Data and Web Mining Project

*Abstract*—Cardiovascular diseases and problems are the leading cause of death globally with heart failure being a key factor in this. This research aims to develop a data mining solution that can be used to predict death by heart failure by using classification. The dataset used was taken from Kaggle.com and processed for use with the naïve models used and the K Nearest Neighbors algorithm. This research did not provide a good solution for predicting death by heart failure by using classification as the accuracy for each test was too low

Keywords—Heart failure, diabetes, high blood pressure, smoking, KNN(K Nearest Neighbors), Data mining, naïve models, data

# Introduction

In Ireland today, there are roughly 90 thousand people living with heart failure as reported by the HSE, you can develop it at any age but there is a higher risk of it developing as you become older. About 1% of under 65s are affected with 15% of over 85s being affected [1]. Cardiovascular diseases are the number one cause of death globally according to the World Health Organization accounting for 16% of total global deaths. Since 2000 the number of deaths related to this disease has risen from 2 million to 8.9 million in 2019, this is the biggest increase in deaths related to any disease [2]. The chosen dataset contains 12 features which we can use to predict the chance of death as a result of heart failure. The main objective of this project is to predict mortality as a result of heart failure. We will set out to achieve our objective by means of prediction models which are shown below. Cardiovascular diseases can be prevented by proper diet and regular activity, however, people with conditions such as diabetes and hypertension are at a higher risk of heart failure therefore early detection by means of a machine learning model would be a massive aid to these people [3] [4]. With heart failure being a common cause of death across the world we were interested to investigate any trends with this type of death when you look at factors such as diabetes, high blood pressure and other factors such as levels of serum creatinine and serum sodium in the blood.

This paper is structured as follows. A list of 10 related works is listed in the next section where they are individually discussed and explored. Section III is where the methodology is done with the implementation done in section IV. Our work is evaluated in section V, here we discuss how we used our methodology to explore our question regarding heart failure. We finished off our report with a conclusion where we summarised our findings and talked about any future work which can be done in this area.

# Related Work

## Evaluating risk prediction models for adults with heart failure: A systematic literature review

[5] This paper looked to identify and assess the quality of other published prediction models for heart failure. Some of the factors identified in relation to heart failure were diabetes, creatinine, age, sex, and blood pressure. These factors were also part of our chosen dataset which made this paper of interest to us. Diabetes was shown to be the most common comorbidity with Hypertension also being reported frequently, this led us to investigate these conditions for ourselves in our analysis. Overall, we felt this was a well put together paper, it showed us what these 40 papers did and did not do well and how they reported their data. It had provided us with a good base to start our project.

## Gender based survival prediction models for heart failure patients: A case study in Pakistan

[6] This next study looked at predicting heart failure based on risk factors which are gender specific. The Cox Model was used in this study with a likelihood ratio test being used to test goodness of fit of the chosen model. This study did a good job at showing how there are different factors which affect each gender, for example, males were more affected by smoking and anemia whereas females were more affected by factors such as sodium and platelets count.

## Systematic examination of a heart failure risk prediction tool: The pooled cohort equations to prevent heart failure

[7] This study focused on heart failure in black and white men and women. The study looks at a 10-year predicted heart failure using the Pooled Cohort Equation to do so. With both gender and race being used the results were an interesting read, for example, the study mentioned using a hypothetical 40-year-old. The probability of heart failure differed in each group with 0.1% to 9.7% chance for a White male, 0.5% to 12.3% chance for a Black male, <0.1% to 9.3% chance for a White female, and a 0.2% to 28.0% chance for a Black female. This study gave us an interesting insight into the heart failure problem as we are not using gender or race as variables in our study.

## Detecting Congestive Heart Failure by Extracting Multimodal Features and Employing Machine Learning Techniques

[8] This study looked at proposing an automated system to analyse heart rate variability (HRV) signals. This is slightly different to our report, but this study makes use of KNN which is one of the models we have chosen. This study gave us a good insight into how the KNN model can be adapted to different areas regarding Heart Failure.

## A prediction model for sudden cardiac death in patients with heart failure and preserved ejection fraction

[9] This study looks at a prediction model for sudden cardiac deaths for patients with heart failure and preserved ejection fraction. This study was of interest to us as ejection fraction is one of the factors in our dataset. A risk model for sudden cardiac death was done using cox regression. The best thing about this study is that the patients were followed up on 4 years later to see how and if some of them died. They ultimately concluded that their model would be of use for selecting patient for SCD prevention trials as there were not methods to identify patients at high risk of SCD at the time of this study.

