## Data Science Clinical Data Analysis

## Cardiovascular Heart Failure Features Analysis

Matthew Yeseta, April 23, 2021

1. **Introduction to Dataset Analysis for Cardiovascular heart failures**

The problem area is health and the heart. The critical issue for any analysis to be performed can elevate awareness in data behavior and to identify health measures that are notable for solutions to this problem by finding patterns that may help to stem the issue associated with worldwide risk of cardiovascular diseases and cardiovascular heart failure.

In the health industry, serum creatinine and ejection fraction measurements are widely known to be leading indicators for predicting outcome following acute kidney injury cardiac surgery, and the high risk of heart failure or cardiomyopathy. Analysis visualization for these two measurements shall be ascertained for their impact on cardiovascular patients, together with a prediction model to evaluate sufficient accuracy for a future machine learning model for predicting cardiovascular heart risks.

1. **Background Issues**

Since Cardiovascular heart failures effect an estimated 17.9 million lives each year, which unfortunately represents approximately 31% of deaths worldwide, this is a recipe for disaster if people do not take heed to the health issues of heart failure and the documentable risky behavior that attributes to heart failures. The outcome of people survival is at stake. Analysis visualization can aid in revealing the importance and scope of this pending issue of heart failure in anyone that exhibits or is associated with any behavior or health indicators (Serum Creatinine, Ejection Fraction)

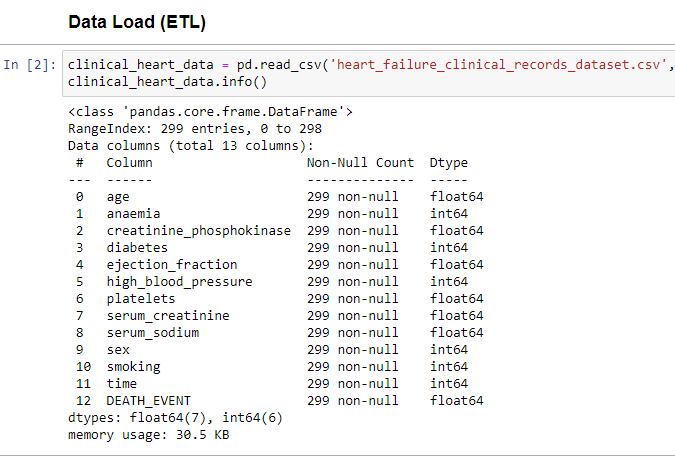
1. **Methodology of Study**

The approach to bring light to this analysis for cardiovascular heart failure risks is a multi-process deployment which includes design and implementation to build a data pipeline and to run visualizations and in the end an evaluation of accuracy scores and cross validation models from a machine learning predictive model.

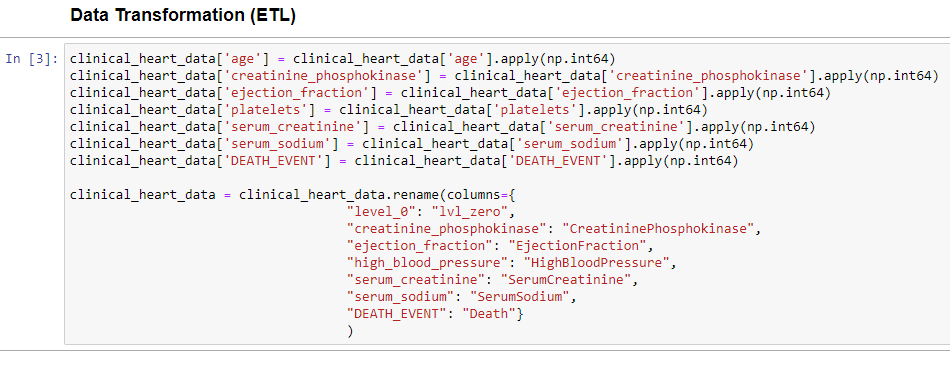
Data. First, a suitable big data store for the collection and store of Cardiovascular heart failures data for current and future repository to offer visibility to address the health and behavior problem. Since this data is critical to the survival of individuals, in aggregate statistics, it must therefore be a flexible repository store that can offer flexible unstructured data store. Therefore, a Document data store structure was selected, for it supports this list of requirements for storage and format options.

