Heart Failure Survival - Final Project

CS 4563 - Linda Sellie

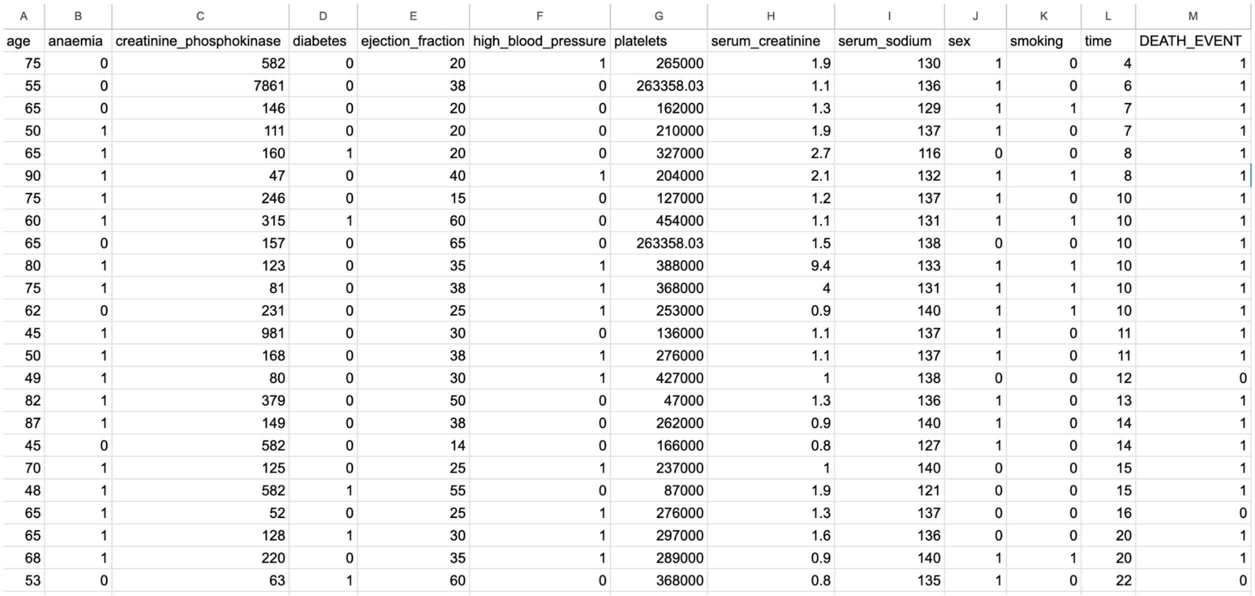
Pablo O’Hop

po2038

Introduction:

Heart failure, one of the main leading causes of death worldwide, presents significant challenges in predicting patient outcomes, especially due to its complex etiology and natural progression. This project seeks to use machine learning to enhance the predictive accuracy of survival outcomes in patients who experienced heart failure. The goal of the project is to utilize a dataset filled with clinical data about heart failure patients and predict the survival rates of these patients, using both unsupervised and supervised learning models. The analysis will include the use of PCA, K-means clustering, and logistic regression, using a dataset found on Kaggle (found in the bibliography section on the last page).

Dataset:



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

The dataset comprises critical clinical markers such as age, blood pressure, serum creatine levels, and ejection fraction among others. In preparing the dataset for this project, data cleaning and pre-processing were employed using the Panda and Sklearn libraries. Since all the data was relevant to the goal of predicting survival from heart failure, the data required standardization to ensure that each variable contributed equally to the analysis, thereby avoiding any undue influence due to scale differences. This involved using StandardScaler (from sklearn) to normalize continuous variables to a standard Gaussian distribution. Afterward, categorical variables were encoded into a numerical format using Panda's "get_dummies" function. Missing values were addressed by employing imputation techniques, like mean imputation or median imputation, which was determined based on the distribution of each variable, with outliers handled by median replacement. Exploratory analysis involves examining the distribution of key variables and their correlation with the target outcome, survival.

