COMP 542 Machine Learning

**Project Final Report**

**Title: Type2Heart (T2H)**

Team Member’s Names:

MOHAMMAD ZARAK SHAH JI

JAHANZAIB MALIK

1. **Introduction**

* 1. **Problems:**

Heart failure is a potential complication of type-2 diabetes, high blood pressure and age leading to disability or death. Recently, scientists have found that a new class of medications may be helpful for preventing heart failure in patients with diabetes but identifying those with the highest risk can be challenging.

Therefore, we tried to develop a risk score that would classify whether the patient has highest risk of heart failure or not. The risk score predictor is going to be named as WDC-MATH which stands for {Weight [BMI], Diabetes, Creatinine, myocardial infarction (MI), Age, high blood pressure, HDL-C (High-density lipoprotein (HDL) cholesterol), all these parameters being the major reason behind a heart failure.

**Reason for doing this project:**

The above-mentioned problem of heart attack due to type-2 diabetes and high blood pressure has been a personal problem in family as father of Mohammad Zarak has suffered from this problem recently and at that point of time if there would have been a model that could have predicted the possibility of a HF, could have certainly helped in preventing such situation.

Thus, the reason for doing this project is that any other person who can have same complications as Mohammad’s dad had could be prevented.

* 1. **Related (prior) works:**

Although several risk predictions models have become available in T2DM, none to our knowledge have been specific to heart failure HF risk. This machine learning–based approach has unique advantages over traditional risk prediction. Approaches adequately handling these issues may be especially important for complex phenotypes such as HF.

Many risk scores based on statistical knowledge have been developed for predicting individual’s risk of developing T2DM, such as risk evaluation formula, Archimedes trial-validated diabetes model, the diabetes risk score, genetic risk score, the New Chinese Diabetes Risk Score and the American Academy of Family Physicians risk model. These methods made the implicit assumption that each risk factor was linear to the outcome. The complex relationships between nonlinear interaction factors might be oversimplified, leading to the potential loss of related information. Moreover, when the number of variables increased, the hypothesis testing method became complicated. In contrast to traditional methods, machine learning can learn the nonlinear interactions iteratively from large amounts of data using computer algorithms, which have been applied in various fields, such as disease risk assessment and prediction.

* 1. **Brief approach and result:**

In this project, we analyzed a dataset of 299 patients with heart failure collected in 2015. We applied random survival forest (RSF) for the training purpose of our data and predicted a risk-score for survival of the patients comprising the dataset we have used, and then we ranked the features corresponding to the most important risk factors using Permutation-based Feature Importance.

The permutation-based selection uses the variable importance (VIMP) metric of the RSF. A high VIMP suggests that misspecification worsens the predictive accuracy in the forest, while a low VIMP suggests noise is more informative than the observed variable.

Briefly, an RSF is an ensemble classification method that determines a consensus prediction by averaging the results of many individual decision trees. Each individual tree is fitted using randomly selected data using a subset of the observations.

We make use of random survival forests, a random forests method for the analysis of right-censored survival data. Here survival splitting rules for growing survival trees are used, as is a new missing data algorithm for imputing missing data. A conservation-of-events principle for survival forests is implemented here and used to define ensemble mortality, a simple interpretable measure of mortality that can be used as a predicted outcome.

Our result tries to find out the factors within our dataset that mostly affect the risk of Heart Failure and we select the two of those features (which turn out to be serum creatinine and high blood pressure) in order to optimally predict the risk score pertaining to Heart Failure in the individuals. We also try to visualise the Survival Probability of those individuals.

This discovery has the potential to impact on clinical practice, becoming a new supporting tool for physicians when predicting if a heart failure patient will survive or not. Indeed, medical doctors aiming at understanding if a patient will survive after heart failure may focus mainly on serum creatinine and high blood pressure.

