

Project Report

Team #12

WANG Zeyu

YANG Xirui

Wu Tianxiao

April 7, 2025

Contents

1 Problem definition

2 Method

- Data Prepare
- Data Analysis

3 Solution

- Solution - Task 1
- Solution - Task 2
- Solution - Task 3

4 Summary

Problem definition

- **Task 1:** Train models based on 10 features, including academic metrics, aptitude scores, and soft skill ratings, to predict whether a student will be successfully placed in a job.
- **Task 2:** Train models based on 18 anonymity features and labeled with 5 classes, then predict the labels based on the training set.
- **Task 3:** Train models based on 28 anonymity numerical with the transaction amount, to indicate whether it is a fraudulent transaction or a legitimate transaction.

Correlation coefficient^{[1][2]}

- **Pearson correlation coefficient:** Measures linear correlation between two sets of data;
- **Kendall correlation coefficient:** Measures the rank correlation by counting the concordant pairs;
- **Spearman correlation coefficient:** Measures the rank correlation based on the Pearson correlation coefficient;

^[1]Hervé Abdi (2007). "The Kendall rank correlation coefficient". In: *Encyclopedia of measurement and statistics* 2, pp. 508–510.

^[2]Jan Hauke and Tomasz Kossowski (2011). "Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data". In: *Quaestiones geographicae* 30.2, pp. 87–93.

k -nearest neighbors imputer^{[3][4]}

The k -nearest neighbors imputer estimating the missing values using the k nearest neighbors which:

- Works with both numerical and categorical data;
- Don't need any assumption about the data distribution;
- robust in many applications.

^[3]Olga Troyanskaya et al. (2001). "Missing value estimation methods for DNA microarrays". In: *Bioinformatics* 17.6, pp. 520–525.

^[4]Afaq Juna et al. (2022). "Water quality prediction using KNN imputer and multilayer perceptron". In: *Water* 14.17, p. 2592.

Dimension raising

Given the scalar data x_i and a given degree d , we compute the new data as

$$(x_i, x_i^2, \dots, x_i^d).$$

Combination of different features are also used in dimension raising, e.g. given scalar x_i and y_i , another kind of new data is computed as

$$(x_i, y_i, x_i y_i).$$

Dimension raising

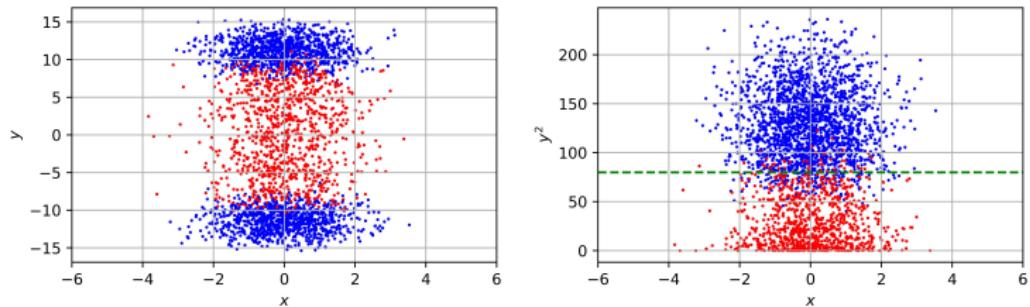


Figure: Example for dimension raising. The left shows that origin data, where the data is not linearly separable, while the right shows the data after dimension raising, where the y^2 is easy to be separated by a line.

Data Analysis

- **The logistic regression**^{[5][6]} is a widely used linear classification model, which gives a probability value ranging between 0 and 1.
- **The decision tree**^{[7][8]} is a supervised learning method used for classification which predict the label with piecewise constant approximation.
- **The multilayer perceptron (MLP)**^{[9][10]} is a basic kind of neural network which learns a function $f: \mathbb{R}^n \mapsto \mathbb{R}^m$ to approximate the input and output, with the ability of hierarchical feature extraction.

[5] Strother H Walker and David B Duncan (1967). "Estimation of the probability of an event as a function of several independent variables". In: *Biometrika* 54.1-2, pp. 167–179.

[6] Maja Pohar, Mateja Blas, and Sandra Turk (2004). "Comparison of logistic regression and linear discriminant analysis: a simulation study". In: *Metodoloski zvezki* 1.1, p. 143.

[7] Paul E Utgoff (1989). "Incremental induction of decision trees". In: *Machine learning* 4, pp. 161–186.