Unsupervised Analysis:

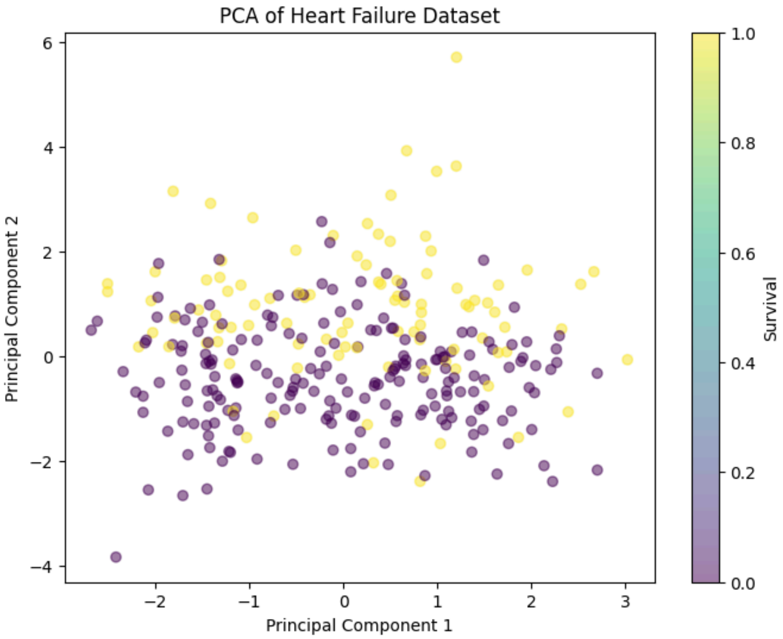


Figure 2: PCA Scatter Plot

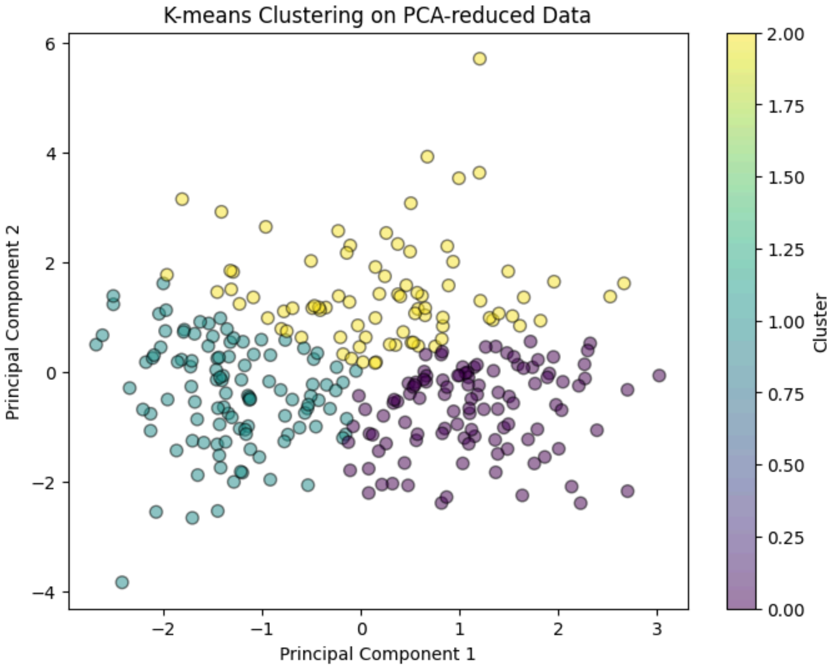


Figure 3: K-Means Cluster Scatter Plot

For the unsupervised analysis, PCA was used to reduce the dataset's dimensionality, thus simplifying the complex data into principle components that capture the most significant variance. Since the dataset contained so many variables, PCA was crucial for transforming these variables into a smaller set of uncorrelated components, which enabled a more efficient exploration and visualization of the data. The data was represented in two dimensions, which contained data on a binary scale, specifically the yellow dots indicating that the patient survived (represented with a 1) and the purple dots indicating that the patient passed away (represented with a 0). There were a few key takeaways from the PCA gradient, specifically that there was no clear and distinct separation between survival outcomes, which suggests that survival cannot be predicted by these two components alone or that the survival patterns were extremely complex and non-linear. K-means clustering was also used to explore potential patterns or grouping within the data. This method was chosen because it could potentially reveal distinct patient groups based on similarities in their clinical profiles, which might correspond to different survival risks.

