Mortality Predictive Analysis on Acute Pancreatitis in Intensive Care Unit Patients - MIMIC-IV Dataset

Danlan Chen[1], Edoardo De Duro[1], Lulu Qurotaini[1], Muhan Guan[1]
[1]University of Melbourne, Australia

1 Abstract (100 words)

Acute Pancreatitis have been one of the biggest contributors to mortality in ICU departments. The rate keeps increasing over the years, especially among patients with unclear risk factors. In this study, by adopting a data-driven approach, we aim to predict the mortality of patients with Acute Pancreatitis (AP). Using MIMIC-IV dataset, we extract a cohort of patients according to the medical criteria to assess AP, and we use ~ 19 *different features as possible indicators for the mortality prediction task. Surprisingly, the ICD cohort obtained the best performance using the XGBoost model.*

2 Introduction

Acute Pancreatitis (AP) is a severe inflammation of the pancreas and it is one of the most common causes of hospitalization admission in the USA[1]. Despite the mortality caused by AP having recently decreased, this disease remains a big cause of death in ICU departments (1%-5%[2]) and the epidemiology is predicted to increase over the next generation[3]. Remarkably, the mortality rates seem to be higher in subgroups of patients characterized by unclear clinical and laboratory risk factors[1]. Public health specialists should help reduce the burden of Pancreatitis[4]. In our case, we aim to predict mortality and identify the strongest predictors of death due to AP by utilizing MIMIC-IV[5] database.

Addressing these tasks using data-driven methods is relevant as it would help to choose the most appropriate treatment options and optimize resources in hospitals[6]. Additionally, it would allow the identification of possible predictors of adverse outcomes allowing for early treatments of patients with AP.

3 Methods

The section comprises our methodology in detail. It includes a description of our dataset, clustering, and predictive analysis. A brief summary of our methodology is shown in Figure 1.

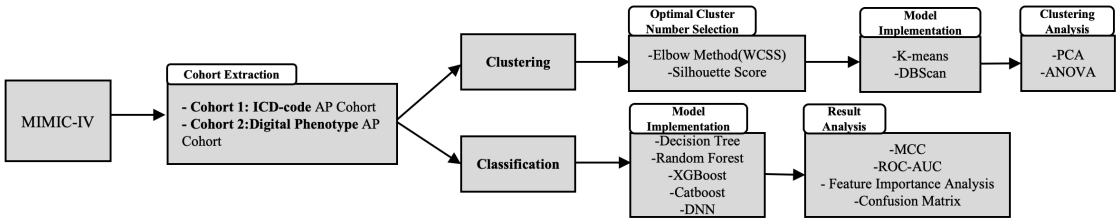

Figure 1. Flowchart summing up the analysis process.

3.1 Dataset

Extraction of the Digital Phenotype AP Cohort: Patients were extracted from ICU ward records based on the following criteria for AP diagnosis[7]:

[1]https://www.uptodate.com/contents/predicting-the-severity-of-acute-pancreatitis

1. Abdominal pain linked to AP (filtering the 'value' field in the table 'chartevent' for keywords)
2. Amylase and/or lipase levels exceeding three times the normal threshold (calculating from the two corresponding test indicators in table 'chartevent')
3. CT scans revealing significant AP imagery (filtering the 'text' field in the table 'radiology_note' for keywords)

Patients satisfying any two of the above criteria were added to a temporary cohort. Subsequently, duplicates were removed from this cohort to form the final list.

Extraction of the ICD AP Cohort: Patients with an ICD-9 code of 577.0 and recorded in the ICU dataset (table 'icustays') were selected.

For the feature selection part, SQL was employed to extract data from the MIMIC-IV database. General information such as age and sex were gathered and clinical and laboratory variables were taken within the first 24 hours post-admission. The outcome variable was in-hospital mortality (extracted from the 'admission' table).

Feature Type	Feature Name
Categorical Features	Gender
	Alcoholism
Numeric Features	Age
	Non Invasive Blood Pressure systolic
	Non Invasive Blood Pressure diastolic
	White Blood Cells(WBC)
	Lipase
	Heart rate Alarm High
	Heart rate Alarm low
	Platelet(PLT)
	Amylase(Serum)
	Creatinine(serum)
	Glucose (serum)
	Hematocrit(serum)
	Blood Urea Nitrogen(BUN)
	Anion gap
	Prothrombin time
	Alanine aminotransferase (ALT)
	Total Bilirubin
	Weight
	Height
	Lactate Dehydrogenase (LDH)

Figure 2. Features extracted.

If a variable was recorded multiple times within the initial 24 hours, only the first record was used. Data from only the first admission of each patient was considered. Patients with more than 5% of their data missing and those below 18 years of age were excluded[8]. Finally, features with more than 20% missing values were removed from both cohorts. The remaining missing values were then imputed based on the distribution of the respective feature types (continuous or categorical) to produce the final datasets for both cohorts.

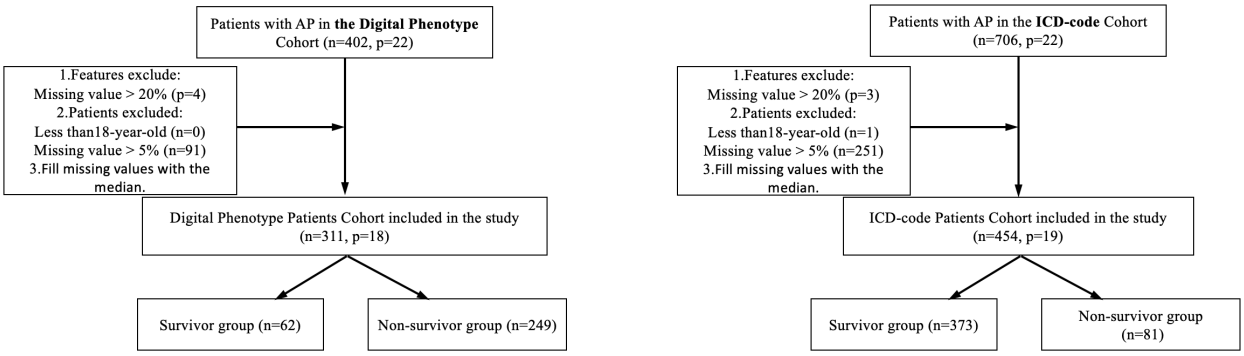

Figure 3. Cohort extraction flowchart.