[8] Sotiris B Kotsiantis (2013). "Decision trees: a recent overview". In: *Artificial Intelligence Review* 39, pp. 261–283.

[9] Frank Rosenblatt (1958). "The perceptron: a probabilistic model for information storage and organization in the brain.". In: *Psychological review* 65.6, p. 386.

[10] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams (1986). "Learning representations by back-propagating errors". In: *nature* 323.6088, pp. 533–536.

Task 1 - Overview

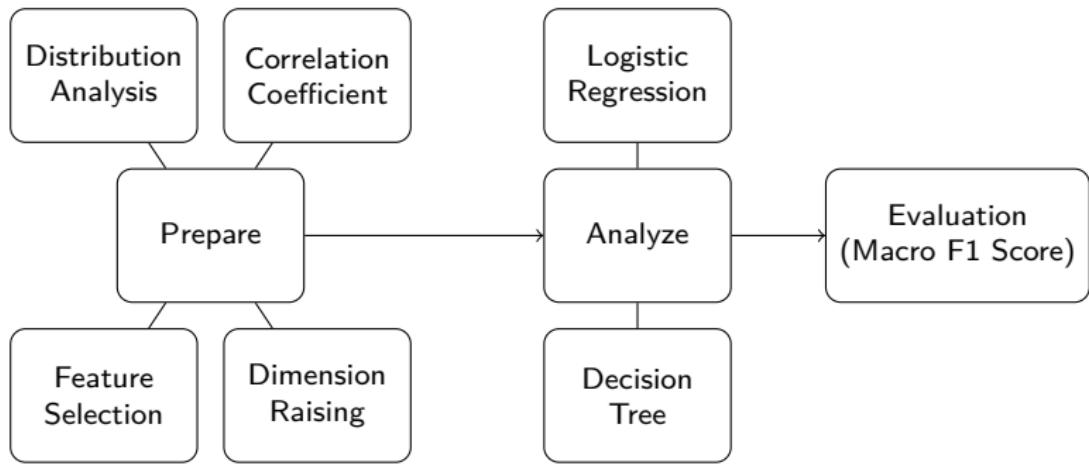


Figure: Overview flowchart for task 1.

Task 1 - Data Distribution

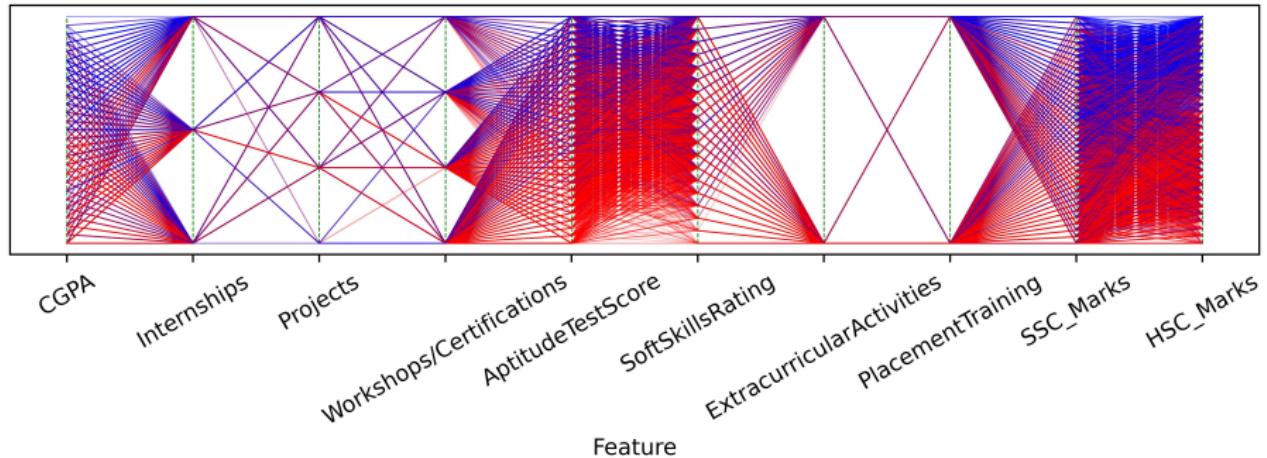


Figure: The parallel coordinate plot for each features.