## Comparative Study of Classifier for Chronic Kidney Disease prediction using Naive Bayes, KNN and Random Forest

[10] This study looks at predicting chronic kidney disease using Naive Bayes, K-Nearest Neighbours (KNN), and Random forests. This study appealed to us as it showed how KNN can be applied to different aspects of disease prevention. This study did a good job at representing the performance of the three models by use of graphs and tables. It also gave us a good idea as to how we approach our objective of predicting heart failure, starting with the sample data, and ending with evaluating the performance of the model.

## How visiting nurses detect symptoms of disease progression in patients with chronic heart failure

[11] The purpose of this study was to explore how nurses detect symptoms of the disease progression of heart failure patients. While the type of study and method used (KJ Method) may have been different to this report it gave us an interesting insight to how healthcare professionals are dealing with heart failure patients. The results mentioned how nurses find it difficult to detect disease progression due the patient comorbidities, these comorbidities are addressed in our data, for example diabetes. The results show the importance of a machine learning approach to disease prevention. The researcher did a good job at gathering the information as it was done firsthand rather through interviews rather than by using datasets.

## Practice patterns in the management of congestive heart failure and post discharge quality of life: A hospital based cross sectional study

[12] This paper aimed to identify the risks to heart failure patients after they have been discharged from hospital. The researcher conducted the study by getting in contact with the patients by telephone and asking a series of questions. The results showed that smoking was a major factor in patients’ health deteriorating further implying that there is an improvement needed in the management of this condition after patients are discharged. The researcher did a good job at conveying the results with the information clearly shown and easy to read graphs displayed. This paper was an interesting read as it showed what happens if patients are not fortunate enough to have had their disease detected before it is too late.

## Heart Failure with Preserved Ejection Fraction and 30-Day Readmission

[13] This study sets out to investigate if a particular type of heart failure contributes more to hospital readmission than others. The researcher used statistical tests rather than putting together prediction models the way this report does. The report ultimately concluded that Heart Failure with Preserved Ejection Fraction (HFpEF) was a major factor in patients being readmitted to hospital. This relates to our study as ejection fraction is one of the variables in our dataset. This type of study could be carried out after putting together prediction models to try and predict the death of a patient, anyone who survives could move onto this study to see if they are at risk of readmission to hospital.

## Predicting heart failure using deep neural network

[14] This study looks to achieve the same goal as we are, however, this study used Deep Neural Network (DNN). This gave us an interesting insight into how our project can be carried out in a different way but still strive to get the same result. The researcher achieved results of 88% accuracy again proving to us that the objective we are trying to carry out can be adopted many ways. After going through this study we were further motivated to carry out our own analysis as the variable types used in this study are similar to the ones we proposed using.

# Methodology

The methodology followed in this research was the Knowledge Discovery in Databases (KDD) methodology. KDD was chosen due to the emphasis of discovering insights in given data which corresponds to our objective of developing a data mining solution to discover any correlation between the medical conditions contained in the data and risk of heart failure.

## Data Selection and Description

The chosen dataset works well for the objectives of this research as it easy to apply data mining methods on due to all of the data either being numerical or a Boolean (in this case 1 or 0 for positive and negative). There were no missing values or invalid entries in the data so the data was ready to be used instantly. One of the main reasons this dataset was chosen was because of it’s relevance and functionality to a real world problem, while it is a relatively small dataset with only 300 entries it provides a good sample to use to create predictions and gain an insight into whether there is any correlation between heart failure and diabetes, high blood pressure, anaemia, etc.

This data was downloaded from Kaggle.com and contained 13 attributes with 6 being Boolean attributes and the remaining 7 being numerical attributes. In total there are 300 entries for

this dataset and the file size being relatively small at 13kb. Details on the data can be seen in figures 1 and 2.

## Data Processing and Transformation

With there being no missing or invalid entries in this data only a small amount of processing and transformation was carried out. The tidyverse package for R Studio was helpful for this stage as it is a larger collection of packages for R Studio that contains packages like readr, tidyr and dplyr as well as other packages like ggplot2 which will be used later. The attribute “DEATH\_EVENT” was converted from a Boolean attribute to a factor and assigned the labels “yes, no” as “1,0”.

## Data Mining

The goal of the research is to use the data provided to predict the likelihood of there being a link between the attributes within the data and heart failure by using classification. Classification is a supervised learning task that was chosen to be used on this data due to the nature of what we are trying to predict and the makeup of the provided data.

Naïve models were used to measure performance on simple models to get an understanding of how well the models initially work.

The KNN algorithm was also used to classify entries grouping them by similar attributes. KNN was chosen because of the ability to calculate Euclidian distance within the attributes thus making it easier to find patterns that occur within the data and is a good step up from the simpler naïve models.

# Implementation

## Prerequisites

After the data was chosen, processed and transformed we moved on to performing the main objective of our research which is applying our learning task and classifying the data. Before beginning all relevant packages were installed, including: tidyverse, caret, reshape, gmodels, dummies, class.