Based on the flowchart, these two final cohorts for our study have 'n' patients and 'p' features.

3.2 Categorical visualization analysis

The Kaplan-Meier survival analysis segmented by age groups unveils distinct survival patterns[9].

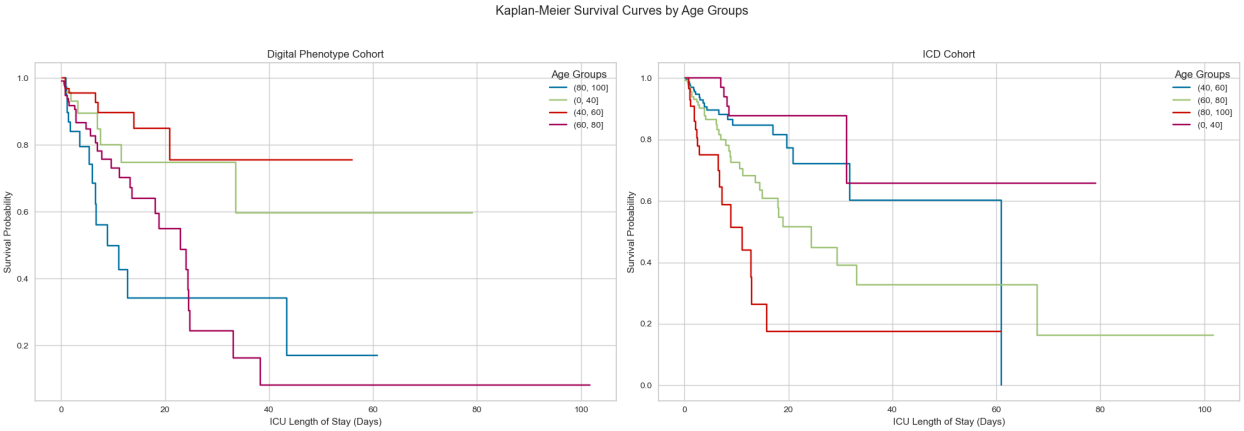

Figure 4. Kaplan-Meier Survival Analysis for Two Cohorts

Figure 4 reveals both cohorts show similar survival trends, patients aged 20-60 show a slower initial drop in survival, suggesting a lower mortality risk, while those aged 60-100 face a sharp decrease. Figure 5 shows more male patients in the ICD Cohort than females, while in the phenotype cohort, males have a higher mortality rate compared with females. Summary statistics (of continuous and categorical variables) of the two cohorts are presented in Appendix A.

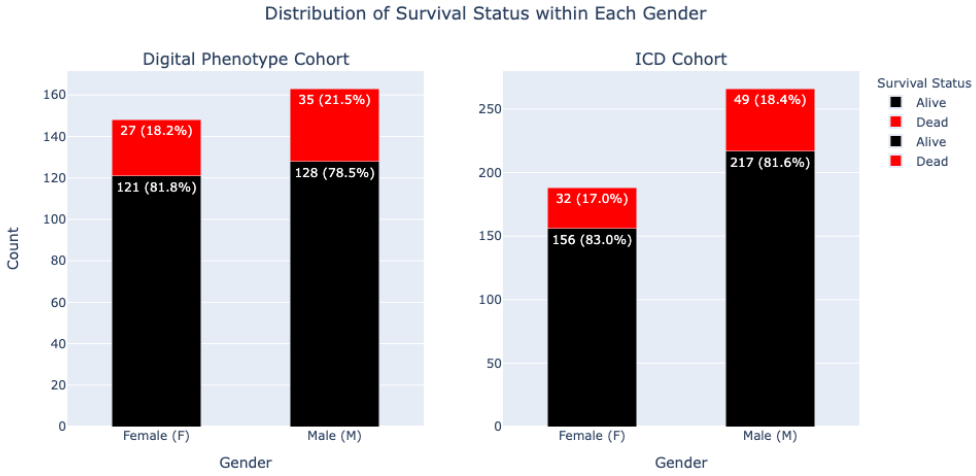

Figure 5. Survival Status within Genders

3.3 Clustering analysis

Our objective is to characterise patients effectively, providing insight into underlying mortality symptoms. In this case, we employ the K-Prototype algorithm ('Cao'[10] initialization). The method handles both numerical and categorical features by integrating dissimilarity measures for diverse data types[11]. Attributes `gender`, `alcoholism`, and `mortality` are defined as categorical, while numerical features are retained, excluding the unique `hadm_id` attribute which was removed from analysis. Additionally, we explore DBSCAN ($\epsilon = 3$), a non-parametric method known for discovering clusters based on density. We aim to compare the parametric and non-parametric clustering methods to provide additional insights of reliability.

$$WCSS = \sum_{i=1}^{k} \sum_{j=1}^{n_i} d(x_j, m_i) \tag{1}$$

$$S(x_i) = \frac{b(x_i) - a(x_i)}{\max\{a(x_i), b(x_i)\}} \tag{2}$$

$$a(x_i) = \frac{1}{|C_i| - 1} \sum_{x_j \in C_i, x_j \neq x_i} d(x_i, x_j)$$

$$b(x_i) = \min_{j \neq i} \left(\frac{1}{|C_j|} \sum_{x_k \in C_j} d(x_i, x_k) \right)$$

where:

n_i : number of points in cluster C_i

m_i : centroid of cluster C_i

C_j : the j-th cluster

$d(x_i, x_j)$: distance between data points x_i and x_j

To determine the optimal number of clusters k ($k \in [2, 15]$) for K-Prototye, we utilise elbow method and silhouette score with adjusted distance matrix for different data types. We aim to find low within-cluster sum of squares (WCSS; 1) and high silhouette scores (2). Upon clustering, independent parametric testing (ANOVA) and principal component analysis (PCA) are employed for analysis.

3.4 Predictive analysis

The mortality risk prediction is expressed as supervised learning (probabilistic classification), assigning binary labels to individuals; high mortality rate (y = 1) or not (y = 0). Five machine learning models are applied in the phenotype and ICD cohort described above for mortality risk prediction. Decision Tree is employed as the baseline compared with ensemble learning techniques (XGBoost, Random Forest, CatBoost) and neural network (Deep Neural Network). Each dataset is divided into two subsets: training and test set (80/20). Before training the model, the models' hyperparameters are tuned using grid-search cross-validation. We also employ Adaptive Synthetic Sampling (ADASYN), an oversampling method to address the imbalance in the dataset.