Task 1 - Data Distribution

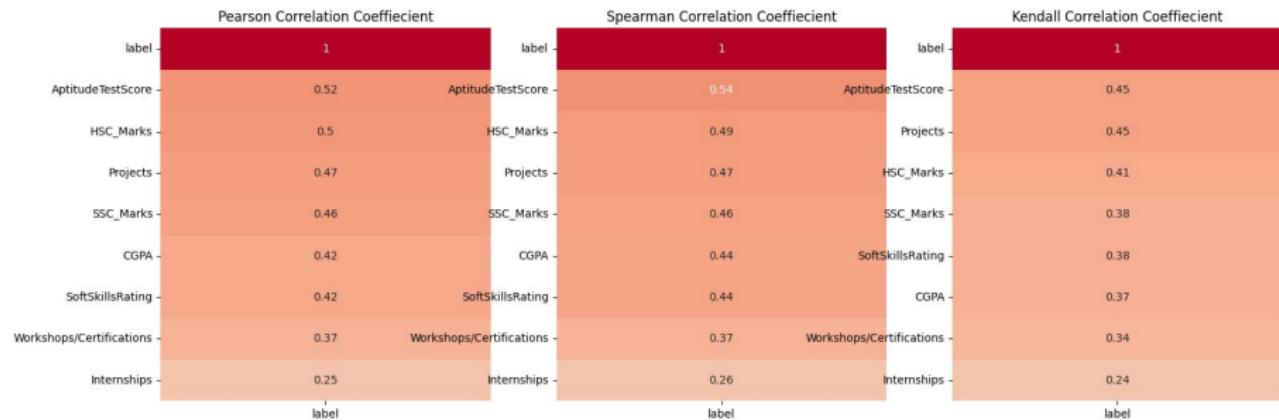


Figure: Correlation coefficient for each features.

Task 1 - Performance - Logistic Regression

Penalty	Dimension Raising (Degree)	F1 Score (Test/Train)	Time/Mem (s/MB)
None	None	0.7909/0.7933	1.55/735
None	2	0.7905/0.7936	1.41/730
None	3	0.7909/0.7936	1.28/741
Lasso	None	0.7912/0.7927	1.66/736
Ridge	None	0.7915/0.7928	1.65/735

Table: Preformance for logistic regression. We tested each option with 5 times cross validation (80% for training set, 20% for training set) and 1 times that using all data as training set. The f1 scores showed in the table is the average f1 score for cross validation and using all data as training set (in this case, we just tested the model on training data) and using all data as training set (in this case, we just tested the model on training data), the time is the total time for 6 run, and the memory usage is tested using all the data as training set.

Task 1 - Performance - Decision Tree

Max Depth	Number of Threshold	Criterion	F1 Score (Test/Train)	Time/Mem (s/MB)
10	64	gini	0.7502/0.8677	17.5/497
10	128	gini	0.7485/0.8698	16.4/498
12	64	gini	0.7266/0.9089	25.2/497
10	64	entropy	0.7428/0.8500	16.4/501

Table: Preformance for decision tree. We tested each option with 5 times cross validation (80% for training set, 20% for training set) and 1 times that using all data as training set. The f1 scores showed in the table is the average f1 score for cross validation and using all data as training set (in this case, we just tested the model on training data) and using all data as training set (in this case, we just tested the model on training data), the time is the total time for 6 run, and the memory usage is tested using all the data as training set.

Task 2 - Overview

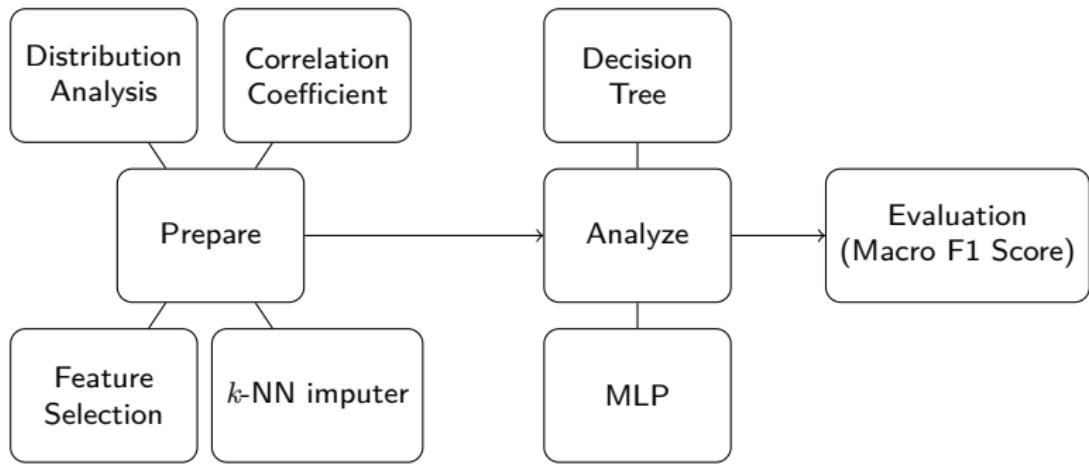


Figure: Overview flowchart for task 2.