1. **Background**

Cardiovascular diseases kill approximately 17 million people globally every year, and they mainly exhibit as myocardial infarctions and heart failures. Heart failure (HF) occurs when the heart cannot pump enough blood to meet the needs of the body. Available electronic medical records of patients quantify symptoms, body features, and clinical laboratory test values, which can be used to perform biostatistics analysis aimed at highlighting patterns and correlations otherwise undetectable by medical doctors. Machine learning can predict patients’ survival from their data and can individuate the most important features among those included in their medical records.

Given the importance of a vital organ such as the heart, predicting heart failure has become a priority for medical doctors and physicians, but to date forecasting heart failure-related events in clinical practice usually has failed to reach high accuracy [1].

To understand the implementation of this project the reader must be familiar with some concepts that are associated with this project, which include the ones in the problem itself and the concepts given in the solution provided. So, the problem is simple (but the solution is not), we have some of the patients who are suffering from a disease called diabetes and are at a risk of having a heart attack which is one of the comorbidities arising due to presence of diabetes. So, our goal is to find out at how much risk an individual is based on the numerous factors or what we call features, that are part of our data. We want to find out based on time how much a person is at a risk of getting a Heart Failure. We also want to find out what are the factors in the dataset that most affect the risk of getting a Heart Failure.

In order to solve this kind of problem we have tried to use a machine learning model based on a data set which consisted of real-world data containing various features relating to various factors contributing towards Heart Functioning. Now to understand the implementation of Machine Learning the reader has to be familiar with Machine Learning in general especially the Random Forest classifiers which is the most feasible solution to this kind of problem. This problem has a twist that that we want to find “time to a certain event”, which in this case is the Heart Failure of individuals. Thus, using just the Random Forest classifier won't be enough as it is not any other data set, here we have a thing called censoring or we have censored data which is any data for which we do not know the exact event time. This type of problem is called a survival problem. Thus, we use a special Machine Learning model called Random Survival Forests which deals with the problems in which we have to find or predict a certain time to an event.

In this project, we analyze a dataset of 299 patients with heart failure collected in 2015. We apply machine learning models to both predict the patient's survival and rank the features corresponding to the most important risk factors. Feature ranking approach clearly identify serum creatinine and high blood pressure as the two most relevant features.

1. **Data and Model:**

* 1. **Dataset:**

We analyzed a dataset containing the medical records of 299 heart failure patients collected at the Faisalabad Institute of Cardiology and at the Allied Hospital in Faisalabad (Punjab, Pakistan), during April–December 2015. The patients consisted of 105 women and 194 men, and their ages range between 40 and 95 years old (Table shown below). All 299 patients had left ventricular systolic dysfunction and had previous heart failures that put them in classes III or IV of New York Heart Association (NYHA) classification of the stages of heart failure.

Table

Meanings, measurement units, and intervals of each feature of the dataset.

**Table: Meanings, measurement units, and intervals of each feature of the dataset:**

**mcg/L: micrograms per liter. mL: microliter. mEq/L: milliequivalents per litre**

The dataset contains 13 features, which report clinical, body, and lifestyle information (Table above), that we briefly describe here. Some features are binary: anaemia, high blood pressure, diabetes, sex, and smoking (Table above). The hospital physician considered a patient having anaemia if haematocrit levels (the ratio of the volume of red blood cells to the total volume of blood) were lower than 36%. Unfortunately, the original dataset manuscript provides no definition of high blood pressure.

