Assessment 1: Research Proposal

Measuring the Limits of Machine Learning Algorithms in Predicting Patient Prognosis

|  |  |
| --- | --- |
| Name: | Supervisor: |
| University ID: |  |
| Word Count: 3242 (Without References) |  |
| Date: |  |

Table of Contents

[Table of Figures 2](#_Toc70167797)

[Abstract 3](#_Toc70167798)

[Introduction 3](#_Toc70167799)

[Background 3](#_Toc70167800)

[Aim/Objective 4](#_Toc70167801)

[Research Questions 4](#_Toc70167802)

[Ethical Considerations 4](#_Toc70167803)

[Literature Review 4](#_Toc70167804)

[Project Timeline 11](#_Toc70167805)

[Methodology 11](#_Toc70167806)

[Project Evaluation 12](#_Toc70167807)

[Conclusion 12](#_Toc70167808)

[References 12](#_Toc70167809)

# Table of Figures

[Figure 1 – Results (Bengtsson and Lindblad, 2021) 5](#_Toc70167810)

[Figure 2 - Machine Learning Taxonomy 6](#_Toc70167811)

[Figure 3 – Results (Venkatesh, Balasubramanian, and Kaliappan, 2019) 7](#_Toc70167812)

[Figure 4 - Data Statistics (Park, Cho, Kim, Wall, Stern, Lim, Yoo, Kim and Cha, 2020) 8](#_Toc70167813)

[Figure 5 - Performances for Models Trained (Park, Cho, Kim, Wall, Stern, Lim, Yoo, Kim and Cha, 2020) 9](#_Toc70167814)

[Figure 6 - Top Performing Features (Park, Cho, Kim, Wall, Stern, Lim, Yoo, Kim and Cha, 2020) 9](#_Toc70167815)

[Figure 7 - Figure 6 - Performance of AD predictive models trained on NHIS-NSC by using balanced samples (Park, Cho, Kim, Wall, Stern, Lim, Yoo, Kim and Cha, 2020). 10](#_Toc70167816)

[Figure 8 - Performance of AD predictive models trained on NHIS-NSC by using balanced samples (Park, Cho, Kim, Wall, Stern, Lim, Yoo, Kim and Cha, 2020). 10](#_Toc70167817)

[Figure 9 - Project Timeline 11](#_Toc70167818)

[Figure 10 - Feature Trend for the Dataset. 12](#_Toc70167819)

# Abstract

Today, wellbeing expectation in current life turns out to be a lot of fundamental. Enormous information examination assumes an urgent part to foresee future status of wellbeing and offers superior well-being result to individuals. To mechanize the cycle and foresee infections all the more precisely AI strategies are acquiring notoriety in research local area.

The advent of technology has made it easier for humans to automate the process of patient prognosis and deduction of disease/issue from symptoms and report analysis. However, Machine Learning is only as capable as the data and the resources it has. It cannot expand and learn at will. But the capacity of digesting huge amounts of data and statistically analysing to provide accurate predictions in the blink of an eye is what draws the researchers to using ML for this purpose.

This research aims to render out the limits of Machine Learning in achieving prognosis for patient health using Machine Learning. There are many possible techniques and paradigms for use in this research. Classification algorithms like Multi-layer Perceptron, Decision trees, SVM or Regression algorithms like Logistic Regression. The output of this research will be the experimental proof and analysis of how much a Machine Learning algorithm can perform given a medical dataset to predict patient health.

**Keywords:** Prognosis, Classification, Regression, Multilayer Perceptron, Decision trees, SVM, Logistic Regression

# Introduction

Since Centuries humanity has found different demonstrated medical services frameworks. To robotize the interaction and anticipate sicknesses all the more precisely AI techniques are acquiring fame in research local area. AI strategies work with improvement in the knowledge to a machine for more efficient performance later on utilizing insight from learning (Santosh A. Shinde and P. Raja Rajeswari, 2018).