Task 2 - Data Distribution

Feature	Number Of Miss Value	Feature	Number Of Miss Value	Feature	Number Of Miss Value
X01	0	Y01	0	Z01	0
X11	0	Y11	0	Z11	0
X21	0	Y21	0	Z21	0
X31	355	Y31	355	Z31	355
X41	1604	Y41	1604	Z41	1604
X51	6634	Y51	6634	Z51	6634

Table: The number of missing values for each feature in training data. (The total number of data is 40000.)

Task 2 - Data Distribution

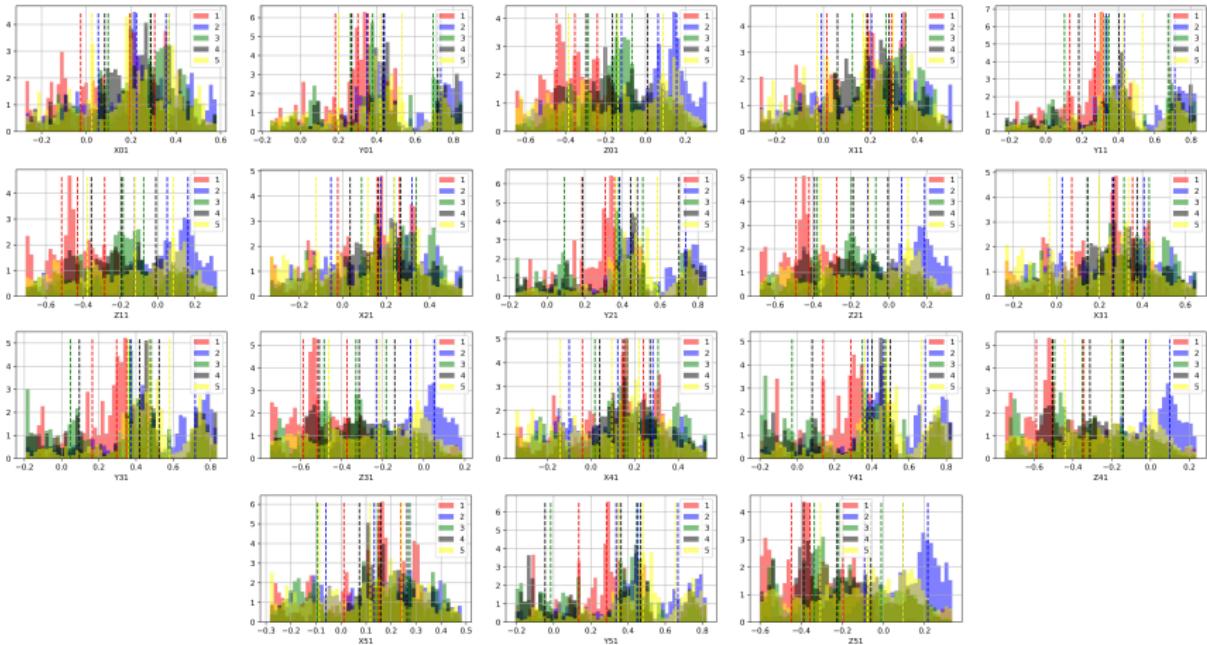


Figure: The distribution for each feature (the top and bottom 5% points are considered as outliers, and are not shown). The vertical lines shows the position of first, second and third quartiles.