Regarding the features, the creatinine phosphokinase (CPK) states the level of the CPK enzyme in blood. When a muscle tissue gets damaged, CPK flows into the blood. Therefore, high levels of CPK in the blood of a patient might indicate a heart failure or injury. The ejection fraction states the percentage of how much blood the left ventricle pumps out with each contraction. The serum creatinine is a waste product generated by creatine when a muscle breaks down. Especially, doctors focus on serum creatinine in blood to check kidney function. If a patient has high levels of serum creatinine, it may indicate renal dysfunction. Sodium is a mineral that serves for the correct functioning of muscles and nerves. The serum sodium test is a routine blood exam that indicates if a patient has normal levels of sodium in the blood. An abnormally low level of sodium in the blood might be caused by heart failure. The death event feature, that we use as the target in our binary classification study, states if the patient died or survived before the end of the follow-up period, that was 130 days on average. The original dataset unfortunately does not indicate if any patient had primary kidney disease and provides no additional information about what type of follow-up was carried out. Regarding the dataset imbalance, the survived patients (death event = 0) are 203, while the dead patients (death event = 1) are 96. In statistical terms, there are 32.11% positives and 67.89% negatives.

**Table: Statistical quantitative description of the category features:**

Table

Description automatically generated

**#: number of patients. %: percentage of patients. Full sample: 299 individuals. Dead patients: 96 individuals. Survived patients: 203 individuals.**

**Table: Statistical quantitative description of the numeric feature**

Table

Description automatically generated

**Full sample: 299 individuals. Dead patients: 96 individuals. Survived patients: 203 individuals. *σ*: standard deviation**

* 1. **Learning Models:**

Regarding learning model, we focused on Random Survival Forests, because as it turns out to be the top performing model for our problem to be addressed

The Random Survival Forests is based on Survival Analysis. The objective in survival analysis — also referred to as reliability analysis in engineering — is to establish a connection between covariates and the time of an event.  Survival analysis is a type of regression problem (one wants to predict a continuous value), but with a twist. It differs from traditional regression by the fact that parts of the training data can only be partially observed – they are *censored*. (**any data for which we do not know the exact event time)**

As an example, for explaining survival analysis, consider a clinical study, which investigates cardiovascular disease and has been carried out over a 1-year period as in the figure below.

Chart

Description automatically generated

Patient A was lost to follow-up after three months with no recorded cardiovascular event, patient B experienced an event four and a half months after enrollment, patient C withdrew from the study three and a half months after enrollment, and patient E did not experience any event before the study ended. Consequently, the exact time of a cardiovascular event could only be recorded for patients B and D; their records are *uncensored*. For the remaining patients it is unknown whether they did or did not experience an event after termination of the study. The only valid information that is available for patients A, C, and E is that they were event-free up to their last follow-up. Therefore, their records are *censored*.

Formally, each patient record consists of a set of covariates 𝑥 ∈ ℝ^𝑑, and the time t > 0 when an event occurred or the time c > 0 of censoring. Since censoring and experiencing and event are mutually exclusive, it is common to define an event indicator 𝛿∈ {0;1} δ∈ {0;1} and the observable survival time y > 0. The observable time 𝑦 of a right censored sample (right censoring occurs **when a subject leaves the study before an event occurs**) is defined as: Diagram

Description automatically generated with low confidenceConsequently, survival analysis demands for models that take this unique characteristic of such a dataset into account.

**Measuring the Performance of our Survival Model:**

Once we fitted our survival model, we wanted to assess how well our model can predict survival. Our test data was usually subject to censoring too, therefore metrics like root mean squared error or correlation were unsuitable. Instead, we used generalization of the area under the receiver operating characteristic (ROC) curve called Harrell’s concordance index or c-index. The area under the receiver operating characteristics curve (ROC curve) is a popular performance measure for binary classification task. In the medical domain, it is often used to determine how well estimated risk scores can separate diseased patients (cases) from healthy patients (controls). Given a predicted risk score f^, the ROC curve compares the false positive rate (1 - specificity) against the true positive rate (sensitivity) for each possible value of f^

The interpretation is identical to the traditional area under the ROC curve metric for binary classification: - a value of 0.5 denotes a random model, - a value of 1.0 denotes a perfect model, - a value of 0.0 denotes a perfectly wrong model. Our model’s c-index indicates that the model clearly performs better than random but is also far from perfect.

**Feature importance**:

One of the most basic questions we asked of a model is: What features had the biggest impact on predictions?