Nowadays, wellbeing expectation in current life turns out to be a lot of fundamental. Huge information examination assumes a vital part to anticipate future status of well-being and offers superior well-being result to individuals. A great deal of examination is going on prescient investigation utilizing AI methods to uncover better dynamic. Large information investigation cultivates extraordinary freedoms to anticipate future well-being status from wellbeing boundaries and give the best results (Venkatesh, Balasubramanian and Kaliappan, 2019).

This research project aims to utilize Machine Learning for Patient Prognosis and Diagnosis using Data Analysis. The main aim behind this is to measure the limits i.e. how much Machine Learning can achieve in terms of Patient Prognosis when provided with the data resources.

## Background

Technology development has taken healthcare to zeniths never seen before. The stage now is of automation. The necessity comes with the involvement of Machine Learning in the medical domain. The huge amount of data required to process for predicting patterns and identifying disease symptoms can only be targeted using Machine Learning. This data is normally used for patient prognosis.

Therefore, researchers around the globe are working on the automation of patient prognosis. The analysis of patient data and reports to predict whether the patient has a disease or not and provide a report on it is what has got the attention of the researchers in the medical domain. This research is focused on testing the mettle of Machine Learning algorithms to the limit given a dataset for prognosis.

## Aim/Objective

This research project aims to see how much Machine Learning can achieve in terms of predicting patient prognosis. To achieve this, the following steps are traced:

* Literature Review of papers already written in the same domain
* Dataset Finalization, Comprehension and Analysis
* Data Pre-processing and Machine Learning Algorithm Design
* Performance Metrics for the Machine Learning Ensemble

## Research Questions

The question that provides the drive to this research is as follows:

1. What are the Limitations of using Machine Learning for the automation of patient prognosis?

## Ethical Considerations

Ethics is a complicated subject that has only become more prominent during the advent of Big Data. The UK Data Service department also provides guidelines for ethical research with specific relation to Big Data. These guidelines will form the basis for this reports ethical approach. Some of the concerns that will be addressed are:

* Maintaining confidentiality in line with Birmingham City University (BCU) and DC guidelines,
* Anonymising information that violates group privacy,
* Ensuring transparency in reasons for data collection,
* Ensuring data is only used for the direct purpose it has been requested,
* Referencing sources for all information used within the research project,
* Ensuring all data is stored in the correct location. DC information must remain on DC servers.

## Literature Review

The process of reviewing, studying and understanding research done by peer researchers in the same domain through their research papers is known as Literature Review (LR). A strong LR provides validity to the integrity of the research using proven, published facts. The papers selected for LR in this research are summarized as follows.

Missing information is the most outstanding issue experienced by AI topic specialists while investigating certified information. In different applications going from quality verbalization in computational science to inspect reactions in humanistic frameworks, missing information is open to different degrees (Bertsimas, Pawlowski and Daisy Zhuo, 2018). As different quantifiable models and AI assessments depend upon complete illuminating records, it is significant to dealing with the missing information suitably. To a great extent, major methods of reasoning may take care of business to oversee missing information. For instance, an absolute case appraisal utilizes essentially the information that is known and obstructs all experiences with missing qualities to incite irrefutable evaluation (Bertsimas, Pawlowski and Daisy Zhuo, 2018).

These limits excellently a few observations contain missing qualities, and when the information is missing totally at irregular, complete case appraisal doesn't quick lopsided results (Bertsimas, Pawlowski and Daisy Zhuo, 2018). Then, several AI assessments normally address missing information, and there is no need for pre-dealing with it. For example, CART and K-surmises have been adjusted to issues with missing information. In different conditions, missing attributes should be credited to going before running quantifiable appraisals on irrefutably the instructive document (Bertsimas, Pawlowski and Daisy Zhuo, 2018).

Examiners have been working on this issue for quite a while previously and have composed innovative, new and novel approaches to manage to deal with Data Imputation. The Literature Review for this investigation fixates on approaches to manage Data Imputation after 2014. The philosophies for Data Imputation are generally quantifiable, yet more imaginative and novel techniques are appearing in the Literature i.e., significant learning, generative opposing learning, fuzzy logic, etc.

