DATA 2204: Discriminant Analysis – Assignment #3

Name: Raj Dholakia

Student ID: 100813041

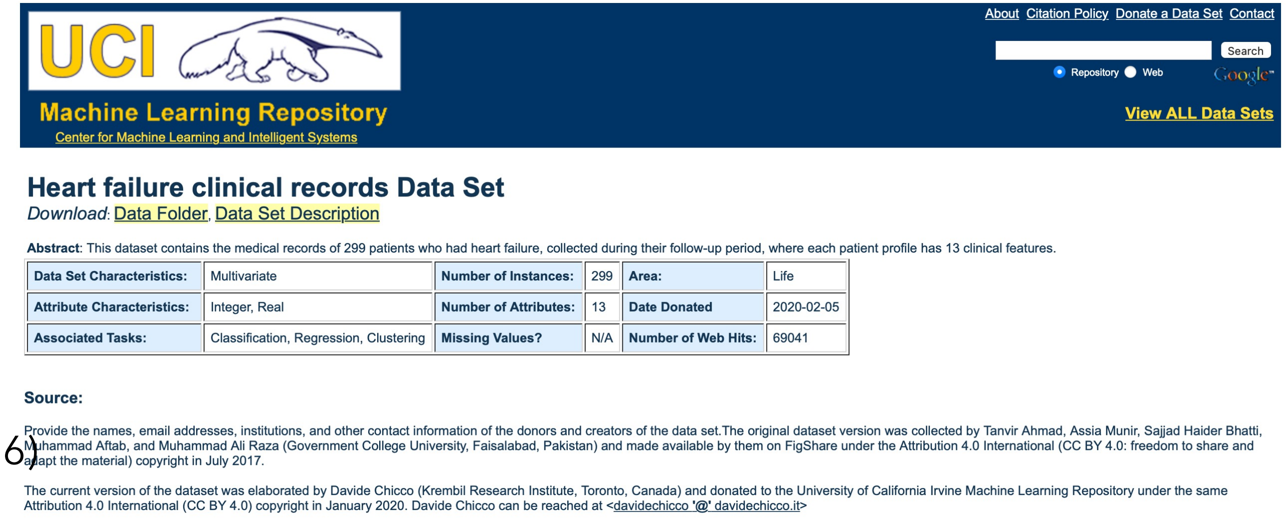
Table of Contents

Analysis Statement

Develop an **LDA model** and evaluate its performance in comparison with a Logistical Regression model with SMOTE for the heart failure dataset and carry out additional analysis to better understand the **LDA Model**.

Key Insights from the Pandas Profiling Report

- The dataset consists of twelve (12) independent variables and one dependent – ‘DEATH_EVENT’.
- The twelve (12) independent variables are made up of six (6) numerical (NUM) variables and six (6) categorical (BOOL) variables – hence a mixed dataset.
- Only three (3) out of the six (6) continuous variables have a normal distribution. The other three (3) have skewed or uniform distributions.
- There are a total of **299 samples** in the dataset (a small dataset) and is and contains **no missing values**.
- There is a **large imbalance in the dataset** as the number of deceased patients examples are 96, while the number of alive patients are 203.
- From the correlation plot, it is evident that there are **very weak or no correlations between variables** in the dataset. One exception could be a slight correlation between the variables `time` and `DEATH_EVENT`.


About Citation Policy Donate a Data Set Contact

UCI Machine Learning Repository

Center for Machine Learning and Intelligent Systems

Search

Repository Web

View ALL Data Sets

Heart failure clinical records Data Set

Download: Data Folder, Data Set Description

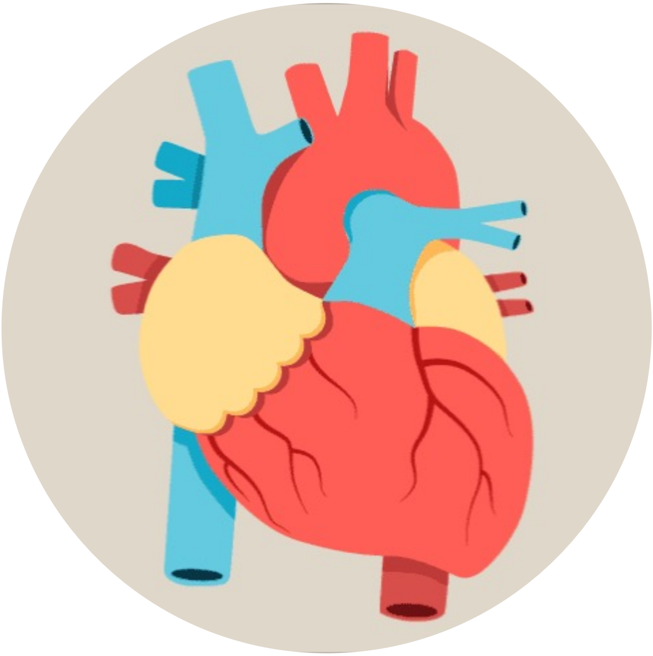
Abstract: This dataset contains the medical records of 299 patients who had heart failure, collected during their follow-up period, where each patient profile has 13 clinical features.