Therefore, after extensive researching we found out multiple ways to measure feature importance. Some approaches answer subtly different versions of the question above. Other approaches have documented shortcomings. We kept our focus on **permutation importance**. Compared to most other approaches, permutation importance is:

* fast to calculate,
* widely used and understood, and
* consistent with properties we would want a feature importance measure to have.

Permutation importance uses models differently than anything you've seen so far, and many people find it confusing at first. Permutation feature importance measures the increase in the prediction error of the model after we permuted the feature’s values, which breaks the relationship between the feature and the true outcome. The concept is straightforward: We measure the importance of a feature by calculating the increase in the model’s prediction error after permuting the feature. A feature is “important” if shuffling its values increases the model error, because in this case the model relied on the feature for the prediction. A feature is “unimportant” if shuffling its values leaves the model error unchanged, because in this case the model ignored the feature for the prediction. The permutation feature importance measurement was introduced by Breiman (2001) for random forests.

1. **Experiments and Analysis of Results:**

We implemented the idea using Random Survival Forests introduced in scikit-survival 0.11.

First, we needed to load the data and transform it into numeric values. Next, we acquired all the features for the model and generated a feature list and transformed feature variable. After that, we acquired and preprocessed the Event and 'time to event' variable for the model. To be fully compatible with scikit-learn, Status and Survival-in-days needed to be stored as a structured array with the first field indicating whether the actual survival time was observed or if was censored, and the second field denoting the observed survival time, which corresponds to the time of death (if Status == 'dead', 𝛿=1) or the last time that person was contacted (if Status == 'alive', 𝛿=0). where 𝛿= an event indicator 𝛿∈ {0;1}. Next, the data splitting was done so that we could determine how well our model generalizes. According to the features and outcome variable the split was decided to be 90% for training and 10% for testing so that the test data contains the patients equally comprising of having a right censored outcome and un-censored outcome.

Several split criteria have been proposed in the past, but the most widespread one is based on the log-rank test, which we probably know from comparing survival curves among two or more groups. Using the training data, then we fit a Random Survival Forest comprising 240 trees. Our test data was subject to censoring too, therefore metrics like root mean squared error or correlation were unsuitable. Instead, we used generalization of the area under the receiver operating characteristic (ROC) curve called Harrell’s concordance index or c-index.

The interpretation is identical to the traditional area under the ROC curve metric for binary classification: - a value of 0.5 denotes a random model, - a value of 1.0 denotes a perfect model, - a value of 0.0 denotes a perfectly wrong model. The concordance index of our model turned out to be 0.70, which indicates that the model clearly performs better than random and is a good a value as per the results reported in the Random Survival Forests paper**.**[12]

**Predicting:** For prediction, a sample was dropped down each tree in the forest until it reaches a terminal node. Data in each terminal was used to non-parametrically estimate the survival and cumulative hazard function using the Kaplan-Meier and Nelson-Aalen estimator, respectively. In addition, a risk score is being computed that represents the expected number of events for one particular terminal node. The ensemble prediction is simply the average across all trees in the forest. We selected a couple of patients from the test data according to the binary number of diabetes and high blood pressure. The predicted risk scores indicated that risk for a Heart failure (HF) in the patients with risk score greater than 140 is more as compared to the patients with score less than that if we are taking into consideration the dataset we have used. We then used predict survival function to have more detailed visual insight. It showed that the biggest difference occurs roughly within the first 125 days. The patient labelled 3 in the predicted set we took is being an anomaly because of the values of age, diabetes and high blood pressure and is likely on greater risk as compared to other patients in the dataset.