MCC (Matthews Correlation Coefficient) and ROC-AUC (Receiver Operating Characteristic-Area Under Curve) are used to evaluate models' performances. MCC ranges between -1 and +1, with -1 representing the worst prediction and +1 representing the best prediction [12]. The ROC-AUC as primary metric is used. With a higher AUC comes a better detection performance [13]. Given the confusion matrix, the metrics are calculated as,

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{3}$$

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR}^{-1}(t))\, dt \tag{4}$$

4 Results

In the results section, key findings from the analysis are presented. The study's outcomes are summarized, revealing insights regarding the research questions.

4.1 Clustering analysis

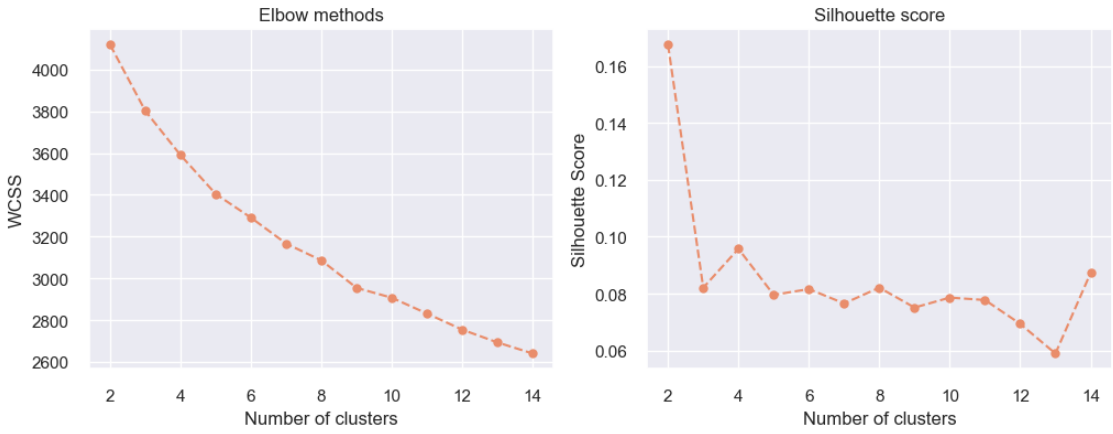

Figure 6. Evaluation result of K-Prototype on digital phenotype cohort

In the phenotype cohort, DBSCAN identifies 3 distinct clusters with a silhouette score of 7.8%. In contrast, K-Prototype analysis suggests 4 clusters as optimal, achieving a maximum silhouette score of 9.6% and WCSS of 3589.7 (91.3% variance explained) (Figure 6). Table 1 provides an overview of each cluster's characteristics, calculated from the average distribution (for a detailed one, refer to Appendix B). Notably, clusters with the highest mortality rates are PHENO_DB1, PHENO_KP1, and PHENO_KP2, sharing common traits such

as elevated levels of aminotransferase, creatinine, blood urea nitrogen (BUN), and anion gap. High glucose, aminotransferase, and prothrombin time are observed consistently across all clusters. Additionally, both algorithms highlight the significance of BUN, creatinine, white blood cells, platelet count, aminotransferase, bilirubin, anion gap, and weight (Appendix C).

Table 1. Characteristics of clusters on digital phenotype cohort

Cluster ID	#	Age	Gender	Alcoholism	Mortality (%)	Key Features
PHENO_DB1	97	Old	Male	No	39.18	Higher BUN, creatinine, white blood, aminotransferase, bilirubin, anion gap, prothrombin time, and weight.
PHENO_DB2	162	Old	Female	No	7.41	Lower weight.
PHENO_DB3	18	Young	Male	Yes	0.00	Higher bilirubin and weight.
PHENO_KP1	131	Old	Male	No	14.44	Higher hematocrit, white blood, and weight.
PHENO_KP2	90	Young	Male	No	33.33	Higher aminotransferase, prothrombin time, creatinine, bilirubin, anion gap, and weight.
PHENO_KP3	47	Old	Female	No	13.74	Lower hematocrit, weight.
PHENO_KP4	9	Old	Male	No	34.04	Higher BUN, creatinine, white blood, anion gap. Lower hematocrit.

The ICD cohort yields an agreement with the phenotype cohort in the number of clusters. K-Prototype, as depicted in Figure 7, favours 4 clusters with 9.8% silhouette score and 5440.0 WCSS, increases from the phenotype cohort. DBSCAN performs at 6.2% silhouette score. Appendix D provides detailed cluster characteristics. Table 2 reveals clusters associated with higher mortality rates, including ICD_DB1, ICD_KP2, and ICD_KP3, have higher BUN, creatinine, aminotransferase, and anion gap. Additionally, all clusters demonstrate high aminotransferase, lipase, and glucose serum. Both algorithms emphasize the significance of BUN, creatinine, blood pressure, glucose serum, aminotransferase, bilirubin, and anion gap (Appendix E).

