CIND 820 – Literature Review

Sang Won Baek

Student Id: 501144120

Supervisor: Uzair Ahmad, Ph.D

Date: June 6, 2022

Table of Contents

[Revised Abstract - Introduction 3](#_Toc452457337)

[**Literature Review 3**](#_Toc452457338)

[**Brief Descriptive Statistics of Selected Dataset 6**](#_Toc452457339)

[**Tentative Methodology**](#_Toc452457339) **7**

[**GitHub Repository Link**](#_Toc452457339) **7**

[**References**](#_Toc452457339) **8**

# Revised Abstract – Introduction

Cardiovascular diseases are the leading cause of death globally causing an estimated 17.8 million deaths per year (Kaptoge et al., 2019). One of the United Nations sustainable development goals is to drastically reduce incidence of premature mortality from non-communicable diseases by a third by the year 2030 (Kaptoge et al., 2019). Therefore, to achieve the goal set out by the United Nations, it is imperative to reduce the incidence of CVD (cardiovascular diseases) worldwide. Towards this end, it is imperative to develop technologies and techniques for early detection of individuals at risk of developing cardiovascular disease (Dahlöf, 2010). Unfortunately, the detection of at-risk patients is a difficult task as most adverse cardiovascular events occur in individuals with modest and (often) unnoticed elevations in numerous risk factors, rather than a single factor being an outlier (Dahlöf, 2010). Thus, the advancement of machine learning techniques can greatly aid in the effort to identify individuals at risk for cardiovascular disease, leading to a reduction of cardiovascular disease mortality (Al’Aref et al., 2019).

The focus of my capstone project will be on utilizing classification (supervised learning) to develop a prediction algorithm. The goal will be to develop both a decision tree-based and a logistic regression model to predict heart disease events. The efficacy of both models will be compared to determine which model is more effective in predicting heart disease events. The project will also include data exploration and visualization. The “Heart Failure Prediction Dataset” found publicly on Kaggle.com will be used to train and test the models (fedesoriano, 2021). Lastly, the project will be programmed in python using Jupyter notebook and various libraries such as numpy, pandas, matplotlib, seaborn, and sklearn.

# Literature Review

As stated before, the main focus of my capstone project is to answer the question of whether supervised learning techniques could be utilized to aid in the diagnoses of adverse heart disease events. Towards answering this question, my project will utilize both a decision tree-based and a logistic regression model to predict heart disease events. Since the 1960s, there have been an exploration of applying computers, algorithms and artificial intelligence information systems in the field of medicine (Mathur, Srivastava, Xu, & Mehta, 2020). The recent advancements in machine learning technologies have spurred a large number of research literature exploring the usability of machine learning techniques in the diagnosis of cardiovascular diseases (Mathur et al., 2020). Most of the research literature have implemented several machine learning concepts composed of both supervised, unsupervised, and deep learning methods for the diagnosis of cardiovascular disease (Mathur et al., 2020). As there have been numerous research papers published, the following paragraphs will review those papers in relation to my proposed capstone project.

First, a research paper published in the Institute of Electrical and Electronic Engineers outlined a prediction model to predict whether individuals have a cardiovascular disease (K. G. Dinesh, K. Arumugaraj, K. D. Santhosh, & V. Mareeswari, 2018). The paper used a dataset obtained from the Cleveland, Hungarian, Switzerland, Long Beach VA heart disease database from the UCI machine learning repository to train their model (K. G. Dinesh et al., 2018). The dataset originally consisted of 920 records with 76 attributes related to the patient’s medical records, after the attribute selection process the research team ended up using 14 attributes to train their models (K. G. Dinesh et al., 2018). The language R was used as the basis of their operating environment to perform statistical computation and graphical representation in relation to data analysis (K. G. Dinesh et al., 2018). After preprocessing the data, the research team designed five different classification algorithms (K. G. Dinesh et al., 2018). The various machine learning algorithms used by the research team consisted of logistic regression, naïve bayes, random forest, support vector machine, and gradient boosting (K. G. Dinesh et al., 2018). The research team then used an accuracy module to calculate the maximum accuracy for each model (K. G. Dinesh et al., 2018). The results of the paper displayed that the logistic regression model resulted in the highest overall accuracy of 0.865, while the support vector machine model resulted in the lowest overall accuracy of 0.798. In relation to my capstone project, this research paper highlights the potential uses of machine learning in diagnosing cardiovascular diseases. The methodology of the research paper is similar to my proposed capstone project as they used both logistic regression and decision tree (random forest) models to predict cardiovascular disease. The main difference between their research paper and my capstone project would be the operating environment as I will be using Python instead of R. The difference in operating environment will help test if there are any benefits or restrictions of using python over R when designing classification algorithms.