Missing information is universal in massive information clinical starter. Through different evaluations don't unequivocally report how they handle missing information, some obvious techniques are utilized in quantifiable programming. As such, various packs may shockingly oversee missing information (or the default frameworks are phenomenal) and results may not be copied definitively by utilizing arranged quantifiable programming gatherings (Zhang, 2016). Now and then, this may not incite on an extremely fundamental level various outcomes, yet the steady adequacy of the assessment is attacked. The best practice is to unequivocally state how missing attributes are managed. For ease, different specialists kill inadequate case (Listwise prohibition), which is comparatively the default strategy in different lose the faith packs (Zhang, 2016).

This technique gets reliable outcomes precisely when the measure of missing qualities isn't enormous and the missing model is missing eccentrically (MCAR) or missing MAR. Another affront to finish case appraisal is data affliction. This can be a critical issue when there is an enormous number of (segments) (Zhang, 2016). A liberal number of cases can be erased because annulment depends on missing qualities on at any rate one components. Also, an absolute case assessment can prompt erratic propensity (Zhang, 2016). The reaction to this issue is attribution. Missing qualities are uprooted by ascribed values. Since credit is a space of dynamic evaluation, there are various methodologies and gatherings made for attribution (Zhang, 2016).

The missing characteristics are by and large evaluated using central tendency assessments like mean, centre and mode in various kinds of investigation (Zhang, 2016). The mean and standard deviation are lopsided. Attributions with mode and focus work thusly and they are left to customers for planning (Zhang, 2016). However, unforgiving attribution gives fast and central systems to missing qualities, it puts down change, manages the relationship among factors, and propensities synopsis encounters. In this manner barbarous attributions ought to be utilized when a subtle bunch of qualities are missing, they are not for general use (Zhang, 2016).

A couple of experts use Listwise Deletion, Predictive Mean Matching and Poisson Imputation for dealing with the data attribution issue (Bengtsson and Lindblad, 2021). The results for these systems from the paper by (Bengtsson and Lindblad, 2021) are according to the following:

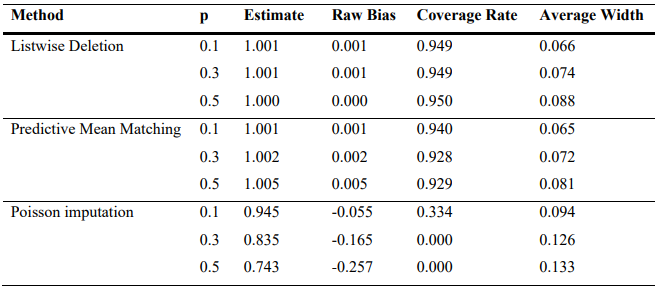


Figure 1 – Results (Bengtsson and Lindblad, 2021)

Most scientists produce papers that make the following contributions to the domain:

* Orders classifications of AI techniques.
* Orders best in class research in wellbeing hazard forecast
* Frameworks or infection forecast frameworks.
* Gives future exploration headings.

Figure 1 shows the most popular understanding of Machine Learning paradigms from research papers (Shinde and Rajeswari, 2018).

**Supervised Learning:** Supervised learning techniques work on known assumptions on the marked info dataset.

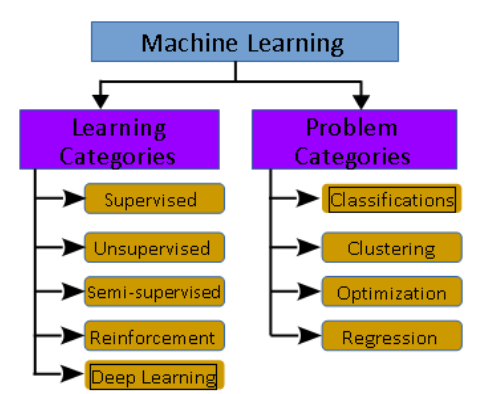


Figure 2 - Machine Learning Taxonomy

**Unsupervised Learning:** These techniques focus to break down the design of information in the provided input unlabelled dataset and construct planning between the info and yield credits, while yield credits are obscure before the examination.