Task 2 - Data Distribution

Pearson			Spearman			Kendall			Pearson			Spearman			Kendall			
label-1	1	label-1	1	label-1	1	label-1	1	label-1	label-2	1	label-2	1	label-2	1	label-2	1	label-2	
X11	-0.028	X51	-0.03	X51	-0.024	Z41	0.42	Z41	Z41	0.42	Z41	0.34	Z51	0.41	Z51	0.34	Z51	
X41	-0.039	X41	-0.042	X41	-0.035	Z51	0.42	Z51	Z51	0.41	Z51	0.34	Z31	0.4	Z31	0.33	Z31	
X31	-0.045	X31	-0.05	X31	-0.041	Z31	0.4	Z31	Z31	0.4	Z31	0.3	Z21	0.38	Z21	0.32	Z21	
X21	-0.056	X21	-0.061	X21	-0.05	Z21	0.39	Z21	Z21	0.4	Z21	0.31	Z11	0.38	Z11	0.31	Z11	
X11	-0.072	X11	-0.074	X11	-0.06	Z11	0.37	Z11	Z11	0.38	Z11	0.31	Z01	0.37	Z01	0.3	Z01	
X01	-0.14	X01	-0.14	X01	-0.11	Z01	0.36	Z01	Z01	0.37	Z01	0.3						
Y51	-0.2	Y51	-0.24	Y51	-0.19	Y41	0.25	Y41	Y41	0.25	Y41	0.2	Y51	0.25	Y51	0.2	Y51	
Y01	-0.2	Y01	-0.24	Y01	-0.2	Y51	0.25	Y51	Y01	0.21	Y01	0.17	Y31	0.24	Y31	0.2	Y31	
Y41	-0.22	Z51	-0.26	Z51	-0.21	Y31	0.24	Y31	Y21	0.23	Y21	0.19	Y11	0.21	Y11	0.17	Y11	
Y31	-0.23	Y41	-0.26	Y41	-0.21	Y21	0.23	Y21	Y11	0.21	Y11	0.17	X21	0.041	X21	0.034	X21	
Y21	-0.23	Z41	-0.27	Z41	-0.22	Y01	0.21	Y01	Y01	0.21	Y01	0.17	X01	0.04	X01	0.039	X01	
Y11	-0.23	Y31	-0.28	Y31	-0.22	X21	0.041	X21	X21	0.041	X21	0.034	X31	0.037	X31	0.035	X31	
Z51	-0.25	Y21	-0.28	Y21	-0.23	X01	0.04	X01	X01	0.04	X01	0.032	X51	0.032	X51	0.032	X51	
Z41	-0.26	Y11	-0.28	Y11	-0.23	X31	0.037	X31	X31	0.033	X31	0.029	X41	0.021	X41	0.016	X41	
Z01	-0.26	Z31	-0.29	Z31	-0.23	X51	0.033	X51	X51	0.033	X51	0.027						
Z31	-0.27	Z01	-0.29	Z01	-0.23	X11	0.032	X11	X11	0.032	X11	0.026						
Z21	-0.28	Z21	-0.3	Z21	-0.24	X41	0.021	X41	X41	0.021	X41	0.016						
Z11	-0.3	Z11	-0.31	Z11	-0.26													
	label-1	label-1	label-1	label-1	label-1				label-2	label-2	label-2	label-2						
Pearson			Spearman			Kendall			Pearson			Spearman			Kendall			
label-3	1	label-3	1	label-3	1	label-3	1	label-3	label-4	1	label-4	1	label-4	1	label-4	1	label-4	
X11	0.15	X21	0.16	X01	0.13	X41	0.11	X41	X41	0.099	X41	0.081	Z51	0.097	Z51	0.079	Z51	
X21	0.15	X11	0.16	X11	0.13	X31	0.11	X31	X31	0.093	X31	0.076	Z01	0.093	Z01	0.071	Z01	
X01	0.13	X01	0.14	X01	0.11	X21	0.085	X21	X21	0.087	X21	0.071	X11	0.057	X11	0.053	X11	
X31	0.12	X31	0.13	X31	0.1	X01	0.081	X01	X01	0.071	X01	0.065	X21	0.056	X21	0.046	X21	
X41	0.099	X41	0.1	X41	0.081	X11	0.057	X11	X11	0.037	X11	0.031	X31	0.033	X31	0.031	X31	
X51	0.041	X51	0.039	X51	0.032	X21	0.04	X21	X21	0.021	X21	0.016	X31	0.029	X31	0.013	X31	
Z01	0.0013	Z01	-0.0058	Z01	-0.0047	X01	0.035	X01	X01	0.021	X01	0.012	X01	0.0096	X01		X01	
Z01	-0.033	Z11	-0.04	Z11	-0.032	X11	0.035	X11	X11	0.016	X11	0.013	Z21	-0.071	Z21	-0.061	Z21	
Z11	-0.053	Y01	-0.052	Y01	-0.042	X21	0.021	X21	X21	0.015	X21	0.012	Z31	-0.042	Z31	-0.035	Z31	
Z21	-0.094	Y11	-0.093	Y11	-0.076	X31	0.015	X31	X31	0.012	X31	0.009	Z41	-0.042	Z41	-0.035	Z41	
Y11	-0.1	Z21	-0.095	Z21	-0.077	X41	-0.037	X41	X41	-0.029	X41	-0.024	Z51	-0.042	Z51	-0.035	Z51	
Z51	-0.11	Z51	-0.1	Z51	-0.085	Z01	-0.042	Z01	Z01	-0.042	Z01	-0.035	Z11	-0.071	Z11	-0.058	Z11	
Z31	-0.12	Z31	-0.12	Z31	-0.095	Z11	-0.068	Z11	Z11	-0.075	Z11	-0.061	Z21	-0.094	Z21	-0.077	Z21	
Z41	-0.14	Z41	-0.14	Z41	-0.11	Z21	-0.075	Z21	Z21	-0.075	Z21	-0.061	Z31	-0.094	Z31	-0.077	Z31	
Y21	-0.19	Y21	-0.17	Y21	-0.14	Z31	-0.095	Z31	Z31	-0.094	Z31	-0.077	Z41	-0.1	Z41	-0.084	Z41	
Y51	-0.22	Y31	-0.2	Y31	-0.16	Z41	-0.1	Z41	Z41	-0.094	Z41	-0.077	Z51	-0.11	Z51	-0.12	Z51	
Y31	-0.22	Y51	-0.21	Y51	-0.17	Y51	-0.11	Y51	Y51	-0.15	Y51	-0.12						
Y41	-0.25	Y41	-0.22	Y41	-0.18													
	label-3	label-3	label-3	label-3	label-3				label-4	label-4	label-4	label-4						