Data Pipeline. Secondly, an appropriate data pipeline approach was necessary to design and implement for capture / ingest the data. This pipeline must also be flexible enough to be able to support data feeds from future streamed senor data collection devices or batch data loads from server sources.

Data Pipeline Extract. It was decided that the data pipeline shall be developed to interface with MongoDB using the publicly available Python-Mongo library: ‘pymongo’. Therefore, for the data pipeline affectionally known in developer circles, the: ETL Extract. This implementation for the ETL ‘extract/download/read’ was a designed as a simple python read extract from file formatted dataset source. This Pipeline permitted future replacement in one module routine for building suitable streaming data functionality to be added. The target destination for this extract to load, in following steps, was a MongoDB instance.



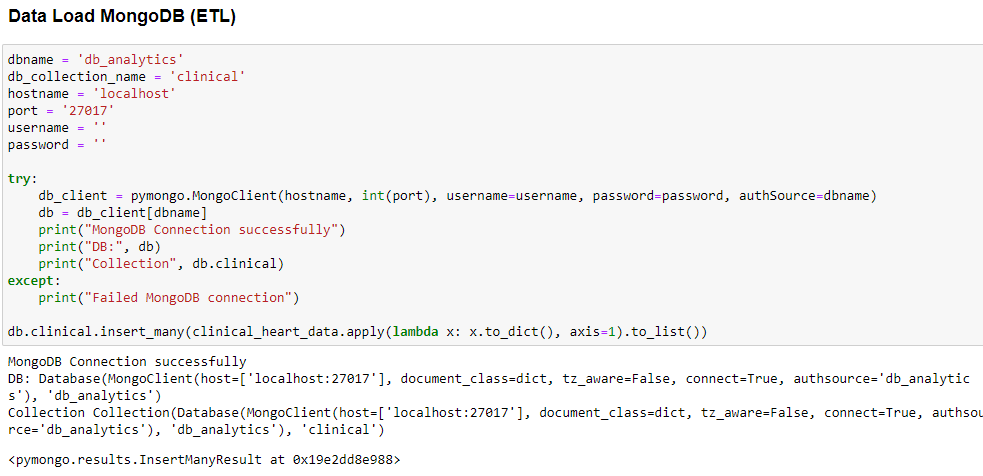
Data Pipeline Transform. The ETL ‘data transformation’ was implemented to provide and ensure that data features were consistent in data types to be more fluid in data type treatment in processing and data visualization. The data transformation implemented a few small conversion routines (python) for the express purpose to cast data types from real numbers in dataset source; and casting them to integer number data types. Specifically, the features that had data type casting included the features that would be reported on in the visualizations.



Storage Model. Thirdly, a storage model for the Cardiovascular heart failure features data was implemented with freely available storage model. The technical storage architecture for this Cardiovascular data storage system design in purpose to be implemented with a Big Data Document database to store Cardiovascular heart failure features data for processing and analysis. The Document data store permits scalability for unstructured big data storage, which can offer future batch computing for analysis of unstructured and structured data. Additionally, a storage model can be fulfilled in a stream computing model for stream analysis of unstructured and structured data in real-time views of new data points from system sensor monitored Cardiovascular heart failure data.