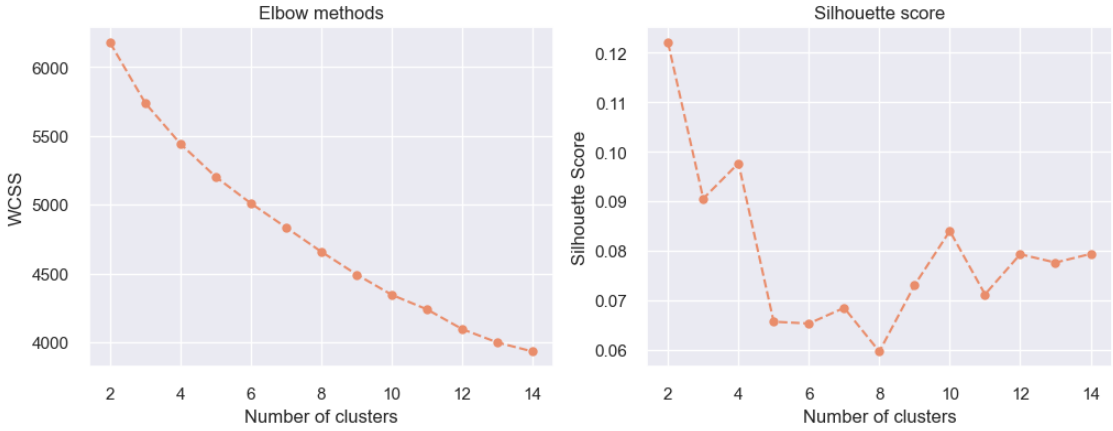

Figure 7. Evaluation result of K-Prototype on ICD cohort

Table 2. Characteristics of clusters on ICD cohort

Cluster ID	#	Age	Gender	Alcoholism	Mortality (%)	Key Features
ICD_DB1	118	Young	Male	No	40.68	Higher BUN, creatinine, white blood, aminotransferase, prothrombin time, bilirubin, anion gap
ICD_DB2	271	Young	Male	No	2.95	-
ICD_DB3	5	Old	Male	No	100.00	Higher BUN, white blood. Lower weight.
ICD_KP1	186	Old	Female	No	11.83	-
ICD_KP2	40	Old	Male	No	30.00	Higher BUN, creatinine, prothrombin time, aminotransferase, bilirubin, anion gap
ICD_KP3	56	Old	Male	No	28.57	Higher BUN, anion gap, and weight
ICD_KP4	112	Young	Male	No	9.82	Higher white blood, and weight

4.2 Predictive analysis

The prediction performance is summarised in Table 3. In phenotype cohort, the model performs an average of 0.84 AUC-ROC and 0.68 MCC on training set. Meanwhile, ICD cohort shows 0.87 AUC-ROC and $0,75$ MCC. CatBoost outperforms in training set on both phenotype and ICD cohort. However, the decision tree and XGBoost model exhibit better performance on the phenotype and ICD testing set, with 0.62 and 0.63 ROC-AUC respectively.

Table 3. Prediction Performances of Different Models

Phenotype cohort	Training		Testing	
Model	AUC-ROC	MCC	AUC-ROC	MCC
Decision Tree	0.77	0.54	**0.62**	**0.25**
Random Forest	0.88	0.77	0.48	-0.03
XGBoost	0.86	0.72	0.46	-0.10
CatBoost	**0.90**	**0.80**	0.42	-0.19
DNN	0.79	0.59	0.54	0.10
ICD cohort	Training		Testing	
Model	AUC-ROC	MCC	AUC-ROC	MCC
Decision Tree	0.79	0.59	0.62	0.24
Random Forest	0.90	0.80	0.55	0.14
XGBoost	0.90	0.80	**0.63**	**0.28**
CatBoost	**0.92**	**0.85**	0.54	0.09
DNN	0.86	0.73	0.58	0.17

As seen in Figure 8, the overall performance of ICD cohort is above the phenotype, aligned with metric performance result. While models in phenotype cohort, except decision tree, are deemed underperformed, all models in ICD cohort exhibit a better result. Additionally, Figure 9 and 10 show the outcomes of the predictions for phenotype and ICD cohort respectively. We observe bias towards negative predictions with greater false negatives and fewer true negatives in all models.

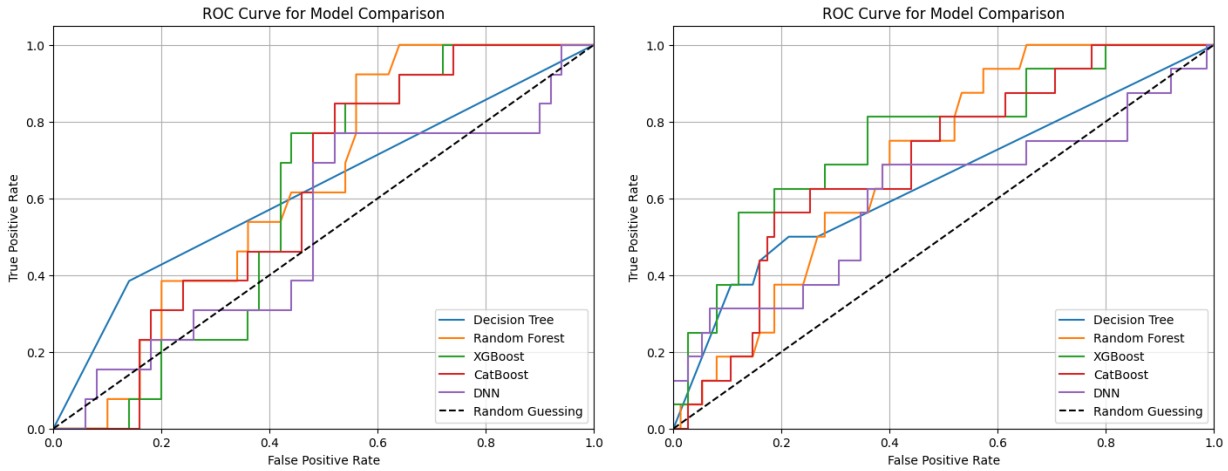

Figure 8. ROC curve for model on digital phenotype (left) and ICD (right)

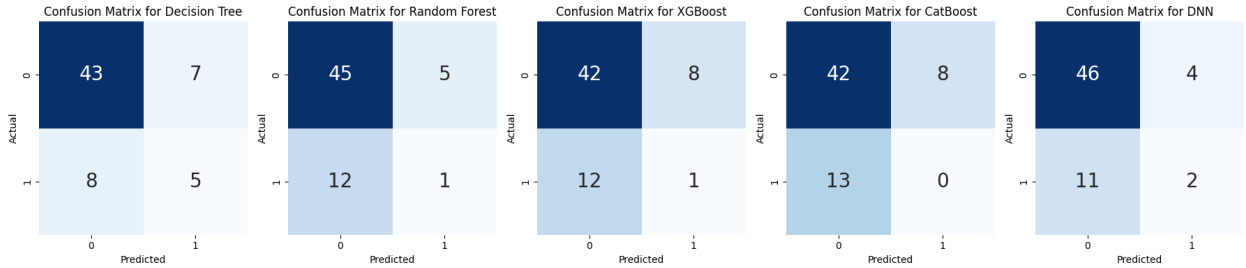

Figure 9. Confusion matrix for prediction insights on digital phenotype cohort

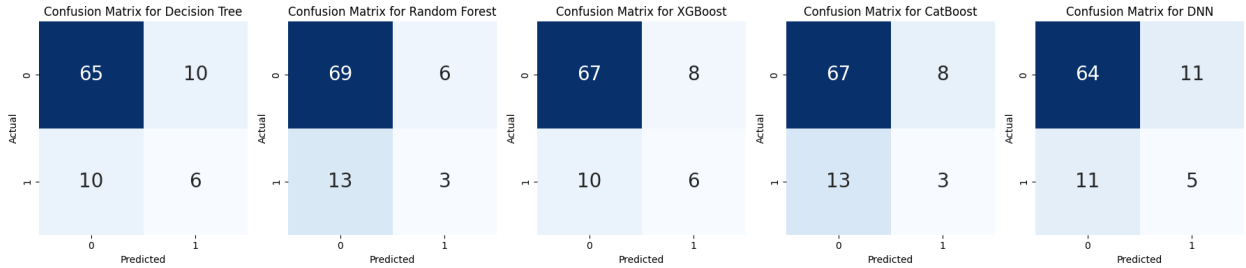

Figure 10. Confusion matrix for prediction insights on ICD cohort

5 Discussion

In this section, we interpret the findings and explore their implications. We consider the broader context, encouraging several literatures for a deeper understanding.