Figure: Correlation coefficient for each features with label 1, 2, 3, 4.

Task 2 - Performance - Decision Tree

Max Depth	Num of Threshold	Criterion	Feature Selection	F1 Score (Test/Train)	Time/Mem (s/MB)
10	64	gini	False	0.8097/0.8359	153.6/606
10	128	gini	False	0.8043/0.8512	168.5/650
12	64	gini	False	0.8511/0.9106	172.8/601
10	64	entropy	False	0.8184/0.8596	147.7/608
10	64	gini	True	0.8052/0.8334	80.7/585

Table: Preformance for decision tree. We tested each option with 5 times cross validation (80% for training set, 20% for training set) and 1 times that using all data as training set. The f1 scores showed in the table is the average f1 score for cross validation and using all data as training set (in this case, we just tested the model on training data) and using all data as training set (in this case, we just tested the model on training data), the time is the total time for 6 run, and the memory usage is tested using all the data as training set.

Task 2 - Performance - MLP

Dropout	Hidden Size	Feature Selection	Batch Size	F1 Score (Test/Train)	Time/Mem (s/MB)
0.3	128	False	512	0.9617/0.9718	219.4/856
0.3	64	False	512	0.9173/0.9253	212.9/812
0.3	256	False	512	0.9767/0.9939	221.7/856
0.3	128	False	1024	0.9629/0.9743	148.6/857
0.3	128	False	512	0.9644/0.9750	381.4/859
0.3	128	True	512	0.9474/0.9559	171.7/848

Table: Preformance for MLP. We tested each option with 5 times cross validation (80% for training set, 20% for training set) and 1 times that using all data as training set. The f1 scores showed in the table is the average f1 score for cross validation and using all data as training set (in this case, we just tested the model on training data), the time is the total time for 6 run, and the memory usage is tested using all the data as training set.