Second, research conducted in Nirma University, India used the WEKA software tool to design a heart disease prediction system via decision tree classifiers with 10-fold cross validation (Patel, TejalUpadhyay, & Patel, 2015). The research team used the Cleveland dataset consisting of 76 attributes (of which 13 were used) and 303 records from the UCI repository (Patel et al., 2015). Four different types of decision tree classifiers were used throughout the paper, the J48 algorithm (an open-source Java implementation of the C4.5 algorithm contained in WEKA), J48 with reduced error pruning, logistic model tree algorithm, and random forest algorithm (Patel et al., 2015). The research discovered that the J48 with reduced error pruning model generated the best performance as it showed a test error of 0.167 compared to 0.238 (logistic model tree) and 0.2 (random forest algorithm) (Patel et al., 2015). However, the research team concluded that their model only achieved marginal success and that there is a need for further research in combinational and more complex models to increase the accuracy of predicting early onset heart disease (Patel et al., 2015). In contrast to my capstone project, the research team focused their efforts on decision tree classifiers while I will utilize both decision tree and logistic regression models. As the researchers have pointed out the importance of continual exploration of machine learning techniques in predicting heart disease, I hope to add on to the conversation through my project.

Third, research conducted in 2019, explored data driven approaches through the utilization of supervised machine learning models to detect patients at risk for diabetes and cardiovascular disease (Dinh, Miertschin, Young, & Mohanty, 2019). The researchers used the National Health and Nutrition Examination Survey (NHANES) dataset to train multiple machine learning models (Dinh et al., 2019). Logistic regression, random forest, support vector machines, and gradient boosting models were evaluated on their classification performance (Dinh et al., 2019). The researchers ultimately combined the models to develop a weighted ensemble model to leverage the performance of the separate models to increase detection accuracy (Dinh et al., 2019). The results of the ensemble model for cardiovascular disease achieved an Area Under-Receiver Operating Characteristics (AU-ROC) score of 83.1 percent (using no laboratory results) and 83.9 percent accuracy (with laboratory results) (Dinh et al., 2019). This research paper conveys potential next steps for my capstone project. Larger scale implementation of my capstone project could possibly include the development of more complex machine learning models through combining multiple models (logistic regression, decision tree models). In contrast to my project, the research paper focused on a wider scope through including both cardiovascular disease and diabetes.

Fourth, another research article published in 2019 found that biomedical researchers were able to quickly build competent machine learning classifiers to predict cardiovascular disease risk, without having in depth knowledge on the underlying algorithms through the use of auto machine learning (AutoML) libraries (Padmanabhan, Yuan, Chada, & Nguyen, 2019). The researchers compared manually created machine learning classifiers (designed by a graduate student using scikit-learn) to classifiers built by an auto machine learning library (auto-sklearn) (Padmanabhan et al., 2019). The results indicated that the automatic machine learning took one hour to build classifiers that perform better than the ones built by the graduate student that took one month (Padmanabhan et al., 2019). The experiment was completed using python and the Heart UCI dataset which contain 76 attributes (only 12 were used) and 303 records (Padmanabhan et al., 2019). Overall, the paper aimed to propose the adoption of AutoML in the healthcare domain by breaking the stereotype that machine learning is only accessible to trained experts (Padmanabhan et al., 2019). The sentiment of this research paper mirrors the same sentiment I hold with my capstone project. The integration of machine learning in the clinical setting is crucial towards the advancement of patient care and diagnosis.

Fifth, research conducted in McMaster University designed a system for monitoring cardiovascular disease using mobile devices and machine learning techniques (Boursalie, Samavi, & Doyle, 2015). The system used wearable sensors that collected observable vital signs from patients, the data would be transferred to a support vector machine (SVM) instead of a health care professional to be analyzed (Boursalie et al., 2015). The SVM would analyze features extracted from clinical databases and the data collected from the wearable sensors (the patient) to classify a patient as “continued risk” or “no longer at risk” for cardiovascular disease (Boursalie et al., 2015). Overall, using a synthetic clinical database of 200 patients, the experiment was able to classify a patient’s cardiovascular disease risk with a 90.5 percent accuracy (Boursalie et al., 2015). This research clearly displays a proof of concept of how machine learning can be implemented in real life clinical settings in the present time. Using a combination of monitoring devices for cardiovascular disease patients that not only delivers data to health care providers but can also process raw data into meaningful information can greatly improve patient care.