5.1 Clustering analysis

Upon clustering analysis, we observe the consistency of patients with higher mortality rates, including higher levels of BUN, serum creatinine, prothrombin time, aminotransferase, and white blood cells (Figure 11). These features are aligned with the diagnostic assessment of AP severity levels[14]. High aminotransferase,

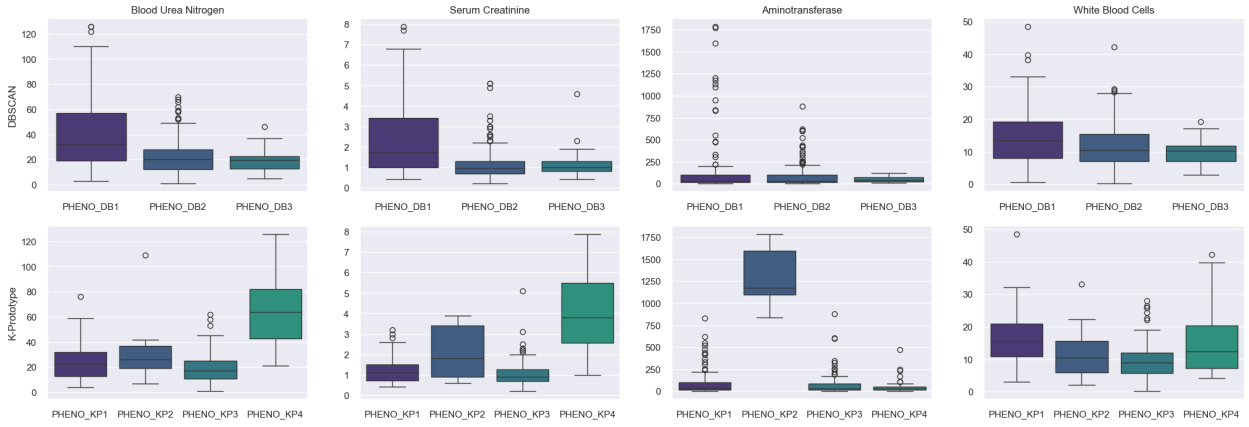

Figure 11. Boxplot of overall clusters on phenotype cohort

when related to AP, may happen due to damage in the metabolism, such as liver disease, which is more common in younger age[14]. Although we find the case in older people, we notice the levels of aminotransferase are extremely elevated in the younger ones. Additionally, it is always followed by elevated prothrombin time, which indicates longer clotting for blood which is anticipated in severe liver disease[15]. When this happens, the mortality rate is at 40%. Meanwhile, higher creatinine and BUN usually indicate renal organ failure. On our results, we observe people experiencing this may have $> 30\%$ mortality rate.

From the characteristics, mortality may happen due to severe complications resulting in organ failure, including metabolism and urinary. This is observed on clusters PHENO_DB1, ICD_DB1, PHENO_KP4, and ICD_KP3. On a moderate level of severity, we observe several clusters with relatively less complex abnormalities, such as PHENO_KP3 and ICD_KP4. Lastly, there are also clusters without complication in chart events, yet mortality still exists, such as ICD_DB2 and PHENO_DB2. In this case, cluster analysis provides a decent segmentation of severity levels in AP.

Finally, in comparison, we remark the phenotype cohort has a better structurisation when it comes to clustering. It exhibits less WCSS with more representative clusters. Although DBSCAN provides a worse silhouette score, we notice really strong cluster representations with immediate insights into the severity and mortality. In Figure 12, DBSCAN remarks the cluster with 0% mortality by gathering "outlier" points. With the K-Prototype in Figure 12, we observe that 4 clusters are redundant in which PHENO_KP2 and PHENO_KP4 are located near each other.

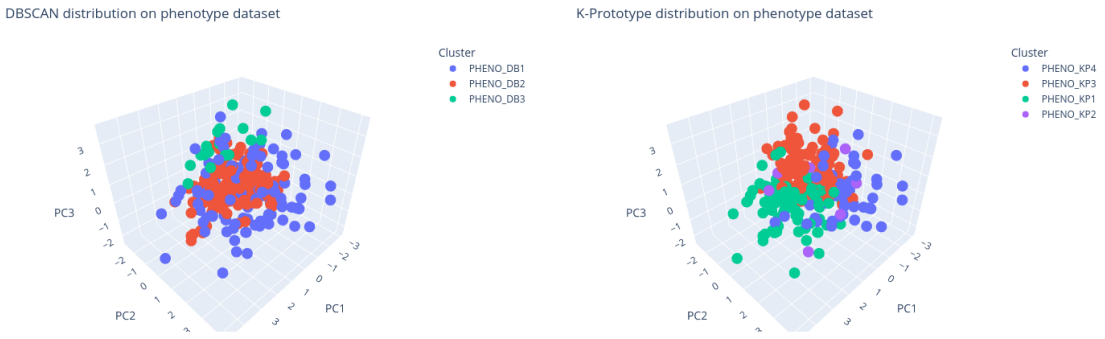

Figure 12. Distribution of clusters from DBSCAN and K-Prototype on phenotype cohort

5.2 Predictive analysis

On model performance, we remark the same pattern in complex models, having high performance on training yet failing to predict testing set. We presume the oversampling has increased the likelihood of overfitting, given a relatively insufficient and imbalanced dataset, as anticipated by literature [16]. Interestingly, this overfitting issue is absent in the simpler decision tree model. We suppose this happens as simpler model is less likely to deeply learn, and thus more flexible, given the repeating pattern of oversampling dataset.

