ML Final Project

By: Pablo O'Hop

Introduction

- Heart failure:
  - Significant mortality rates and predictive challenges
- Project goal:
  - Improve prediction accuracy using machine learning
- Methodology:
  - PCA + K-means
  - Logistic regression
- Data source:
  - Dataset from Kaggle

Dataset

- Utilized Pandas for data cleaning and preprocessing
- Applied StandardScaler from Scikit-learn for normalization
- Transformed categorical data using get_dummies
- Aim: Create a uniform feature scale for accurate analysis
-

A	B	C	D	E	F	G	H	I	J	K	L	M
age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	sex	smoking	time	DEATH_EVENT
75	0	582	0	20	1	265000	1.9	130	1	0	4	1
55	0	7861	0	38	0	263358.03	1.1	136	1	0	6	1
65	0	146	0	20	0	162000	1.3	129	1	1	7	1
50	1	111	0	20	0	210000	1.9	137	1	0	7	1
65	1	160	1	20	0	327000	2.7	116	0	0	8	1
90	1	47	0	40	1	204000	2.1	132	1	1	8	1
75	1	246	0	15	0	127000	1.2	137	1	0	10	1
60	1	315	1	60	0	454000	1.1	131	1	1	10	1
65	0	157	0	65	0	263358.03	1.5	138	0	0	10	1
80	1	123	0	35	1	388000	9.4	133	1	1	10	1
75	1	81	0	38	1	368000	4	131	1	1	10	1
62	0	231	0	25	1	253000	0.9	140	1	1	10	1
45	1	981	0	30	0	136000	1.1	137	1	0	11	1
50	1	168	0	38	1	276000	1.1	137	1	0	11	1
49	1	80	0	30	1	427000	1	138	0	0	12	0
82	1	379	0	50	0	47000	1.3	136	1	0	13	1
87	1	149	0	38	0	262000	0.9	140	1	0	14	1
45	0	582	0	14	0	166000	0.8	127	1	0	14	1
70	1	125	0	25	1	237000	1	140	0	0	15	1
48	1	582	1	55	0	87000	1.9	121	0	0	15	1
65	1	52	0	25	1	276000	1.3	137	0	0	16	0
65	1	128	1	30	1	297000	1.6	136	0	0	20	1
68	1	220	0	35	1	289000	0.9	140	1	1	20	1
53	0	63	1	60	0	368000	0.8	135	1	0	22	0

Figure 1: First 25 Rows of Dataset

Unsupervised Analysis: PCA + K-Means Cluster

- Applied PCA for dimensionality reduction
- PCA:
  - No clear separation in survival outcomes identified
- Implemented K-Means to detect patient clusters
- K-Means:
  - Revealed overlapping groups, underscoring heart failure complexity

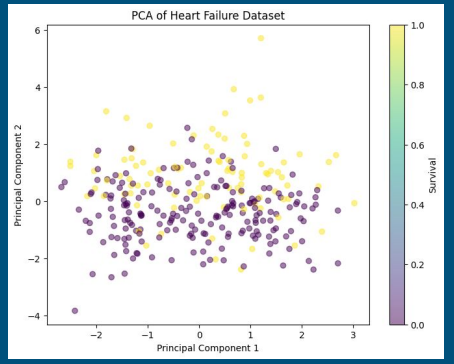

Figure 2: PCA Scatter Plot

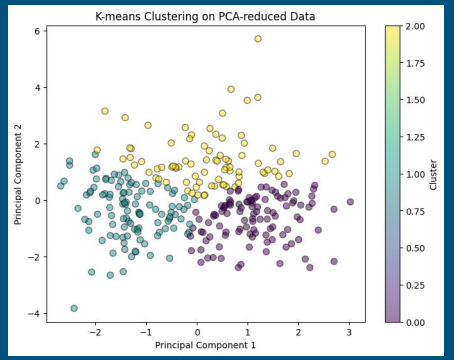

Figure 3: K-Means Cluster Scatter Plot

Supervised Analysis: Logistic Regression

- Developed three logistic regression models:
  - No Regularization:
    - C=1e9 (approx. no reg.)
  - Ridge (λ=0.1):
    - C=10 (mild reg.)
  - Ridge (λ=1):
    - C=1 (strong reg.)
- Focused on prediction of patient survival

Logistic Regression: No Regularization

- Training Accuracy: 87.45%
- Validation Accuracy: 80%
- Precision: 93.33%
- Recall: 56%
- F1 Score: 70%

	Actual Positive	Actual Negative
Predicted Positive	14 (TP)	1 (FP)
Predicted Negative	11(FN)	34 (TN)