**Semi-supervised Learning:** These strategies utilized both named and unlabelled datasets to create models for insight surmising.

**Reinforcement Learning:** This techniques objective is to maximize the awards from the outcome. That is support learning technique creates a grouping of choices that help to secure the most noteworthy prizes.

**Deep Learning:** These techniques centre on bringing together fake knowledge with AI. It chips away at regular information to give significant bits of knowledge. It deals with an input dataset that has less marked information and addresses issues grouped under semi-administered figuring out how to fabricate complex neural organization models.

AI calculations or methods are additionally characterized utilizing learning issues as Classification, Clustering, Optimization, and Regression (Shinde and Rajeswari, 2018).

1. **Classification:** It is a gathering method that relies upon the given estimation of target and dataset. As per the provided target esteem, it measures and orders the dataset.
2. **Clustering:** It is a method that takes just the dataset as an input and recognizes fascinating examples to infer knowledge. When contrasted with the order, in the grouping objective worth isn't given as a piece of information or it is an obscure boundary.
3. **Regression:** In this strategy knowledge or data is gotten from the past learning experience. A condition is inferred that matches with the vast majority of the information focus and the situations where information doesn't fit with the bend, those information focuses we dispose of. This strategy is known as relapse.
4. **Optimization:** It is a technique to improve the exhibition of the framework as far as different ascribes.

The Naïve Bayes approach is executed and the result plainly shows the high precision in many research papers (Venkatesh, Balasubramanian, and Kaliappan, 2019). The Machine learning approach (Naïve Bayes) is incorporated with sparkle climate and demonstrated to be a suitable foreseeing arrangement when managing wellbeing boundaries. The proposed BPA-NB plot involved two phases including the bunching and forecast stage. The bunching stage includes cleaning the acquired information and gathering the information dependent on the illnesses (Venkatesh, Balasubramanian, and Kaliappan, 2019). The prediction stage finds the class marks incorporate Yes\_p and No\_p utilizing the Naïve Bayes method.

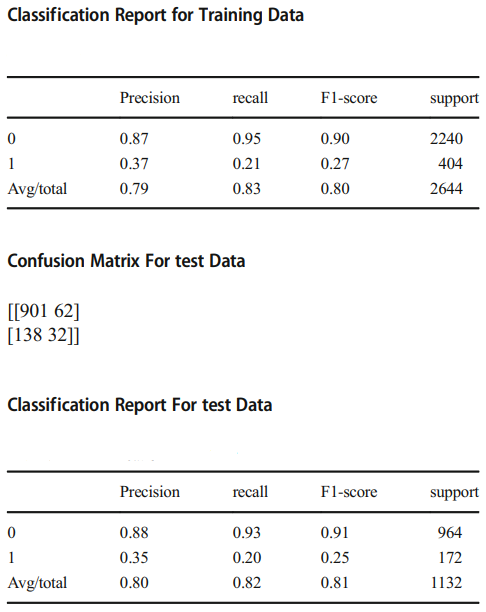


Figure 3 – Results (Venkatesh, Balasubramanian, and Kaliappan, 2019)

The authors in this domain have tried the chance of AI models to anticipate future frequency of Alzheimer's sickness (AD) utilizing enormous scope authoritative wellbeing information (Park, Cho, Kim, Wall, Stern, Lim, Yoo, Kim and Cha, 2020). From the Korean Public Health Insurance Service data set somewhere in the range of 2002 and 2010, the authors got de-recognized wellbeing information in seniors over 65 years (N = 40,736) containing 4,894 interesting clinical highlights including ICD-10 codes, drug codes, research centre qualities, history of individual and family disease and socio-socioeconomics. To characterize episode AD, the authors thought about two operational definitions: "unmistakable AD" with demonstrative codes and dementia prescription (n = 614) and "plausible AD" with just finding (n = 2026).