On comparison, ICD cohort surprisingly provides slightly better results in terms of predictive performance using XGBoost. These results are in contrast with previous studies [17] [18] showing the importance of adopting clinically validated phenotyping. Given the result, ICD is more robust from overfitting than the phenotype, resulting a better results on testing performance. However, both the digital phenotyping and ICD-code system have their downsides: while in the first there might be data quality issues like inconsistency and incompleteness [19], the second might be susceptible to human errors in the coding phase (which most of the time happens at subsequently to patient discharge and mainly for billing purposes [18]).

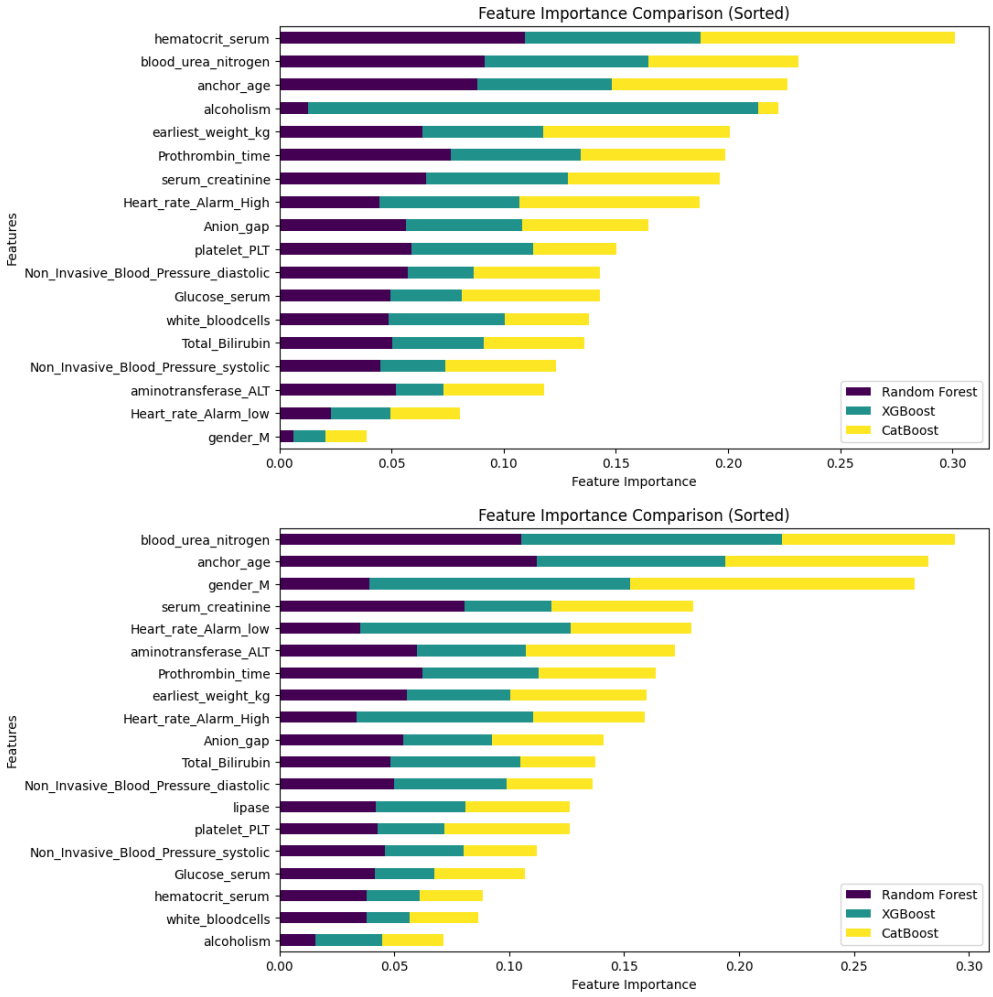

Figure 13. Feature importance for tree-based model on digital phenotype (top) and ICD cohort (bottom)

Upon classification, we explore feature importance as seen in Figure 13. BUN, hematocrit, age, and gender are considered most important in predicting mortality. BUN, as both cohorts agreed on the importance, is aligned with 5.1 result. Age and gender are considered as factors in assessing AP severity levels, in which females over 55 years old will likely experience chronic features [14]. Interestingly, gender importance is only

shown in ICD cohort. Contrary, hematocrit demonstrates more importance in phenotype cohort only, while hematocrit is also proven as a prognostic marker [14]. Additionally, we notice that the predictive model in ICD cohort puts more weight on demographic aspects (age and gender) whilst phenotype focuses more on chart and lab events.

6 Conclusion

The results mentioned above highlight the importance of the rationale behind cohort extraction for machine learning applications in healthcare. In fact, despite focusing on the same disease, the two cohorts extracted are different in terms of (a) composition (overlapping of patients between cohorts of just 80 patients), (b) clustering, giving insights into severity levels of AP, (c) performance in the mortality prediction task and (d) better-performing model in the same task (decision tree vs. XGBoost). Interestingly, the decision tree, despite being the simplest model, performs relatively well in both cases.

In comparison, the ICD cohort provides a better performance than the digital phenotype one. There might be many reasons for the incongruency with the previous literature mentioned above, but what remains, is the importance of clearly stating and designing extraction criteria apt to the tasks. In the future, healthcare AI should put lots of effort into understanding the subtle differences between coding systems and the clinical consequences of choosing one over the other.

Author Contributions: CRediT

Conceptualization: Edoardo De Duro;
Formal analysis: Muhan Guan, Lulu Quortaini, Danlan Chen;
Data curation: Muhan Guan, Lulu Quortaini, Danlan Chen;
Project administration: Lulu Quortaini;
Visualization: Lulu Quortaini, Edoardo De Duro;
Writing review & editing: Edoardo De Duro:
Investigation: Edoardo De Duro, Muhan Guan;
Methodology: Lulu Quortaini, Muhan Guan;
Supervision: Edoardo De Duro,
Validation: Lulu Quortaini, Muhan Guan;
Software: Muhan Guan.

References

1. Lee PJ, Papachristou GI. New insights into acute pancreatitis. Nature reviews Gastroenterology & hepatology. 2019;16(8):479-96.

2. Krishna SG, Kamboj AK, Hart PA, Hinton A, Conwell DL. The changing epidemiology of acute pancreatitis hospitalizations: a decade of trends and the impact of chronic pancreatitis. Pancreas. 2017;46(4):482.