The light blue and yellow clusters appeared to intermingle somewhat, which can imply that while the algorithm has found a way to differentiate the data points to some extent, there still are similarities between these groups that the two principle components do not fully capture. This outcome emphasized the complexity of heart failure as a medical condition, as well as highlighting the challenge of using unsupervised methods alone to delineate patient outcomes based on the available features.

Supervised Analysis:

Building on the insights gleaned from the unsupervised analysis, the project transitioned into a supervised learning analysis using logistic regression, implemented through sklearn logistic regression functions. Logistic regression was used in this project due to its efficiency in handling binary classification tasks, making it ideal for modeling the probabilistic outcomes of patient survival. Three different logistic regression models were created in order to understand how varying degrees of regularization influence the model's performance in predicting heart failure survival outcomes. The three different models were as follows:

1. No Regularization Model
    a. This model was aimed at understanding the baseline performance where the learning algorithm focuses solely on minimizing the loss without any penalty on the feature weights. This was done by setting C to an extremely high value, essentially removing regularization from the equation.
2. Ridge Regularization ($\lambda = 0.1$)
    a. This small level of regularization was implemented to slightly penalize the magnitude of the coefficients, which helps reduce the risk of overfitting

while maintaining a low bias. In this model, C = 10 which was the inverse of lambda. This was done to strike a balance between maintaining model complexity and controlling overfitting.

3. Ridge Regularization (λ = 1)
    a. In contrast to the previous regularization, this setup involved a higher level of regularization with a lambda value of 1, meaning that C = 1 as well. This increases the penalty on the size of coefficients, which creates a simpler model with lower variance. This model was expected to demonstrate improved generalization capabilities at the potential cost of increased bias, making it better for new or unseen data but less flexible in fitting data intricately.

Each model's training involved fitting the logistic regression algorithm to the dataset, which included splitting the data into training and testing subsets to evaluate model performance in an objective manner. The logistic regression was assessed based on several key performance metrics: accuracy, precision, recall, and F1 score, all of which were implemented using sklearns metric functions.

Results:

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 1: Logistic Regression Results