The authors prepared and approved irregular woodland, support vector machine and strategic relapse to anticipate occurrence AD in 1, 2, 3, and 4 resulting years (Park, Cho, Kim, Wall, Stern, Lim, Yoo, Kim and Cha, 2020). For foreseeing future rate of AD in adjusted examples (bootstrapping), the AI models showed sensible execution in the 1-year forecast with AUC of 0.775 and 0.759, because of "unequivocal AD" and "likely AD" results, separately; in 2-year, 0.730 and 0.693; in 3-year, 0.677 and 0.644; in 4-year, 0.725 and 0.683. The outcomes were comparative when the whole (uneven) tests were utilized. Significant clinical highlights chose in strategic relapse included haemoglobin level, age and pee protein level. This investigation may reveal an insight into the utility of the information-driven AI model-dependent for enormous scope authoritative wellbeing information in AD hazard forecast, which may empower better choice of people in danger for AD in clinical preliminaries or early discovery in clinical settings (Park, Cho, Kim, Wall, Stern, Lim, Yoo, Kim and Cha, 2020).

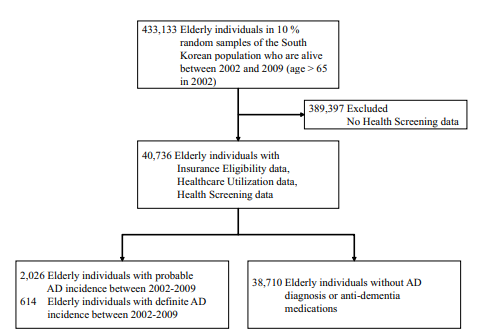


Figure 4 - Data Statistics (Park, Cho, Kim, Wall, Stern, Lim, Yoo, Kim and Cha, 2020)

Of 40,736 people with age over 65 years in 2002, the authors distinguished 614 one of a kind people with AD frequency utilizing the unmistakable AD result, 2026 with AD frequency utilizing the likely Promotion definition, and 38,710 seniors with no AD frequency. The pace of AD in this companion was 1.56% utilizing the positive AD definition, and 4.97% utilizing the plausible AD definition. Demographic attributes showed huge contrasts in age between both AD gatherings and non-AD gatherings and non-significant contrasts in pay and sex.

The outcomes were comparable when the authors utilized the whole, lopsided examples for model preparing and assessment, RF showed the best execution in anticipating a 0-year rate of AD with AUC of 0.887 when utilizing the positive AD definition and AUC of 0.805 when utilizing the likely AD definition. Order execution diminished as the anticipating time frame getting longer; utilizing the unequivocal AD definition, AUC of 0.781 (1 year), 0.739 (long term), 0.686 (long term), and 0.662 (long term); utilizing the likely AD definition, AUC of 0.730 (1 year), 0.645 (long term), 0.575 (long term), and 0.602 (long term). Quantities of highlights and think back periods likewise diminished in the latter year.

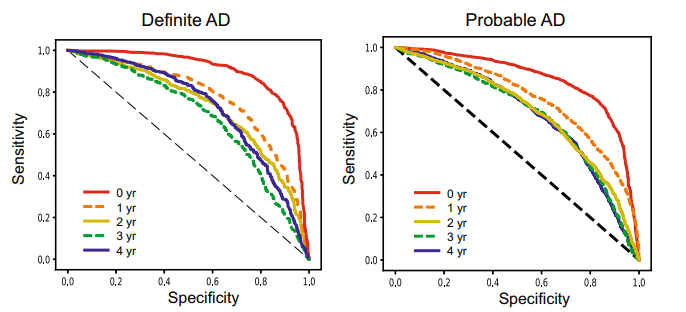


Figure 5 - Performances for Models Trained (Park, Cho, Kim, Wall, Stern, Lim, Yoo, Kim and Cha, 2020)