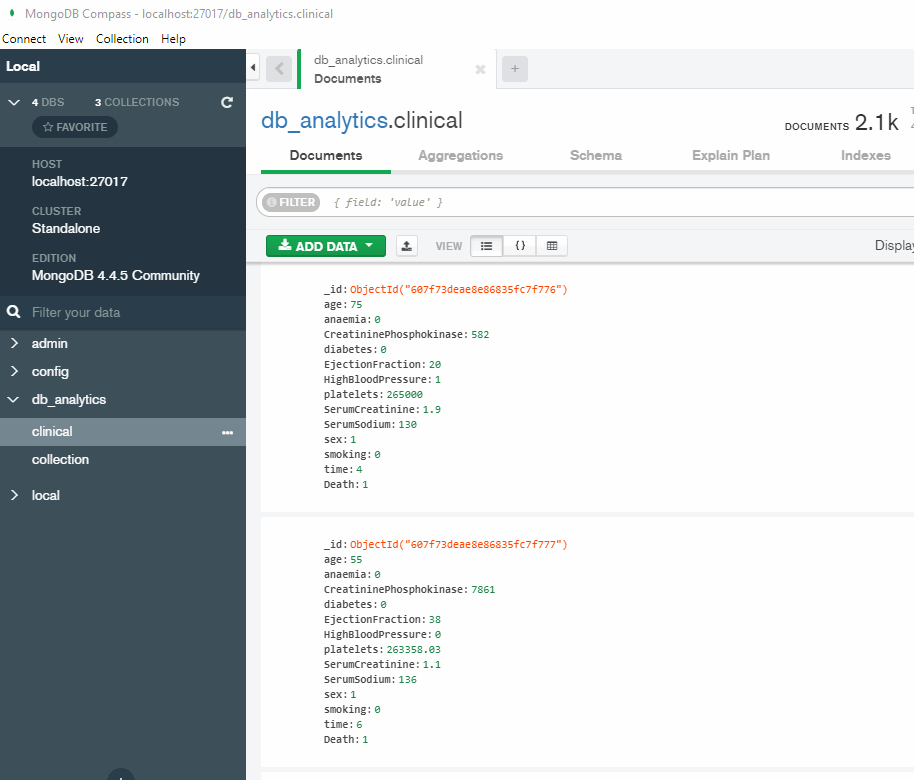
The storage model database was instantiated by creating a database, namely the MongoDB database, ‘db\_analytics’; the collection model was instantiated to be deployed inside the ‘db\_analytics’ database, appropriately names: clinical.

Data Pipeline Load. The associated ETL ‘load’ process shall utilize the storage model, MongoDB database, for loading the Cardiovascular heart failure features data. The target database collection in this MongoDB is db\_analytics.clinical, which was implemented using the MongoDB data store.



In order to perform loading this target storage using the ETL ‘load’ data pipeline process, a Python program script shall connect to MongoDB using the Python-Mongo library: ‘pymongo’ API. An instance of the database is instantiated. Following with the instantiation of the MongoClient, which connects to the MongoDB collection, named “clinical’. A fast single line script issued the entire load in a MongoDB insert function, named ‘insert\_many()’. This function was scripted inside a single apply() function, which acts as a map() function for loading big data instances to an entire data destination, namely the db\_analytics.clinical Document collection. Together in this load function was scripted for all rows to convert each row to a JSON (BSON) format for posting to the MongoDB db\_analytics.clinical Document collection. This was accomplished within the apply() function with a lambda anonymous function to convert values from the Cardiovascular heart failure into a dictionary format to satisfy the JSON (BSON) format of MongoDB.

The MongoDB Compass db\_analyitcs.clinical Document Big Data Database.



1. **Data Pipeline and Analysis Results**

The results computed from our Exploratory Data Analysis (EDA) statistical charts give succinct evidence that when analyzing charts on Cardiovascular heart failure data, such as age/time, density of age, age by gender, age by non-survival (death), serum creatinine, ejection fraction, that that later two Cardiovascular features manifest more evidence to provide more accurate predictions, over and above the balance of the other Cardiovascular dataset features. This concurs with the clinical industry knowledgebase that serum creatinine and ejection fraction are the two most singular clinical features that most adequately evaluate the survival of people with Cardiovascular symptoms.

This Exploratory Data Analysis proved to be an essential technique for analysis patterns that would aid in forecast analytics. This set of analysis charts did reveal patterns of survival or loss to Cardiovascular heart failure.

The outcome analysis from the Data Pipeline process focus’s principally on exploratory data analysis and histogram distribution charts for individual cardiovascular heart failure feature data, namely age/time, density of age, age by gender, age by non-survival (death), serum creatinine, ejection fraction. These bar charts were group aligned to view a summary perspective of the health risk indicators more easily. Further analysis was conducted on some of the categorical nominal univariate features and two of the continuous univariate features which were charted in bar charts. A box plot / swarm plot analysis was plotted for Gender by Age Distribution.

Visible analysis from these bar chart plots it is evident in the revelation of a high risk in Serum Creatinine levels at 61.21% for male replicants. The next prof progressive analysis bar plot was for the entire age range 25-60 in Ejection Fraction measurements. These two levels / measurements: namely, Serum Creatinine and Ejection Fraction which are two measures that provide doctors with very accurate indicators for patient high risk associated with heart failure or cardiomyopathy. These levels of serum creatinine and the other measurements of ejection fraction are two critical positive indicators for high risk of patient heart failure or cardiomyopathy. The other health risk indicators, namely, smoking, anaemia, diabetes, gender sex, do focus primarily on which gender is at most risk for the various behavior practices. The overall summary of at most risk gender is the female with higher percentage ranking in smoking, anaemia, diabetes, and high blood pressure.

