	0	0.042
Percentage of Binary Features	2.521	23.529	4.545	19.231	2	94.483	6.25	14.286	21.429	28.571	10	85.714
Percentage of Numeric Features	34.783	77.778	57.692	86.364	5.517	98	33.333	81.25	35.714	42.857	9.091	80
Percentage of Symbolic Features	22.222	65.217	13.636	42.308	2	94.483	18.75	66.667	57.143	64.286	20	90.909

Table 6: Range of Values for Each Characteristic in Hybrid Datasets (2/2)

Pattern ID / Characteristics	Hb 0		Hb 1		Hb 2		Hb 3		Hb 4		Hb 5	
	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max
Percentage of Instances With Missing Values	0	54.779	0	50	0	100	0	7.494	0	99.66	0	98.246
Max Skewness of Numeric Atts	1.96	78.687	2.611	11.152	7.314	13.883	1.034	5.815	1.27	1.549	-0.147	4.151
Min Skewness of Numeric Atts	-65.615	0.093	-1.514	-0.683	-0.846	1.233	-0.531	0.344	-0.537	-0.284	-2.007	0.171
Quartile 1 Skewness of Numeric Atts	-4.389	0.685	0.044	2.561	0.05	1.259	-0.273	0.433	-0.37	-0.186	-0.975	0.219
Quartile 2 Skewness of Numeric Atts	-2.504	3.144	0.664	3.632	1.342	8.657	0.481	1.094	0.714	0.929	-0.355	1.565
Min Kurtosis of Numeric Atts	-1.841	-1.06	-1.957	-0.1	-0.752	4.073	-1.381	-0.03	-0.59	-0.523	-2.035	0.278
Quartile 1 Kurtosis of Numeric Atts	-1.201	0.954	-0.37	11.536	0.26	5.982	-1.21	0.212	-0.546	-0.171	-1.547	0.655
Min Std Dev of Numeric Atts	0.01	6.184	0.153	0.774	0.067	0.376	0.362	19.514	0	1.161	0.066	20.655