## Naïve Models

When performing classification the data is required to be split into training and testing data. When splitting the data we used a 34% sample for the test data with the other 66% going into the training data. the seed used was 1337 for the results to be replicable. The independent variables chosen from the data were diabetes, high blood pressure and smoking as these factors would naturally be believed to be linked to heart failure. Each independent variable was then put in an independent dataframe with the dependent variable so that a naïve model can be made for each variable. Our dependent variable was the DEATH\_EVENT attribute as we want to examine if a person has died from heart failure while having one of the outlined conditions above. The data was then split into training and test data once again with a 66-34 split for each model. After this we make and evaluate 1000 models based on 1000 samples to see how it performs. variables and visualise this by using a boxplot in R Studio which can be seen in figure 3. After this we can evaluate the performance of our classification by using a confusion matrix and examining the results.

## KNN

When implementing KNN we first summarise our numeric data and then normalise it so that all of the values are in the same interval and Euclidian distance can be measured accurately. After this we set our dependent variable and split that into training and test data, again using a 66-34 split. We then set our k values which were 14, 3 and 8. We chose 14 to get an idea of how KNN ran with a larger k value then 3 to see how it ran with a smaller value and then finally cose 8 to see how it performed in the middle. After this we load each of our KNN results into a confusion matrix to evaluate the results.

# Results/Evaluation

## Naïve Models

1. *Diabetes*

As we can see from the results from figure 4 we got an accuracy of 53.92% which is significantly lower than the no information rate of 71.57%, which means if we were to blindly predict the dominant class (in this case) in every case we would be 71.57% correct this explains our negative kappa value. Some other key values in our results are the sensitivity and specificity values. Sensitivity gives us the rate of true positives, in this case meaning the rate of death from heart failure when the person had diabetes, at 37.93% this is a low value meaning there may be a high number of false negatives and people going undetected who have died with diabetes. Our specificity – which is the rate of true negatives – comes in at 60.27% which is low in the context of this data. This may lead to false positives which may result in people being prescribed treatment which isn’t needed. Overall these are poor results from this model and are very inaccurate for this type of medical data.

1. *High Blood Pressure*

For the high blood pressure model results as seen in figure 5 we have an accuracy of 58.82% which is a slight improvement on the diabetes model but still too low to be considered accurate and is again lower than the no information rate of 71.57%. Another key value in the results is the confidence interval which in this case is 48.64-68.48 which means we can only be between 48.64% and 68.48% confident in these models which is far from a desired value for medical data of this nature. The sensitivity was again 37.93% which remains extremely low and unimpressive but the specificity for this model was 67.12% which is an improvement on the previous model but is still only slightly more accurate. Again these results are poor and inaccurate.

1. *Smoking*

For the smoking model results as seen in figure 6 we are given an accuracy of 59.8% which is again an improvement on the previous model but only very slight. The confidence interval has also slightly increased to 49.63% and 69.39% which is not a big jump but still a small improvement. Sensitivity for this model was 44.12% which is again an improvement on the previous model but closer to 50/50 which is not ideal for this data when a diagnosis could be a coin flip. Specificity came to 67.65% which is virtually identical to the previous model. Once again these results are an improvement on the previous results but not significant enough to make them good or accurate useful.

## KNN

When running KNN we got 3 sets of results based on the different assigned k values.

1. *K = 14*

From the results in figure 7 when k was assigned as 14 we got an accuracy of 64.36% which is not very high but could be expected with the low number of entries in the data. The confidence interval for when k equals 14 is between 54.21% and 73.64%, not a terrible result but not very good. Positively, the sensitivity is 94.2% which is a high result and indicates good performance. However the specificity came out at 0. This could be due to a number of factors that may become clear as we examine the other results.

1. *K =3*

For k equals 3 the results in figure 8 differed from the first set of results. The accuracy dropped to 53.47% which indicates the results may be worse as the k value drops as the confidence interval also dropped down to between 43.27% and 63.45% which almost equates to a drop of 10% across the accuracy and confidence intervals of the results. The sensitivity has dropped to 65.22% which is not a good sign and is a significant drop from the previous result and the specificity of 28.12% is obviously an increase on the previous results but is very poor. There is a clear difference between the 2 sets of results having a larger and smaller k value, next we will try something in between both.

1. *K = 8*

This time in our results as seen in figure 9 we get an accuracy of 58.42%, which is almost in the middle of our previous two results, again shown by the confidence interval between 48.18% and 68.14%. Sensitivity has increased to 76.81% and specificity has decreased to 18.75% which indicates a relationship whereas the sensitivity increases as a result of a larger k value, specificity then decreases.