Task 3 - Overview

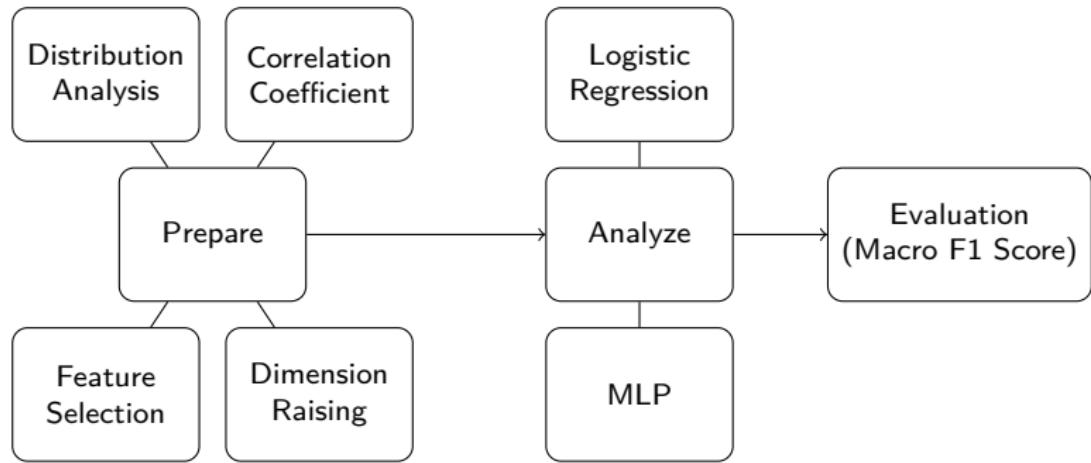


Figure: Overview flowchart for task 3.

Task 3 - Distribution Analysis

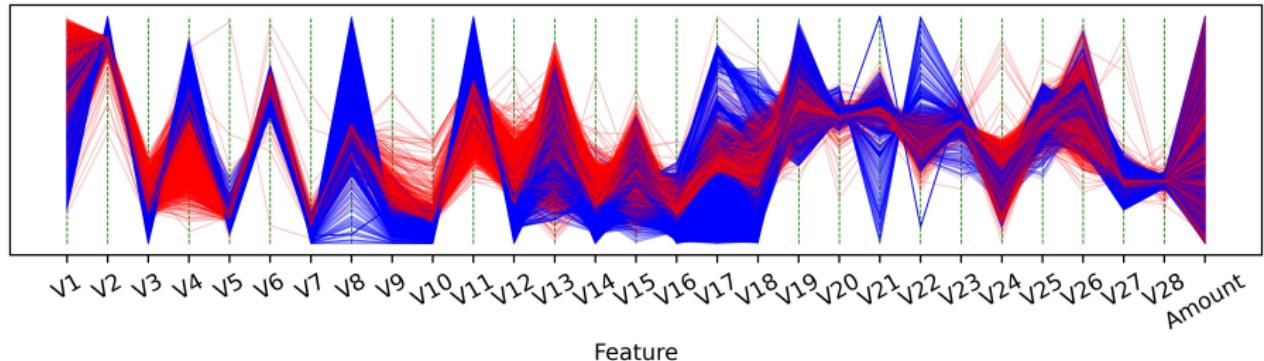


Figure: The parallel coordinate plot for each features (we chose 10% of total data).