This can be seen in the figure below:

Chart, line chart

Description automatically generated

**Prediction values plotted using survival analysis function**

Alternatively, we also plotted the predicted cumulative hazard function to measure the total amount of risk that has been accumulated up to time t. This can be shown as below:

Chart, line chart

Description automatically generated

**Prediction values plotted using cumulative hazard function**

**Feature Importance:**

To check the biggest impact on predictions by the features we have used, we analyzed Permutation-based Feature Importance. The Feature permutation importance is a model-agnostic global explanation method that provides insights into a machine learning model’s behavior. It estimates and ranks feature importance based on the impact each feature has on the trained machine learning model’s predictions.

When interpreting Permutation importances the values towards the top are the most important features in the table generated by it, and those towards the bottom matter least.

The result showed us that the high blood pressure is by far the most important feature. If its relationship to survival time is removed (by random shuffling), the concordance index on the test data drops on average by 0.1193 points.

**Analyzing the results:**

The statistical analysis of our project identified age, EF, creatinine, sodium, anaemia and BP as the significant variables affecting the likelihood of mortality among heart failure patients. Most of the related studies [2-3] supported the male gender as predictor of mortality among heart failure patients. However, like Román et al. [4] in our study odd ratio of men/women is not significant. With respect to significance and importance of variables the findings of the present study are more in lines with Rahimi et al. [5]. The results are found to be similar with other international studies. [6-7].

The experimentation that we did for selecting different values for number of trees to be used showed us that having 1000 trees resulted in less concordance index as compared to selecting 240 trees. Additionally, when we considered the predict survival function to visualize the values, it showed that the biggest difference occurs roughly within the first 125 days.

The patient labelled 3 was being an exception because of the values of age, diabetes and high blood pressure and was likely on greater risk as compared to other patients in the dataset.

One more significant result that we analyzed was that all the patients which had high blood pressure resulted with a risk score of more than 140 and eventually where on a higher risk of mortality caused by a heart failure. With some deep analyzing of the results, we came to find out that the patient with highest risk score of 173.9 in the dataset didn’t have diabetes and the main reasons behind the high-risk score could have been his age of 80 and the presence of high blood pressure.

The findings that seem surprising are **non-significance of smoking and diabetes.** However, similar results concerning diabetes and smoking have been reported in other studies [8–9] as well. The reason behind may be smoking and diabetes are basically causes of heart problem at initial stages. We were only concerned with patients of New York Heart Association (NYHA) class III and IV which are advanced stages of heart failure. Up to these stages, these factors (diabetes and smoking) may probably be controlled by medications and hence these factors do not have significant effect on deaths due to heart failure in class III and IV.

1. **Conclusion:**

It can be concluded that growing age, renal dysfunction (having serum creatinine greater than its normal level 1.5), high BP, presence of anaemia and lower values of ejection fraction (EF) are the key factors contributing towards increased risk of mortality among heart failure patients. Increased level of serum sodium can reduce the odds of death. No significant differences were found due to smoking status, diabetes and gender of patients.

 Moreover, our approach showed that machine learning can be used effectively for binary classification of electronic health records of patients with cardiovascular health diseases.

As a limitation of our present study, we have to report the small size of the dataset (299 patients): a larger dataset would have permitted us to obtain more reliable results. Additional information about the physical features of the patients (height, weight, body mass index, etc.) and their occupational history would have been useful to detect additional risk factors for cardiovascular health diseases. Also, if an additional external dataset with the same features from a different geographical region had been available, we would have used it as a validation cohort to verify our findings.

Regarding future developments, we plan to apply machine learning to alternative datasets of cardiovascular heart diseases [10–11] and other illnesses.