Table 1: Confusion Matrix Model 1

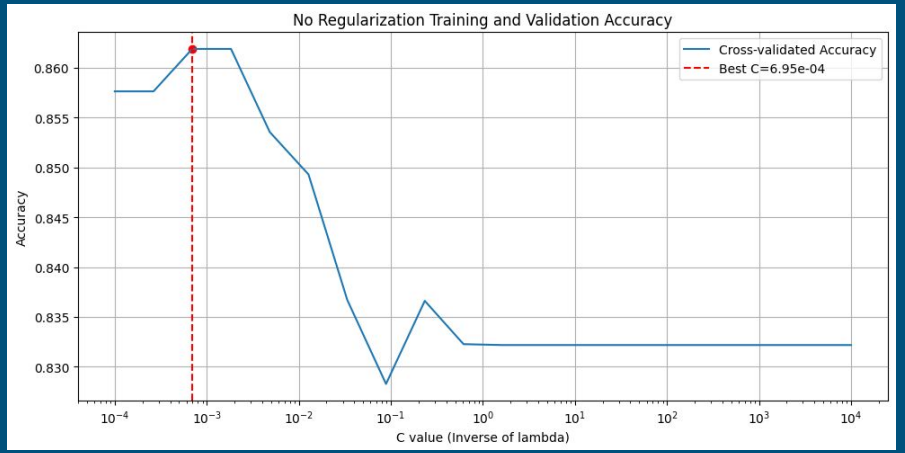

Figure 4: Optimal C Value for Logistic Regression Model 1

Logistic Regression: Ridge (λ=0.1)

- Training Accuracy: 87.45%
- Validation Accuracy: 80%
- Precision: 93.33%
- Recall: 56%
- F1 Score: 70%

	Actual Positive	Actual Negative
Predicted Positive	14 (TP)	1 (FP)
Predicted Negative	11(FN)	34 (TN)

Table 2: Confusion Matrix Model 2

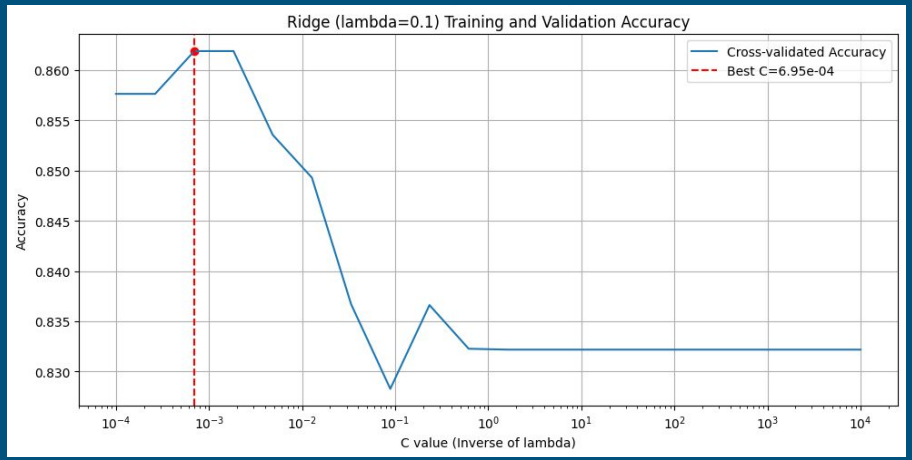

Figure 5: Optimal C Value for Logistic Regression Model 2

Logistic Regression: Ridge (λ=1)

- Training Accuracy: 87.45%
- Validation Accuracy: 81.67%
- Precision: 93.75%
- Recall: 60%
- F1 Score: 73.17%

	Actual Positive	Actual Negative
Predicted Positive	15 (TP)	1 (FP)
Predicted Negative	10(FN)	34 (TN)

Table 3: Confusion Matrix Model 3

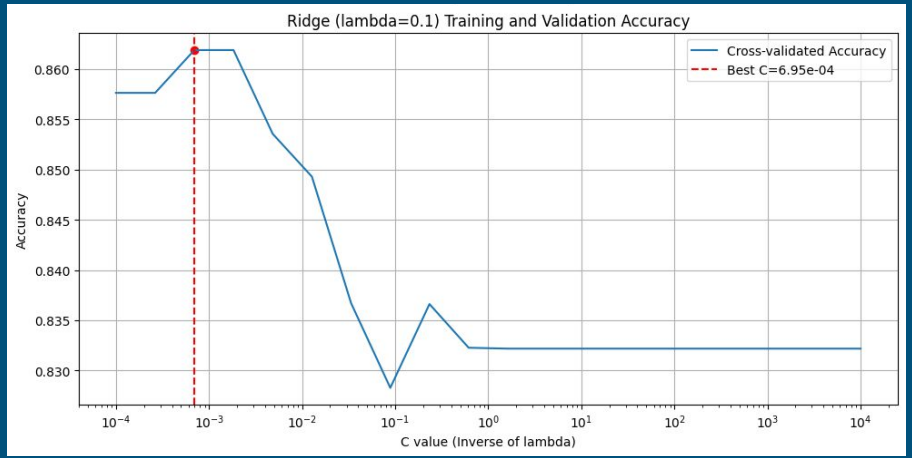

Figure 6: Optimal C Value for Logistic Regression Model 3

LR Model #	**λ**	**Training Accuracy**	**Validation Accuracy**	**Precision**	**Recall**	**F1 Score**
Model 1	N/A	0.8745	0.8	0.9333	0.56	0.7
Model 2	0.1	0.8745	0.8	0.9333	0.56	0.7
Model 3	1	0.8745	0.8167	0.9375	0.6	0.7317

Table 4: Logistic Regression Results

Conclusion

- Concluded Ridge (λ=1) as the most balanced model
  - Improved recall with stronger regularization
  - Cross-validated accuracy peaks
- Importance:
  - Provides predictive tool for medical decision-making

Figure 7: A Machine Learning