Chart, timeline, bar chart

Description automatically generated

The highlight of this Cardiovascular heart failure analysis results in the manifestation of the high risk of various behavior practices and more important for diagnosis of this risk, the critical important levels of serum creatinine and the measurements of ejection fraction. The following two succinct pie charts for this data population shows the population survival rate of patients evaluated with positive / negative serum creatinine level measurements and the population survival rate of patients evaluated with positive / negative for ejection fraction measurements. The result of this analysis shows that the levels of serum creatinine and the measurements of ejection fraction are two of the critical feature data for positive or negative indicators for high risk of patient heart failure or cardiomyopathy.

|  |  |
| --- | --- |
| serum creatinine population analysis | ejection fraction population analysis |
| Chart, pie chart  Description automatically generated | Chart, pie chart  Description automatically generated |

The density distribution charts support the pie charts for population risk with risky serum creatinine levels and ejection fraction levels. Serum Creatinine levels are critical between level when above 1.04 for woman and above 1.55 for men. Higher in this measure it is critical for high-risk heart failure, particularly for woman with extremely high density of heart failure at around level 1.05+; and for men expanding well above the normal high of 1.55, spanning into Serum Creatinine levels of 2-5 and even some between 8-10. Men has normal range for serum creatinine for adult men, 0.74 to 1.35 mg/dL; for adult women, 0.59 to 1.04 mg/dL.

Chart, histogram

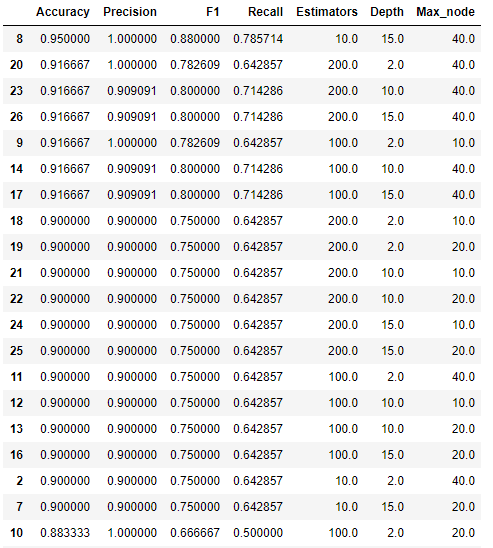
Description automatically generated

Ejection Fraction measures are sufficiently high risk as measurement ranges below 50%; particularly high risk for both women and men. The normal percent level is 55% to 70%. Woman who has measurement levels below 50%, are particularly high risk. The woman's density chart range shows high risk between 20% and 45%. Men who have measurement levels below 50%, are particularly high risk. The men’s density chart range show high risk between 10% and 45%; a very wide deficient range.

Chart, histogram

Description automatically generated

As a final test, the results of the code execution of a Machine Learning Random Forest Accuracy Model (RandomForestClassifier model) which provided very nice accuracy results for reporting predictive accuracy scores for this dataset in the 90th percentile range, in addition, precision reported scores in the in the 90th percentile. The Machine Learning prediction model reported highest accuracy at 95% with precision score at 100%. See table for results.



1. **Visualizations & Interpretations**

The data pipeline process that was learned from this class prepared well for developing this data pipeline solution. In fact, these tools made it easy to develop a simple implementation complete lifecycle for MongoDB and for the data pipeline extract, transform, load using the Python-Mongo library: ‘pymongo’. This basic pipeline was implemented using the Mongo library and created an instance of the database as a data store structure. Using Pymongo was swift in creating a single line script with the MongoDB (insert\_many) to load data into MongoDB instance. The barrier challenge encountered was my computer and its hardware limitations on connection to remote VM was an issue, it is now evident that my PC WIFI component is beginning to fail as I experience significant failed VM operational connections during work session.