3. Iannuzzi JP, King JA, Leong JH, Quan J, Windsor JW, Tanyingoh D, et al. Global incidence of acute pancreatitis is increasing over time: a systematic review and meta-analysis. Gastroenterology. 2022;162(1):122-34.

4. Petrov MS, Yadav D. Global epidemiology and holistic prevention of pancreatitis. Nature reviews Gastroenterology & hepatology. 2019;16(3):175-84.

5. Johnson A, Bulgarelli L, Pollard T, Horng S, Celi LA, Mark R. Mimic-iv. PhysioNet Available online at: https://physionet org/content/mimiciv/10/(accessed August 23, 2021). 2020.

6. Hameed MAB, Alamgir Z. Improving mortality prediction in Acute Pancreatitis by machine learning and data augmentation. Computers in Biology and Medicine. 2022;150:106077.

7. Banks PA, Bollen TL, Dervenis C, Gooszen HG, Johnson CD, Sarr MG, et al. Classification of acute pancreatitis—2012: revision of the Atlanta classification and definitions by international consensus. Gut. 2013;62(1):102-11.

8. Ding N, Guo C, Li C, Zhou Y, Chai X, et al. An artificial neural networks model for early predicting in-hospital mortality in acute pancreatitis in MIMIC-III. BioMed research international. 2021;2021.

9. Goel KPKJ M K. Understanding survival analysis: Kaplan-Meier estimate. International journal of Ayurveda research. 2010;1(4):274–278.

10. Cao F, Liang J, Bai L. A new initialization method for categorical data clustering. Expert Systems with Applications. 2009;36(7):10223-8. Available from: https://www.sciencedirect.com/science/article/pii/S0957417409001043.

11. Huang Z. CLUSTERING LARGE DATA SETS WITH MIXED NUMERIC AND CATEGORICAL VALUES; 1997. Available from: https://api.semanticscholar.org/CorpusID:3007488.

12. Chicco D, Tötsch N, Jurman G. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. BioData Mining. 2021 02;14.

13. Malek Salem SS. A comparison of one-class bag-of-words user behavior modeling techniques for masquerade detection. Security And Communication Networkds. 2012.

14. Lee S, Goh B, Chan C. In: Chapter 55 - Etiology, pathogenesis, and diagnostic assessment of acute pancreatitis; 2016. p. 883-96.e3.

15. Johnston DE. Special Considerations in Interpreting Liver Function Tests. American Family Physician. 1999;59(8):2223-30.

16. Gnip P, Vokorokos L, Drotár P. Selective oversampling approach for strongly imbalanced data. PeerJ Computer Science. 2021;7. Available from: https://api.semanticscholar.org/CorpusID:235766067.

17. Fedyukova A, Pires D, Capurro D. A comparative analysis of sepsis digital phenotyping methods. 2021:1-4.

18. He T, Belouali A, Patricoski J, Lehmann H, Ball R, Anagnostou V, et al. Trends and opportunities in computable clinical phenotyping: A scoping review. Journal of Biomedical Informatics. 2023:104335.

19. Botsis T, Hartvigsen G, Chen F, Weng C. Secondary use of EHR: data quality issues and informatics opportunities. Summit on translational bioinformatics. 2010;2010:1.

7 Appendices

A Data description of the cohorts extracted

Parameter Name	Mean	25th Percentile	75th Percentile	Unit
Anchor_age	59.977	48	73	years
Hematocrit_serum	32.454	27.3	37.1	%
Blood_urea_nitrogen	31.842	14	40.5	mg/dL
Serum_creatinine	1.814	0.8	2.1	mg/dL
Non_Invasive_Blood_Pressure_systolic	123.071	103	140.5	mmHg
Non_Invasive_Blood_Pressure_diastolic	70.527	56	83	mmHg
White_bloodcells	13.817	7.2	17.6	K/uL
Glucose_serum	152.17	103	168	mg/dL
Platelet_PLT	221.177	117.5	290	K/uL
Prothrombin_time	17.748	13.4	18.575	sec
aminotransferase_ALT	195.264	18	106	IU/L
Total_Bilirubin	2.759	0.5	2.5	mg/dL
Anion_gap	15.942	13	18	mEq/L
Heart_rate_Alarm_High	125.37	120	130	bpm
Heart_rate_Alarm_low	52.749	50	60	bpm
Earliest_weight_kg	80.408	64.95	93	kg
ICU_Length_of_Stay(LoS)	8.071	1.709	9.559	days

Table 4. Data description for cohort extracted using digital phenotype.

Parameter Name	Mean	25th Percentile	75th Percentile	Unit
Anchor_age	58.597	46.25	72	years
Hematocrit_serum	33.834	28.7	38.375	%
Blood_urea_nitrogen	29.537	13	37	mg/dL
Serum_creatinine	1.665	0.7	1.8	mg/dL
Non_Invasive_Blood_Pressure_systolic	128.637	110	146	mmHg
Non_Invasive_Blood_Pressure_diastolic	73.229	59	84	mmHg
Lipase	603.171	57.75	772.25	U/L
White_bloodcells	13.884	8.4	17.85	K/uL
Glucose_serum	159.628	102	178	mg/dL
Platelet_PLT	227.474	129.25	286.75	K/uL
Prothrombin_time	17.161	13	17.05	sec
Aminotransferase_ALT	226.8	24	150	IU/L
Total_Bilirubin	2.774	0.5	2.8	mg/dL
Anion_gap	16.194	13	18	mEq/L
Heart_rate_Alarm_High	124.064	120	130	bpm
Heart_rate_Alarm_low	54.46	50	60	bpm
Earliest_weight_kg	84.617	68.75	98.2	kg
ICU_Length_of_Stay(LoS)	8.216	1.806	9.796	days

Table 5. Data description for cohort extracted using icd criteria.