Reference:

1. Buchan TA, Ross HJ, McDonald M, Billia F, Delgado D, Duero Posada JG, Luk A, Guyatt GH, Alba AC. Physician prediction versus model predicted prognosis in ambulatory patients with heart failure. J Heart Lung Transplant. 2019; 38(4):381. ([link](https://doi.org/10.1016%2Fj.healun.2019.01.971))

2. Barlera S, Tavazzi L, Franzosi MG, Marchioli R, Raimondi E, Masson S, et al. Predictors of Mortality in 6975 Patients with Chronic Heart Failure in the Gruppo Italiano per lo Studio della Streptochinasi nell’Infarto Miocardico-Heart Failure Trial Proposal for a Nomogram. Circ Hear Fail. 2013;6: 31–39. pmid:23152490. ([link](https://doi.org/10.1161/CIRCHEARTFAILURE.112.967828))

3. Tavazzi L, Senni M, Metra M, Gorini M, Cacciatore G, Chinaglia A, et al. Multicenter prospective observational study on acute and chronic heart failure one-year follow-up results of in-hf (italian network on heart failure) outcome registry. Circ Hear Fail. 2013;6: 473–481. pmid:23476054. ([link](https://doi.org/10.1161/CIRCHEARTFAILURE.112.000161))

4. Román AV, Shamagian LG, Diéguez MAB, Veloso PR, González-Juanatey JR. Influence of sex on mortality in hospitalized patients with congestive heart failure and preserved or depressed systolic function. Rev Española Cardiol. 2005;58: 1171–1180. ([link](https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0181001))

5. Rahimi K, Bennett D, Conrad N, Williams TM, Basu J, Dwight J, et al. Risk prediction in patients with heart failure: A systematic review and analysis. JACC Hear Fail. 2014;2: 440–446. pmid:25194291. ([link](https://doi.org/10.1016/j.jchf.2014.04.008))

6. Klein L, O’Connor CM, Leimberger JD, Gattis-Stough W, Piña IL, Felker GM, et al. Lower Serum Sodium Is Associated with Increased Short-Term Mortality in Hospitalized Patients with Worsening Heart Failure. Circulation. 2005;111: 2454–2460. pmid:15867182. ([link](https://doi.org/10.1161/01.CIR.0000165065.82609.3D))

7. Varela-Roman A, Shamagian LG, Caballero EB, Ramos PM, Veloso PR, Gonzalez-Juanatey JR. Influence of diabetes on the survival of patients hospitalized with heart failure: A 12-year study. Eur J Heart Fail. 2005;7: 859–864. pmid:15923138. ([link](https://doi.org/10.1016/j.ejheart.2005.01.017))

8. Otero-Raviña F, Grigorian-Shamagian L, Fransi-Galiana L, Názara-Otero C, Fernández-Villaverde JM, del Álamo-Alonso A, et al. Morbidity and mortality among heart failure patients in Galicia, N.W. Spain: the GALICAP study. Int J Cardiol. 2009;136: 56–63. pmid:18617282. ([link](https://doi.org/10.1016/j.ijcard.2008.04.025))

9. Cheraghi M, Sadeghi M, Sarrafzadegan N, Pourmoghadas A, Ramezani MA. Prognostic Factors for Survival at 6-Month Follow-up of Hospitalized Patients with Decompensated Congestive Heart Failure. ARYA Atheroscler 2010. 2010;6: 112–117. ([link](https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0181001))

10. Yunus I, Fasih A, Wang Y. The use of procalcitonin in the determination of severity of sepsis, patient outcomes and infection characteristics. PLoS ONE. 2018; 13(11):0206527. ([link](https://doi.org/10.1371%2Fjournal.pone.0206527))

11. Aushev A, Ripoll VR, Vellido A, Aletti F, Pinto BB, Herpain A, Post EH, Medina ER, Ferrer R, Baselli G. Feature selection for the accurate prediction of septic and cardiogenic shock ICU mortality in the acute phase. PLoS ONE. 2018; 13(11):0199089. ([link](https://doi.org/10.1371%2Fjournal.pone.0199089))

12. Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone, Michael S. Lauer, Random survival forests, Ann. Appl. Stat. 2(3): 841-860 (September 2008). DOI: 10.1214/08-AOAS169 ([link](https://projecteuclid.org/journals/annals-of-applied-statistics/volume-2/issue-3/Random-survival-forests/10.1214/08-AOAS169.full))