Data Set Characteristics:	Multivariate	**Number of Instances:**	299	**Area:**	Life
Attribute Characteristics:	Integer, Real	**Number of Attributes:**	13	**Date Donated**	2020-02-05
Associated Tasks:	Classification, Regression, Clustering	**Missing Values?**	N/A	**Number of Web Hits:**	69041

Source:

Provide the names, email addresses, institutions, and other contact information of the donors and creators of the data set.The original dataset version was collected by Tanvir Ahmad, Assia Munir, Sajjad Haider Bhatti, Muhammad Aftab, and Muhammad Ali Raza (Government College University, Faisalabad, Pakistan) and made available by them on FigShare under the Attribution 4.0 International (CC BY 4.0: freedom to share and adapt the material) copyright in July 2017.

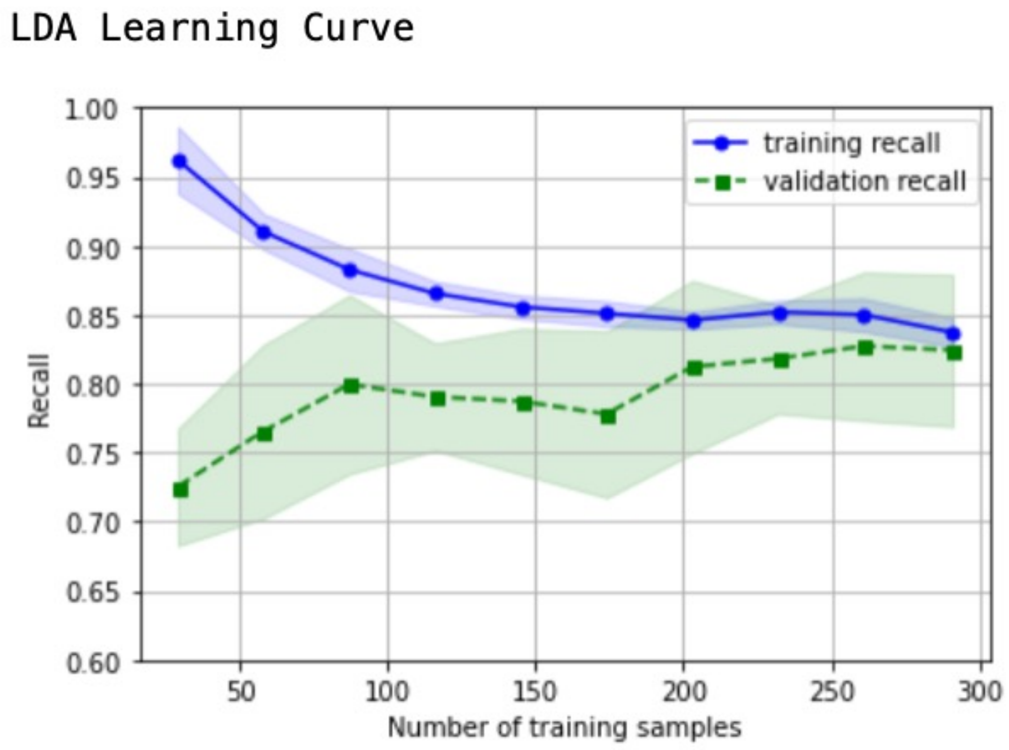
The current version of the dataset was elaborated by Davide Chicco (Krembil Research Institute, Toronto, Canada) and donated to the University of California Irvine Machine Learning Repository under the same Attribution 4.0 International (CC BY 4.0) copyright in January 2020. Davide Chicco can be reached at <davidechicco '@' davidechicco.it>



Understanding the Learning Curve

Some insights gained from the Learning Curve of the Logistical Regression Model:

1. The learning curve seems to be flattening after 250 samples, highlighting the possibility that more samples might not improve the performance of the model.
2. The model has a **very low variance** when trained on 300 samples.
3. The performance of the model is in the acceptable region of 80-85% recall. The model performance can be can work as a base model for development of other resources. This also shows that the bias in the model is at an acceptable level as well.