# Conclusions and future work

##### From the naïve model results we can tell that our models were very inaccurate and would not be useful in predicting heart failure as the key values like sensitivity, specificity and accuracy are not nearly high enough for medical data like this.

For the KNN application on this data we also were not given very reliable results, while k was a higher value we were given better accuracy and sensitivity, the specificity was not returned and this makes the result useless.

##### While we did not achieve the level of accuracy we would have hoped for we are satisfied with the techniques which we employed, we believe that we did not return a high level of accuracy because of the small size of our dataset. We have discussed below how we would do things differently given more time.

##### If we had more time to complete this report, we would look to execute our models on a larger sample of data, the reason for this being a better improvement in the classification accuracy. If we were able to fully devote our time to this work and not have to worry about multiple other college deadlines, we would look to acquire larger datasets from the HSE, NHS or other countries medical agencies to see how results vary from country to country. This would provide us with in-depth knowledge about how people of different societal and living conditions are at risk of dying due to heart failure and other conditions that may be linked with heart failure. We would also look to implement different classification algorithms such as Naïve Bayes and logistic regression to compare to KNN and evaluate which algorithm performs best for this data.

##### References

1. "What is Heart Failure? - HSE.ie", HSE.ie. [Online]. Available: https://www.hse.ie/eng/health/hl/living/heartfailure/. [Accessed: 25- Apr- 2021].
2. "The top 10 causes of death", Who.int, 2020. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death. [Accessed: 20- Apr- 2021].*)*
3. M. C. Thomas, "Type 2 Diabetes and Heart Failure: Challenges and Solutions", Current Cardiology Reviews, vol. 12, no. 3, pp. 249-255, 2016. Available: 10.2174/1573403x12666160606120254.I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
4. [3]C. Krittanawong, "Machine learning prediction in cardiovascular diseases: a meta-analysis", Scientific Reports, vol. 10, no. 1, 2020. Available: 10.1038/s41598-020-72685-1.
5. G. Di Tanna, H. Wirtz, K. Burrows and G. Globe, "Evaluating risk prediction models for adults with heart failure: A systematic literature review", PLOS ONE, vol. 15, no. 1, p. e0224135, 2020. Available: 10.1371/journal.pone.0224135..
6. F. Zahid, "Gender based survival prediction models for heart failure patients: A case study in Pakistan", PLOS ONE, vol. 14, no. 2, p. e0210602, 2019. Available: 10.1371/journal.pone.0210602..
7. A. Bavishi, "Systematic examination of a heart failure risk prediction tool: The pooled cohort equations to prevent heart failure", PLOS ONE, vol. 15, no. 11, p. e0240567, 2020. Available: 10.1371/journal.pone.0240567.
8. L. Hussain, "Detecting Congestive Heart Failure by Extracting Multimodal Features and Employing Machine Learning Techniques", BioMed Research International, vol. 2020, pp. 1-19, 2020. Available: 10.1155/2020/4281243.
9. S. Adabag, "A prediction model for sudden cardiac death in patients with heart failure and preserved ejection fraction", European Journal of Heart Failure, vol. 16, no. 11, pp. 1175-1182, 2014. Available: 10.1002/ejhf.172.
10. R. Devika, S. V. Avilala and V. Subramaniyaswamy, "Comparative Study of Classifier for Chronic Kidney Disease prediction using Naive Bayes, KNN and Random Forest," 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), 2019, pp. 679-684, doi: 10.1109/ICCMC.2019.8819654
11. C. Taniguchi and A. Okada, "How visiting nurses detect symptoms of disease progression in patients with chronic heart failure", International Journal of Qualitative Studies on Health and Well-being, vol. 15, no. 1, p. 1735768, 2020. Available: 10.1080/17482631.2020.1735768.
12. Bansal A, Padappayil RP, Gopal S, Garg M, Joshi R. “Practice patterns in the management of congestive heart failure and post-discharge quality of life: A hospital-based cross-sectional study” J Family Med Prim Care. 2020 Nov 30;9(11):5592-5596. doi: 10.4103/jfmpc.jfmpc\_218\_20. PMID: 33532400; PMCID: PMC7842451.
13. Regmi MR, Bhattarai M, Parajuli P, Lara Garcia OE, Tandan N, Ferry N, Cheema A, Chami Y, Robinson R. Heart Failure with Preserved Ejection Fraction and 30-Day Readmission. Clin Med Res. 2020 Dec;18(4):126-132. doi: 10.3121/cmr.2020.1521. Epub 2020 Apr 27. PMID: 32340982; PMCID: PMC7735447.
14. M. T. Le, M. Thanh Vo, L. Mai and S. V. T. Dao, "Predicting heart failure using deep neural network," 2020 International Conference on Advanced Technologies for Communications (ATC), 2020, pp. 221-225, doi: 10.1109/ATC50776.2020.9255445.