Chart, pie chart

Description automatically generated

A detailed analysis and critical meaning regarding the Creatinine measurements population study data shows that the fully 69.3% of people are positive for Serum Creatinine heart failure risk, while 30.7% of the population has a negative Serum Creatinine measurement. Visible in the "Survival Rate Breakdown of Serum Creatinine' shows that for the Creatinine positive population (27.3%), are positive and do succumb to this heart failure condition. Moreover, 24.36% of population are positive for Creatinine and do not survive. Furthermore, a slim 3.41% of the population are negative for Creatinine and do not survive the acute kidney Creatinine condition. Furthermore, a full 44.7% of the population are positive for Creatinine and do survive this high-risk condition.

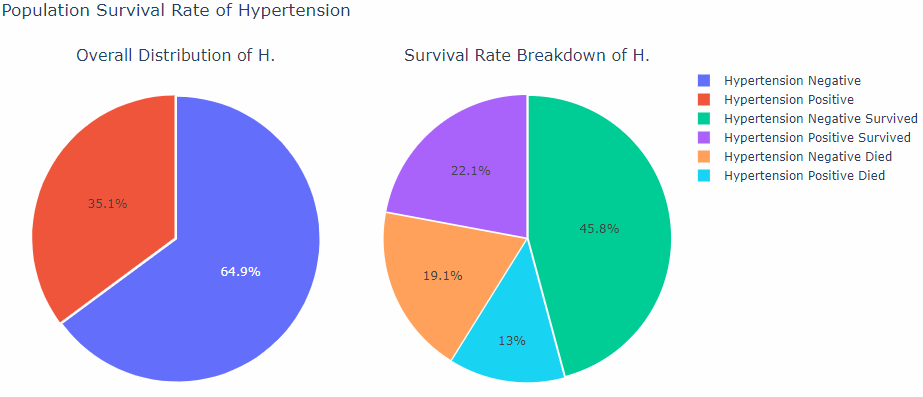
Chart, pie chart

Description automatically generated

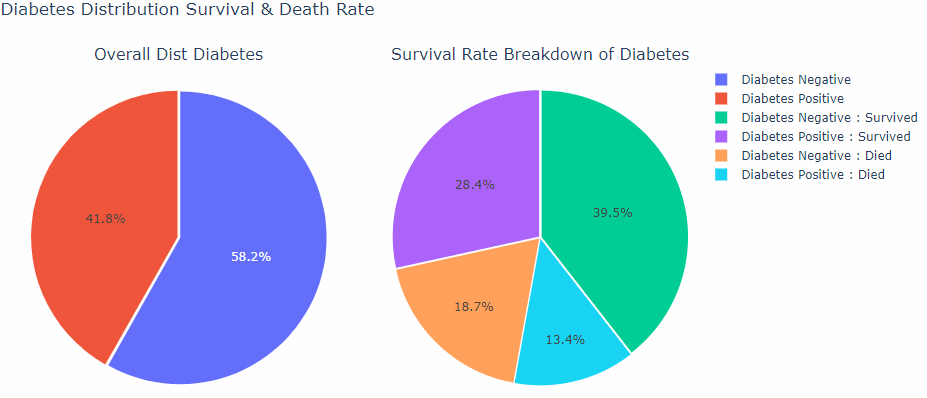
A details analysis and critical meaning regarding the population study that has a rate of Ejection Fraction measurements under 40 percent may be evidence of heart failure or cardiomyopathy. Populations with An Ejection Fraction measurements from 41 to 49 percent may be considered “borderline.” A normal Ejection Fraction measurement is between 55% and 70%. The Ejection Fraction measurements population study data shows that the fully 100% of people are positive for EF heart failure or cardiomyopathy. No one in the population that is measured for EF ranked negative. Visible in the "Survival Rate Breakdown of EF' shows that for the EF positive population (23.7%), do succumb to this heart failure or cardiomyopathy condition. Moreover, 76.3% of population who were EF positive did life and survive. That is a hopeful rating for EF positive people.

The remaining feature variables predominantly focus on people’s behavior practices for potential risk to heart failure. The predominate view for revealing these behavior practice risks was seen data visualization charts. This afforded easy comparison between male patients and female patients and the ranked tests of positive condition or negative condition. Positive is risk of heart failure. Negative is low to no risk.