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

Table 2: Confusion Matrix Model 1

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

Table 2: Confusion Matrix Model 2

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

Table 2: Confusion Matrix Model 3

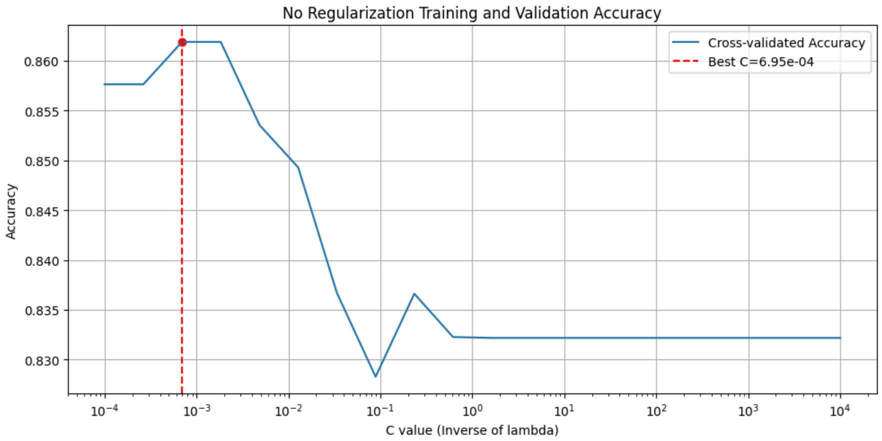


Figure 4: Optimal C Value for Logistic Regression Model 1

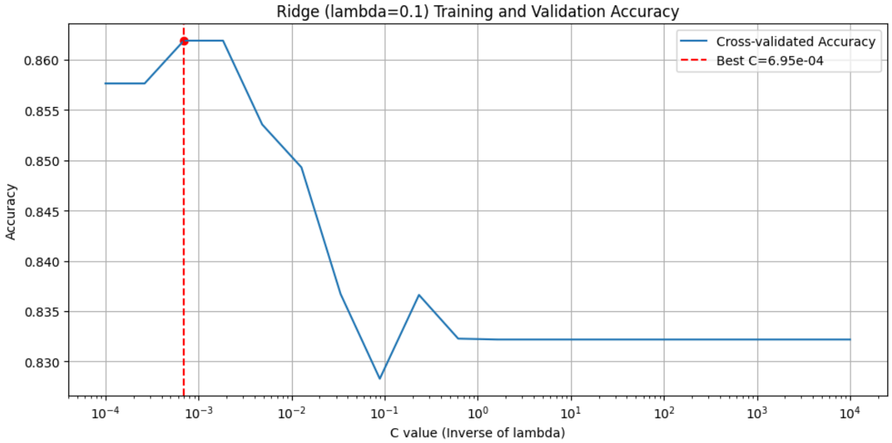


Figure 5: Optimal C Value for Logistic Regression Model 2

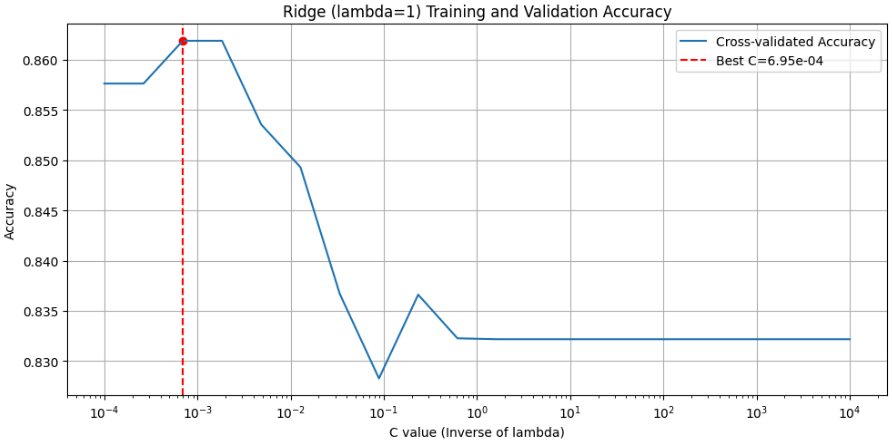


Figure 6: Optimal C Value for Logistic Regression Model 3

Discussion:

The logistic regression model that utilized no regularization and the second model, which applied ridge regularization of λ = 0.1 both achieved a training accuracy of 87.45% and a validation accuracy of 80%. These models demonstrated high precision as well, with 93.33%, but also a relatively low recall of 56%, which indicates a potential overfitting where the models will perform well on training data, but less effectively on unseen data. The F1 scores of 70% for these models suggest that while they are reliable in identifying death cases, they lack sensitivity in detecting positive cases for survival.

The third model, using a stronger ridge regularization λ = 1 showed a definite improvement. Validation accuracy increased to 81.67% and the model obtained a better recall of 60%. The precision increased slightly to 93.75% and the F1 scores increased to 73.17%. This model's performance underscores the effectiveness of increased regularization in reducing overfitting, thus enhancing the model's ability to generalize to new data without having to necessarily sacrifice accuracy. Comparing the two confusion matrices as well, the number of true negatives and false negatives stayed the same, whereas the number of true positives went up slightly and the number of false positives went down. This indicates that the third model had a better balance between true positives and false negatives, which is crucial in medical applications where missing out on true positives could have deathly consequences. The progression from no regularization to a higher degree of regularization illustrates the significant shift toward reducing model overfitting. I also created 3 graphs that showed the optimal C value at which the model performed best before starting to overfit the training data. The graphs for all models peaked at the same C value, which denotes the most effective trade-off between bias and variance achieved through regularization. These models show that for ridge models, especially with $\lambda = 1$, the regularization helps maintain high validation accuracy across a range of C values. It also seems to show that the introduction of regularization seems to smooth out the variability in accuracy across different C values.

Conclusion:

This project experimented with different levels of regularization, demonstrating that appropriate tuning could substantially enhance model performance, especially in terms of reliability. The third model emerged as the most effective, suggesting that higher regularization may be necessary to develop better predictive models in healthcare. In conclusion, the logistic regression analysis demonstrated the importance of regularization in building predictive models that are not only accurate but also reliable in varying medical conditions and settings.

Bibliography

Kharoua, Rabie El. "Predict Survival of Patients with Heart Failure." *Kaggle*, 25 Apr. 2024, www.kaggle.com/datasets/rabieelkharoua/predict-survival-of-patients-with-heart-failure.