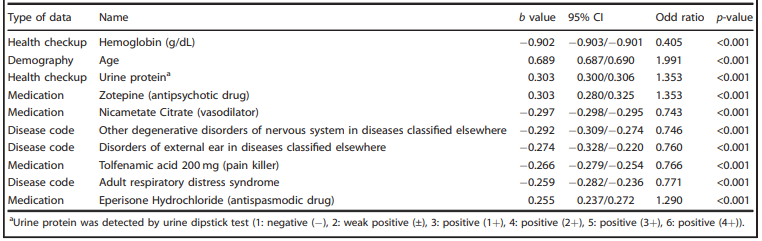


Figure 6 - Top Performing Features (Park, Cho, Kim, Wall, Stern, Lim, Yoo, Kim and Cha, 2020)

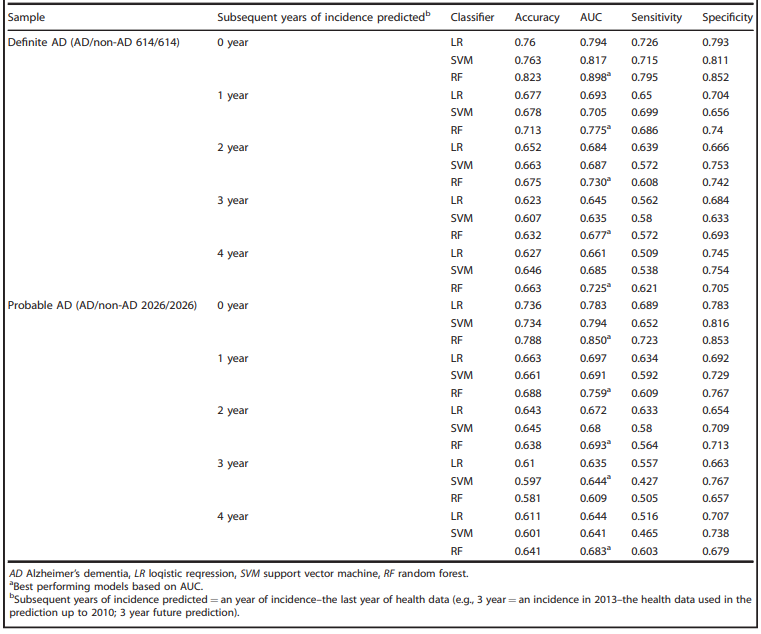


Figure 7 - Figure 6 - Performance of AD predictive models trained on NHIS-NSC by using balanced samples (Park, Cho, Kim, Wall, Stern, Lim, Yoo, Kim and Cha, 2020).

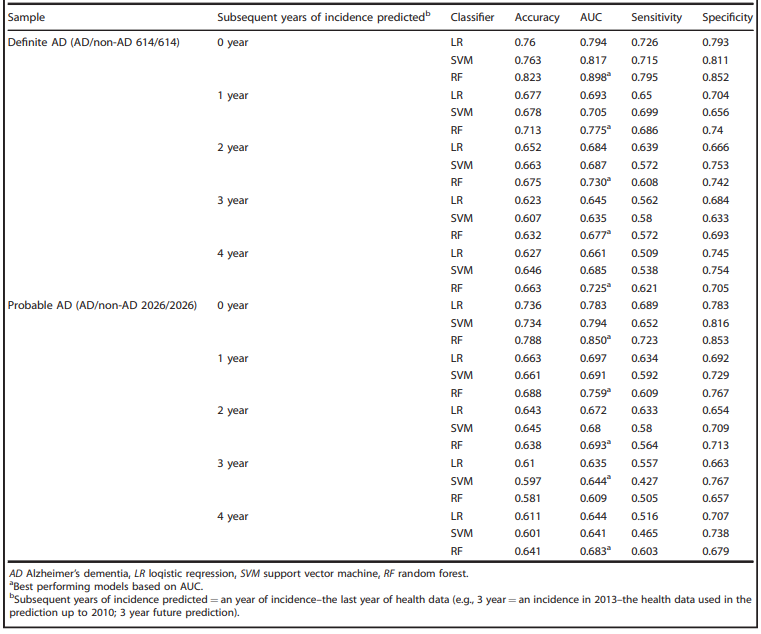


Figure 8 - Performance of AD predictive models trained on NHIS-NSC by using balanced samples (Park, Cho, Kim, Wall, Stern, Lim, Yoo, Kim and Cha, 2020).