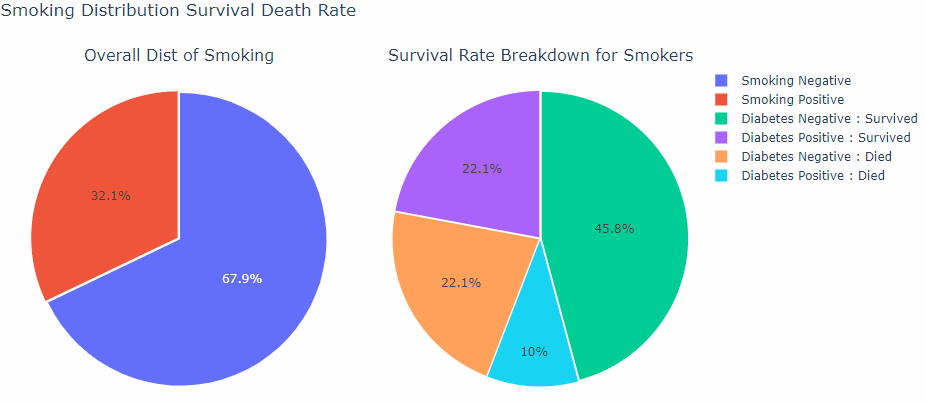
The Hypertension population study data shows that the approximately 35.1% of people are positive for hypertension and do suffer high blood pressure. While the population measure for people who are negative for hypertension is a 64.9% for this study. Within this population, as visible in the "Survival Rate Breakdown of Hypertension' shows that for the hypertension positive population (35.1%), the evidence in the breakdown chart points to a distribution that shows 22.1% people had survived successful from an event of a personal heart failure, while 19.1% of hypertension suffers succumbed to this high blood pressure condition. Furthermore, as for hypertension negative population (65%), the breakdown for hypertension negative who survived an event of a person heart failure did so live; while 13% of population who were hypertension positive did succumb to this high blood pressure condition.



The diabetes population study data shows that the approximately 41.8% of people are positive for diabetes and do suffer diabetes disease. While the population measure for people who are negative for diabetes is a 58.2% for this study. Within this population, as visible in the "Survival Rate Breakdown of Diabetes' shows that for the diabetes positive population (41.8%), the evidence in the breakdown chart points to a distribution that shows 28.4% people had survived having been rated negative for diabetes, while 18.7% of diabetes suffers succumbed to this diabetic condition. Furthermore, as visible for the diabetes negative population (58.2%), the breakdown for diabetes negative who survived diabetes was measured at 39.5%; while for the population who were diabetic positive who measured at 13.4%, did succumb to this diabetic condition.



The smoker’s population study data shows that the approximately 32.1% of people are positive for smokers and do suffer diabetes disease. While the population measure for people who are negative for diabetes is a 67.9% for this study. Within this population, as visible in the "Survival Rate Breakdown of Smokers' shows that for the smoker’s positive population (32.1%), the evidence in the breakdown chart points to a distribution that shows 22.1% people had survived having been rated negative for smokers, while 2.1% of smokers suffers succumbed from a lifetime (or lengthy time) smoking condition. Furthermore, as visible for the smoker’s negative population (67.9%), the breakdown for smoker’s negatives who survived smoking was measured at 45.8%; while for the population who were smokers and were positive measured at 10%, did succumb to this smoker’s lifetime (or lengthy time) condition.



1. **Conclusions**

The data pipeline approach utilized for this project made it easy for a designer developer to streamline the amount of code to a minimum to deliver as a usable yet basic ETL process and deploy data into a big data document store database structure for cardiovascular heart failure dataset data. One of the great successes in working with this pymongo / mongodb / python approach is that a data science engineering team can build a similar, yet simple, and effective ETL and big data solution with very little implementation code coverage for the essential functions. This greatly simplifies that testing time for quicker deployments. From the business perspective, this permitted a more concentrated focus on spend spent on the analysis of the data regarding health issues of heart failure and the attributes of high-risk behavior many patients succumbs to heart failures. Since the outcome of patients is the very survival of their lives, affording less time in data pipeline infrastructure build permits more time in analysis visualization and subsequent machine learning Random Forest Classifier model to focus on analyzing and forecasting predictive accuracy on the importance of this pending issue heart risk that impacts approximately 31% of deaths worldwide. Everyone needs to be aware of all the associated high-risk behavior that leads to unhealthy conditions. As humans, we are all in this together. This is an exciting field to engage in Big Data AI ML predictive & prescriptive analytics.

1. **References**

Chicco, D., Jurman, G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC Med Inform Decis Mak* **20,**16 (2020). <https://doi.org/10.1186/s12911-020-1023-5>.