Lastly, research published in 2017 took clinical data from 378,256 patients from UK family doctors and compared four machine learning algorithms against the established algorithm (American College of Cardiology guidelines) to predict first cardiovascular event over 10 years (Weng, Reps, Kai, Garibaldi, & Qureshi, 2017). The machine learning algorithms used in the research were logistic regression, random forest, gradient boosting machines, and neural networks (Weng et al., 2017). Compared to the established risk prediction algorithm, all the machine learning algorithms improved prediction accuracy (Weng et al., 2017). Logistic regression improved by 3.2 percent, random forest by 1.7 percent, gradient boosting by 3.3 percent, and neural networks by 3.6 percent (Weng et al., 2017). This research has shown that machine learning improves the accuracy of cardiovascular risk prediction, increasing the number of patients identified as high risk who could benefit from early preventive treatment. Unlike the other research papers outlined above and my capstone project, this paper included the use of neural networks into their experimental design. While the utilization of neural networks are outside the scope of my capstone project, it may be an interesting next step to take if I were to increase the scope of my project, especially as the research conducted by Weng et al. have shown the benefits of its uses.

# Brief Descriptive Statistics of Selected Dataset

Graphical user interface, text, application, table

Description automatically generatedImage taken from my google collab notebook.

# Tentative Methodology

|  |  |
| --- | --- |
| Date: | Goal: |
| June 6 to June 12 | Exploratory data analysis – summarize and visualize the data.  Data preparation – clean, select, preprocess and transform the data. |
| June 13 to June 19 | Continue data preparation (if needed).  Design and develop logical regression and random forest (decision tree) models.  Test and document results. |
| June 20 to June 26 | Continue the design and development of the logical regression and random forest (decision tree) models.  Test and document results.  Upload codes to GitHub repository. |
| June 27 | Submit initial results and code. |
| June 28 to July 3 | Finalize/improve machine learning models (if needed).  Record/analyze the effectiveness, efficiency, and stability of the models.  Begin writing final project report. |
| July 4 to July 10 | Write final project report (outlined in course module). |
| July 11 to July 17 | Write final project report (outlined in course module). |
| July 18 to July 24 | Complete final project report (outlined in course module).  Create project demo PowerPoint presentation. |
| July 25 | Submit final results and project report. |
| July 26 to July 28 | Complete project demo PowerPoint presentation. |
| July 29 | Submit project presentation. |

# GitHub Repository Link

<https://github.com/swb1113/CIND820>

References

Al’Aref, S. J., Anchouche, K., Singh, G., Slomka, P. J., Kolli, K. K., Kumar, A., ... & Min, J. K. (2019). Clinical applications of machine learning in cardiovascular disease and its relevance to cardiac imaging. *European heart journal*, *40*(24), 1975-1986.

Boursalie, O., Samavi, R., & Doyle, T. E. (2015). M4CVD: Mobile machine learning model for monitoring cardiovascular disease. *Procedia Computer Science, 63*, 384-391.

Dahlöf, B. (2010). Cardiovascular disease risk factors: Epidemiology and risk assessment.*The American Journal of Cardiology, 105*(1, Supplement), 3A-9A. doi:<https://doi.org/10.1016/j.amjcard.2009.10.007>

Dinh, A., Miertschin, S., Young, A., & Mohanty, S. D. (2019). A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. *BMC Medical Informatics and Decision Making, 19*(1), 1-15.

fedesoriano. (September 2021). Heart Failure Prediction Dataset. Retrieved [May 15, 2022] from <https://www.kaggle.com/fedesoriano/heart-failure-prediction>.

Kaptoge, S., Pennells, L., De Bacquer, D., Cooney, M. T., Kavousi, M., Stevens, G., ... & Di Angelantonio, E. (2019). World Health Organization cardiovascular disease risk charts: revised models to estimate risk in 21 global regions. *The Lancet Global Health*, *7*(10), e1332-e1345.

K. G. Dinesh, K. Arumugaraj, K. D. Santhosh, & V. Mareeswari. (2018). Prediction of cardiovascular disease using machine learning algorithms. Paper presented at the *- 2018 International Conference on Current Trends Towards Converging Technologies (ICCTCT),* 1-7. doi:10.1109/ICCTCT.2018.8550857

Mathur, P., Srivastava, S., Xu, X., & Mehta, J. L. (2020). Artificial intelligence, machine learning, and cardiovascular disease. *Clinical Medicine Insights: Cardiology, 14*, 1179546820927404.

Padmanabhan, M., Yuan, P., Chada, G., & Nguyen, H. V. (2019). Physician-friendly machine learning: A case study with cardiovascular disease risk prediction. *Journal of Clinical Medicine, 8*(7), 1050.

Patel, J., TejalUpadhyay, D., & Patel, S. (2015). Heart disease prediction using machine learning and data mining technique. *Heart Disease, 7*(1), 129-137.

Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M., & Qureshi, N. (2017). Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PloS One, 12*(4), e0174944.