The results from the above literature review show huge jumps and strides in implementing Machine Learning for the medical domain. The popular methods used in Machine Learning for the medical domain have shown great results and promise regarding the accuracy of their prediction. Moreover, it can be deduced that the only limitation of Machine Learning from the Literature Review is the limitation of data. As long as there exists data to feed to Machine Learning algorithms, there will be better results.

## Project Timeline

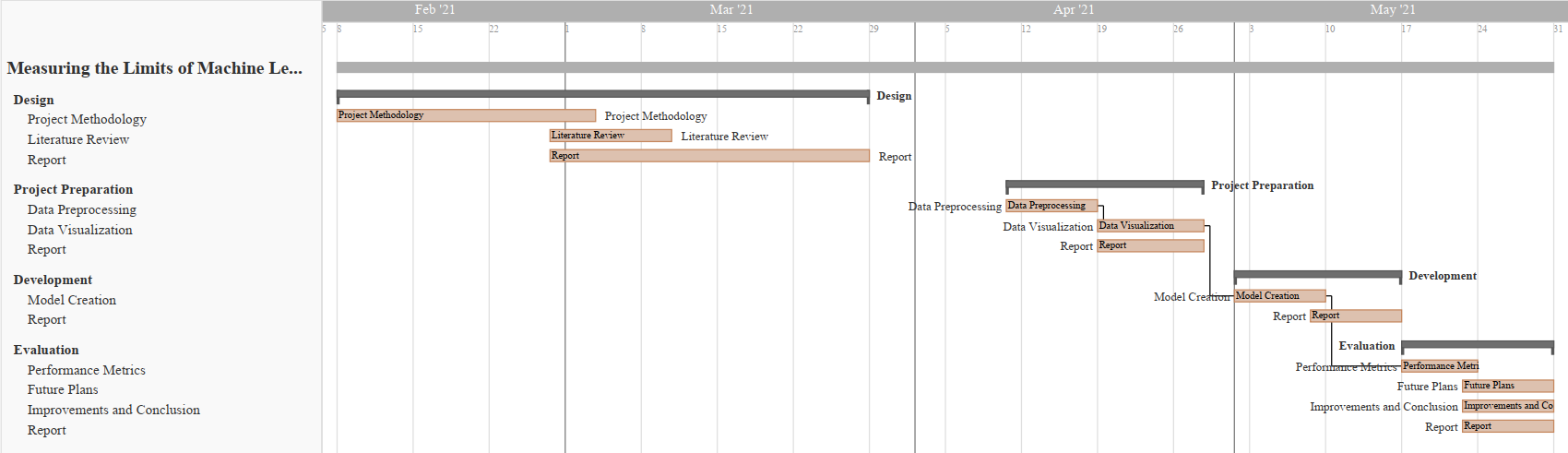


Figure 9 - Project Timeline

# Methodology

The dataset used will be the RAPID (Real-time Adaptive and Predictive Indicator of Deterioration) project that collects and analyses real-time patient data and alerts if the patient health is declining. The paradigm from the analysis of the dataset and the review of the Research Questions has been deduced to be Regression.

The features in the dataset are as follows:

* <Timestamp>: The time at which the data instance was recorded.
* <Lifetouch Heart Rate>: The heart rate of the Patient record.
* <Lifetouch Respiration Rate>: The respiration rate of the Patient record.
* <Oximeter SpO2>: Blood oxygen levels on the Patient record.
* <Oximeter Pulse>: Patient pulse recorded.

This data has been recorded in real-time from the patients. On observation, the dataset has issues like missing values, values that don’t follow the pattern i.e. -1. These issues have to be checked to increase the quality of the dataset and the results of the Machine Learning algorithm. Therefore, data pre-processing is mandatory for this project.