B Detailed characteristics of features for each clusters on digital phenotype cohort

	DBSCAN			K-Prototype			
	-1	0	1	0	1	2	3
Alcoholism	0.16	0.00	1.00	0.10	0.11	0.15	0.11
Gender (%)	M (59.79)	F (54.94)	M (61.11)	M (60.00)	M (66.67)	F (57.25)	M (55.32)
Age	55.51	64.88	49.56	60.86	49.89	59.79	64.43
Hematocrit serum (35-45%)	31.60	33.21	31.91	38.58	32.02	29.76	28.93
Blood urea nitrogen ($\leq$20)	41.55	22.56	20.22	24.13	34.89	19.30	64.57
Serum creatinine ($\leq$1.71)	2.46	1.20	1.27	1.24	2.00	1.03	4.05
Blood pressure (systolic)	125.19	121.33	118.94	138.64	120.56	112.89	118.89
Blood pressure (diastolic)	73.91	68.17	71.28	81.50	69.33	65.11	64.00
White bloodcells (4-11)	15.10	12.21	10.31	16.66	12.72	9.82	15.51
Glucose serum ($\leq$100)	154.09	135.91	150.44	172.23	130.33	125.11	140.62
Platelet (150-450)	233.88	216.87	111.78	291.69	188.11	182.88	168.70
Prothrombin time (10-13)	18.09	15.87	16.47	15.04	20.53	16.29	20.21
Aminotransferase ($\leq$40)	189.72	90.70	47.67	109.79	1283.78	76.53	53.06
Total bilirubin ($\leq$2)	3.14	1.64	3.47	1.47	6.29	2.32	2.98
Anion gap (10-16)	17.40	14.24	14.00	16.00	19.44	13.46	18.49
Heart rate alarm (High)	128.87	122.69	123.06	126.17	130.56	123.74	124.47
Heart rate alarm (Low)	51.91	52.84	54.44	53.33	49.44	52.75	51.49
Earliest weight (kg)	84.21	75.67	85.37	86.43	85.76	73.65	80.11
Mortality rate	39.18	7.41	0.00	14.44	33.33	13.74	34.04

C Significance of features difference among clusters on digital phenotype cohort ($\alpha = 0.01$)

	DBSCAN	K-Prototype
Hematocrit serum	0.218	**0.000**
Blood urea nitrogen	**0.000**	**0.000**
Serum creatinine	**0.000**	**0.000**
Blood pressure (systolic)	0.441	**0.000**
Blood pressure (diastolic)	0.073	**0.000**
White blood cells	**0.005**	**0.000**
Glucose serum	0.087	**0.000**
Platelet	**0.001**	**0.000**
Aminotransferase	**0.004**	**0.000**
Total bilirubin	**0.000**	**0.000**
Anion gap	**0.000**	**0.000**
Heart rate alarm (High)	**0.000**	0.084
Heart rate alarm (Low)	0.161	0.101
Earliest weight (kg)	**0.003**	**0.000**

D Detailed characteristics of features for each clusters on ICD cohort

	DBSCAN			K-Prototype			
	-1	0	1	0	1	2	3
Alcoholism	0.22	0.21	0.00	0.16	0.25	0.11	0.33
Gender (%)	M (58.47)	M (57.56)	M (80)	F (52.15)	M (60)	M (64.29)	M (71.43)
Age	59.83	58.14	73.20	60.46	62.08	67.71	50.54
Hematocrit serum (35-45%)	35.14	33.47	36.76	31.53	31.51	34.42	38.83
Blood urea nitrogen ($\leq$20)	38.30	21.45	28.20	19.08	28.60	63.84	19.71
Serum creatinine ($\leq$1.71)	2.16	1.12	1.44	0.95	1.68	3.62	1.06
Blood pressure (systolic)	131.21	128.57	74.20	119.46	115.20	124.09	151.07
Blood pressure (diastolic)	74.45	73.88	43.20	68.23	61.65	63.73	91.94
Lipase ($\leq$60)	799.12	374.56	57.75	326.86	573.44	571.79	703.32
White blood cells (4-11)	14.29	12.87	15.66	12.36	12.50	13.82	14.98
Glucose serum ($\leq$100)	169.26	138.48	154.40	125.11	126.50	166.38	184.15
Platelet (150-450)	197.13	239.20	185.80	257.11	152.52	188.32	219.15
Prothrombin time (10-13)	18.49	14.78	15.92	14.82	24.71	15.81	14.55
Aminotransferase ($\leq$40)	259.52	108.38	184.20	99.34	615.25	101.48	108.44
Total bilirubin ($\leq$2)	3.10	1.80	1.70	1.63	6.89	1.45	1.82
Anion gap (10-16)	18.36	14.22	17.00	13.24	18.50	19.18	16.31
Heart rate alarm (High)	125.55	123.10	124.00	123.39	121.88	121.96	126.25
Heart rate alarm (Low)	53.98	53.75	54.00	53.06	51.62	53.84	55.85
Earliest weight (kg)	85.38	84.22	67.40	80.25	82.03	88.48	89.93
Mortality rate	40.68	2.95	100.00	11.83	30.00	28.57	9.82

E Significance of features difference among clusters on ICD cohort ($\alpha = 0.01$)

	DBSCAN	K-Prototype
Hematocrit serum	0.074	**0.000**
Blood urea nitrogen	**0.000**	**0.000**
Serum creatinine	**0.000**	**0.000**
Blood pressure (systolic)	**0.000**	**0.000**
Blood pressure (diastolic)	**0.002**	**0.000**
White bloodcells	0.170	**0.022**
Glucose serum	**0.001**	**0.000**
Platelet	0.012	**0.000**
Aminotransferase	**0.000**	**0.000**
Total bilirubin	**0.000**	**0.000**
Anion gap	**0.000**	**0.000**
Heart rate alarm (High)	0.037	**0.003**
Heart rate alarm (Low)	0.913	**0.000**
Earliest weight (kg)	0.217	**0.001**

F Project Management

The link to the GitHub repository is https://github.com/luluilmaknun/COMP90089. The repository contains directory as follows:

1. `dataset/`: This directory should contain the cohort dataset extracted from MIMIC-IV, you may not be able to see the files due to privacy constraint.
2. `exploratory/`: Jupyter notebooks for data exploration, preprocessing.
3. `predictive_analysis/`: Jupyter notebooks for predictive analysis, including feature engineering, modeling, and evaluation.
4. `clustering_analysis/`: Jupyter notebooks for clustering analysis, including feature engineering, modeling, and evaluation.
5. `output/`: Store the trained models, evaluation metrics, and visualizations generated during the analysis.

Our analysis runs on Jupyter Notebook. Prior to running, Jupyter Notebook should have been installed and ready to use. The project runs on Python `>3.9`. Make sure you have the following Python libraries installed:

```
 kmodes==0.12.2
matplotlib==3.8.0
nbformat==5.9.2
numpy==1.26.1
pandas==2.1.1
plotly-express==0.4.1
seaborn==0.13.0
scikit-learn==1.3.1
yellowbrick==1.5.0
```

You can install the dependencies using `pip`:

```
pip install -r requirements.txt
```