Task 3 - Distribution Analysis

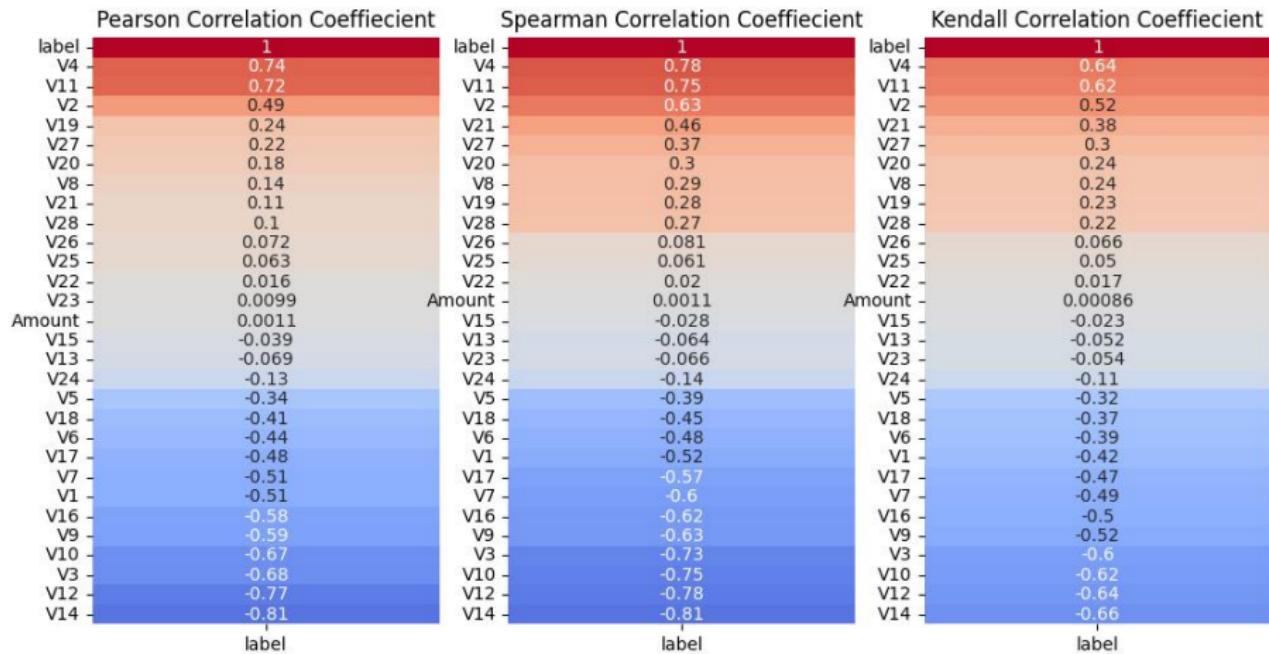


Figure: Correlation coefficient for each features.

Task 3 - Performance - Logistic Regression

Penalty	Dimension Raising (Degree)	Feature Selection	F1 Score (Test/Train)	Time/Mem (s/MB)
None	None	False	0.9586/0.9577	19.5/866
None	2	False	0.9619/0.9610	33.5/915
None	3	False	0.9621/0.9612	49.6/973
Lasso	None	False	0.9579/0.9570	21.5/854
Ridge	None	False	0.9553/0.9544	19.8/853
None	None	True	0.9582/0.9573	16.5/825

Table: Preformance for logistic regression. We tested each option with 5 times cross validation (80% for training set, 20% for testing set) and 1 times that using all data as training set. The f1 scores showed in the table is the average f1 score for cross validation and using all data as training set (in this case, we just tested the model on training data), the time is the total time for 6 run, and the memory usage is tested using all the data as training set.

Task 3 - Performance - MLP

Dropout	Hidden Size	Feature Selection	F1 Score (Test/Train)	Time/Mem (s/MB)
0.1	32	False	0.9522/0.9548	31.9/929
0.2	32	False	0.9000/0.9207	31.1/916
0.1	32	True	0.9481/0.9480	31.3/899
0.1	16	False	0.9503/0.9493	31.7/929
0.1	64	True	0.9500/0.9501	29.7/902

Table: Preformance for MLP. We tested each option with 5 times cross validation (80% for training set, 20% for training set) and 1 times that using all data as training set. The f1 scores showed in the table is the average f1 score for cross validation and using all data as training set (in this case, we just tested the model on training data), the time is the total time for 6 run, and the memory usage is tested using all the data as training set.

Summary

Reference I

-  Abdi, Hervé (2007). “The Kendall rank correlation coefficient”. In: *Encyclopedia of measurement and statistics* 2, pp. 508–510.
-  Hauke, Jan and Tomasz Kossowski (2011). “Comparison of values of Pearson’s and Spearman’s correlation coefficients on the same sets of data”. In: *Quaestiones geographicae* 30.2, pp. 87–93.
-  Juna, Afaq et al. (2022). “Water quality prediction using KNN imputer and multilayer perceptron”. In: *Water* 14.17, p. 2592.
-  Kotsiantis, Sotiris B (2013). “Decision trees: a recent overview”. In: *Artificial Intelligence Review* 39, pp. 261–283.
-  Pohar, Maja, Mateja Blas, and Sandra Turk (2004). “Comparison of logistic regression and linear discriminant analysis: a simulation study”. In: *Metodoloski zvezki* 1.1, p. 143.
-  Rosenblatt, Frank (1958). “The perceptron: a probabilistic model for information storage and organization in the brain.”. In: *Psychological review* 65.6, p. 386.

Reference II

-  Rumelhart, David E, Geoffrey E Hinton, and Ronald J Williams (1986). "Learning representations by back-propagating errors". In: *nature* 323.6088, pp. 533–536.
-  Troyanskaya, Olga et al. (2001). "Missing value estimation methods for DNA microarrays". In: *Bioinformatics* 17.6, pp. 520–525.
-  Utgoff, Paul E (1989). "Incremental induction of decision trees". In: *Machine learning* 4, pp. 161–186.
-  Walker, Strother H and David B Duncan (1967). "Estimation of the probability of an event as a function of several independent variables". In: *Biometrika* 54.1-2, pp. 167–179.