The number of features isn’t that many. Also, from observing the dataset, the paradigm of this machine learning project can be selected as regression. The dataset is time-series data providing diagnosis on the patient. This data will be used to predict the upcoming values for these features in advance. This will help predict if the patient’s health is going to decline or not. Regression algorithms will be implemented for this research project. Machine Learning libraries for Python like scikit-learn provide many Regression algorithms like ANN Regressor, SVM Regressor, ARIMA and many more. These algorithms will be used to train and predict this dataset.

Figure 10 - Feature Trend for the Dataset.

The dataset is certifiably not an enormous dataset with only four features to work with. The example line chart above is of the dataset. It shows how the four features in the dataset return to time. Every component is exceptional and fundamental for the dataset. The avocation this is that there are as of now fewer features to work with. The solitary possible data cleaning for this dataset will be to fill the missing characteristics. That will be done by the settled ML outfit made in this assessment.

# Project Evaluation

The project will be evaluated by:

* Performance Metrics for Regression Algorithms
* Quality of the Statistical Data Analysis

# Conclusion

The advent of technology has brought together medicine and machine learning. Researchers are trying to find better ways and hence develop the technology for the automation of patient prognosis. Various techniques exist but their limitations are not a well-known fact. This research focuses on measuring the limits of Machine Learning algorithms when it comes to patient prognosis based on Real-time Data from RAPID.

# References

* Venkatesh, R., Balasubramanian, C. and Kaliappan, M., 2019. Development of Big Data Predictive Analytics Model for Disease Prediction using Machine learning Technique. Journal of Medical Systems, 43(8).
* Santosh A. Shinde, M. and P. Raja Rajeswari, D., 2018. Intelligent health risk prediction systems using machine learning: a review. International Journal of Engineering & Technology, 7(3), p.1019.
* Park, J., Cho, H., Kim, J., Wall, M., Stern, Y., Lim, H., Yoo, S., Kim, H. and Cha, J., 2020. Machine-learning prediction of the incidence of Alzheimer’s disease using large-scale administrative health data. npj Digital Medicine, 3(1).
* Bengtsson, F. and Lindblad, K., 2021. Methods for handling missing values: A simulation study comparing imputation methods for missing values on a Poisson distributed explanatory variable. Uppsala University, [online] Available at: <https://www.diva-portal.org/smash/get/diva2:1520218/FULLTEXT01.pdf> [Accessed 5 February 2021].
* Jadhav, A., Pramod, D. and Ramanathan, K., 2019. Comparison of Performance of Data Imputation Methods for Numeric Dataset. Applied Artificial Intelligence, 33(10), pp.913-933.
* Jason Anastasopoulos, L. and G.Hunter, S., 2020. Fast and Robust Missing Data Imputation. researchgate.net, [online] Available at: <https://www.researchgate.net/publication/347741781\_Fast\_and\_Robust\_Missing\_Data\_Imputation> [Accessed 5 February 2021].
* Zhang, Z., 2016. Missing data imputation: focusing on single imputation. Annals of Translational Medicine, [online] 4(1). Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4716933/> [Accessed 5 February 2021].
* Amiri, M. and Jensen, R., 2016. Missing data imputation using fuzzy-rough methods. Neurocomputing, 205, pp.152-164.
* Bertsimas, D., Pawlowski, C. and Daisy Zhuo, Y., 2018. From Predictive Methods to Missing Data Imputation: An Optimization Approach. Journal of Machine Learning Research, pp.1-39.
* Duan, Y., Lv, Y., Liu, Y. and Wang, F., 2016. An efficient realization of deep learning for traffic data imputation. Transportation Research Part C: Emerging Technologies, 72, pp.168-181.
* Yoon, J., Jordon, J. and van der Schaar, M., 2018. GAIN: Missing Data Imputation using Generative Adversarial Nets. Proceedings of the 35th International Conference on Machine Learning.