Standard vs Optimized LDA Models

Standard LDA Model:

- The model has achieved an overall **F1 score of 80%**. As the dataset is imbalanced, the overall F1 score does not provide a holistic understanding of the model's performance.
- Considering the model's performance for each class individually, we observe a F1 score of **85%** for positive (deceased) and **70%** for negative (alive) samples (patients).
- There is huge discrepancy (**difference of 15%** in F1 scores) in the model's performance when predicting the two different classes. One of the reasons for this discrepancy can be the **imbalance in the data** (mentioned in Key Features section). However, the use of SMOTE to balance the training dataset has reduced the effects of the imbalance.
- The model has achieved a below average precision of 67% when predicting a positive outcome (deceased), while an outstanding 87% when predicting a negative outcome (alive).

Comparing both the models:

- Both the models have achieved the same results for **all** the metrics – there is no difference in the model performances.
- The identical performance can be underpinned from the confusion matrix, which has the same number at each corner for both the models.

```
Estimator: LDA
[[34  7]
 [ 5 14]]
              precision    recall  f1-score   support

           0       0.87      0.83      0.85        41
           1       0.67      0.74      0.70        19

    accuracy                           0.80        60
   macro avg       0.77      0.78      0.78        60
weighted avg       0.81      0.80      0.80        60
```

```
Optimized Model

Model Name: LinearDiscriminantAnalysis()

Best Parameters: {'clf__solver': 'svd'}

 [[34  7]
 [ 5 14]]
              precision    recall  f1-score   support

   Outcome 0       0.87      0.83      0.85        41
   Outcome 1       0.67      0.74      0.70        19

    accuracy                           0.80        60
   macro avg       0.77      0.78      0.78        60
weighted avg       0.81      0.80      0.80        60
```

Optimized LDA vs Optimized Logistical Regression Models

Comparison:

- The logistical regression model with a F1 score of **83%** performs slightly better than the LDA model that achieved an F1 score of **80%**.
- Comparing performance of the models at predicting each of the outcomes also shows logistical regression model has a better F1 scores (88% for Outcome 0 and 74% for Outcome 1) than the LDA model (85% for Outcome 0 and 70% for Outcome 1).

Reasoning:

- One of the major reasons for logistical regression performing better is the fact that the dataset used is a mixed dataset: contains continuous and categorical variables. To add to it, half of the independent variables are categorical and half are numerical. Hence, logistical regressions should outperform the LDA model significantly.
- However, there logistical regression is only slightly better than the LDA. This is because all the variables in the dataset have the same covariance – supporting one of the assumptions for the LDA model. This also eliminates the possibility of a QDA model outperforming the LDA.
- Finally, if all the continuous variables in the dataset had a normal distribution, the LDA model would have a similar performance to that of the logistical regression model.

```
Optimized Model

Model Name: LinearDiscriminantAnalysis()

Best Parameters: {'clf__solver': 'svd'}

 [[34  7]
 [ 5 14]]

              precision    recall  f1-score   support

   Outcome 0       0.87      0.83      0.85        41
   Outcome 1       0.67      0.74      0.70        19

    accuracy                           0.80        60
   macro avg       0.77      0.78      0.78        60
weighted avg       0.81      0.80      0.80        60
```

```
Optimized Model

Model Name: LogisticRegression(class_weight='balanced', random_state=100)

Best Parameters: {'clf__C': 0.01, 'clf__penalty': 'l2'}

[[36  5]
[ 5 14]]

precision      recall  f1-score   support

Outcome 0      0.88      0.88      0.88      41
Outcome 1      0.74      0.74      0.74      19

accuracy                           0.83      60
macro avg      0.81      0.81      0.81      60
weighted avg   0.83      0.83      0.83      60
```

What should be the next steps?

Mr. John Hughes can take into consideration the following options to better predict if a patient is deceased during the follow-up period ('DEATH_EVENT') using a LDA model:

1. **Improving dataset structure**: though SMOTE is being used to balance the dataset, synthetically improving the balance in the class samples does improve the model's performance, but it can be better with a balanced dataset.
2. **Using more complex algorithms**: deep learning neural nets can perform well with imbalanced datasets, such as this one. Using a more complex algorithm can improve the results obtained, however, there is high chance of overfitting the dataset.
3. **Feature selection**: doing feature selection can reduce the noise for the model, improving its performance.

Thank you

Raj Dholakia

raj.dholakia@dcmail.ca

Student ID: 100813041