# ABSTRACT:

This research aims in predicting the screening colonoscopy numbers across Ireland. Bowel screening aims to detect signs of bowel cancer at an early stage, where there are no symptoms. Bowel Cancer is currently available to all people aged 60 to 69 living in Ireland. The clients who have taken part in the program and whose faecal test shows higher haemoglobin count (than a defined level) will be sent for colonoscopy for polyp detection and other pathology tests to identify cancer. National Cancer Strategy 2021 recommends expansion of the age limit to 55-74. In order to assist in the decision making it is critical to understand the expected colonoscopy counts across the country. The study makes use of the existing screening colonoscopy for each age and gender and combined with the census data 2022, variations of the colonoscopies for each age and gender are identified. In-depth Interviews are done with the subject matter experts, the features that should be used for the study are discussed. Hypothesis testing are done to back this primary research. Co-relation between the features is obtained, several machine learning regression models are compared along with time-series regression models. The best model will be used to predict the colonoscopy numbers. The results and conclusions are provided in the relevant sections in the report.

# Introduction:

The topic for this research is machine learning predictions in bowel screening colonoscopy, the features impacting colonoscopy, statistical analysis, machine learning models for the predicting the colonoscopies. Machine Learning model development involves feature engineering where the features impacting colonoscopy numbers are technically identified. The research makes use of CRISP framework which involves the steps data gathering, data cleansing, data enrichment, statistical analysis, feature engineering, model building, analysing the model, predictions, and deployment. Most of the steps above are iterative in nature to get the best model for prediction. In-depth interviews are used as the primary research methodology for this research. Data enhancements are done based on the outcomes of this primary research. Hypothesis testing are applied to technically test the details suggested in the interviews. Correlation between the variables is verified, and different regression models are tested, accuracies are compared, and the most-appropriate model is used for predictions.

# Objectives:

1. To identify the variations in colonoscopies for each gender and defined age groups, across different months of a year to better plan the colonoscopy capacity in different units using descriptive statistics, furthermore, do a hypothesis testing to test if the colonoscopy counts for male population are less compared to females in order to assist the colonoscopy units better.

2. To identify the features that influence the colonoscopy counts, by undertaking appropriate co-relation studies between the features, and furthermore identify if there are any seasonal factors affecting the colonoscopy numbers.

3. To predict the colonoscopy numbers across Ireland by building an appropriate machine learning model, making use of the historical colonoscopy records as the secondary data source, and the eligible population including the extended age group from the latest census as the primary data source.

# Sampling Strategy:

It is proposed to use three different sampling strategies as part of the research, this includes both probabilistic (simple and stratified) sampling, this is primarily used while attaining the objective stated above. Non-probabilistic sampling (judgemental) is done to support the research and get inputs on the factors influencing the research.

This research project proposes to use two separate populations for achieving the above objectives. The first population are the males who have undergone colonoscopy across different age groups. The second population are the females who have undergone colonoscopy across different age groups.

Three different samples are used to achieve the above objectives and get some insights in addition to the literature review. In the first and second sample, each population(males/females) are divided into different age groups, statistical analysis is done on these age groups. Since the population selected is first grouped together for each age group, and a sample will be collected for further analysis, *stratified sampling* strategy is applied here. This is because the population is divided and grouped to different ages and then random selection is done. Since every record in the sample have undergone colonoscopy, simple sampling is applied on the selected sample. This is a probabilistic sampling, since every record in the sample have undergone colonoscopy, which means every unit has a chance. Also, in the sample the population is further grouped for each age group. The age group that has undergone colonoscopy as part of the colorectal is between 60-69 years. We will divide and group the data as between 60-63, 64-67 and 68-69 for the sake of our analysis. Sample records (30%) across the years, are then selected using the simple random sampling approach. Learnings from the above sampling is understood and applied on the second population for Females. Descriptive statistics are then applied to this selected samples so that the objective of identifying the max, min, average colonoscopies for the population is identified.

The third population that is proposed to be used in the research is the number of eligible clients based on the census 2022 data. *Probabilistic* sampling since is done on this population, since all the clients are equally eligible for colonoscopy. We will club the census data with the actual colonoscopy numbers and then apply suitable machine learning models on the sample records, we will use 70% of the data to train the model and the remaining 30% to test the model. *Simple* random sampling strategy is proposed to be used here since any random records can be selected for the research. This is because in this sample random sampling is done on the population without any age grouping.

Further, to better understand the dependencies of the research topic a mixture of judgement and convenience sampling is applied. Depth interview will be done as detailed in the primary research section below. There are several points in the research that is beyond the literature review and needs input from the experts. The experts with regards to the topic are readily available and convenient hence the strategy applied is convenience strategy. The experts selected would represent the population and hence the judgemental sampling strategy.

Ethical considerations to be aware of are the probability sampling are done considering three populations like male, female, based on the literature reviews. The data is from the clinical database and the issue with regards to the bias are taken care.

# Design and Methodology:

The project management methodology adopted in this research is a combination of KDD (Knowledge Discovery is Database) and CRISP-DM (Cross Industry Standard process for Data Mining).

The activities involved are research understanding that involves primary research and evaluation, data collection, data preparation, Data Analysis, enrichment by collecting additional data, descriptive statistical analysis, machine learning model building and comparison, predicting using the best

model

* 1. Research Understanding:

Consent

Information Leaflet

Identify the Participants

In-Depth Interviews

Interview Questionnaire

Research Understanding

Literature Review

Explore the data, do some basic checks.

Data Understanding

*Fig 1: Chart explaining the Research Understanding phase.*

Before we get into the data collection and further analysis, the research topics are explained in detail. The areas of Bowel Screening Colonoscopy, features impacting colonoscopy, understanding of census numbers for 2022, limitations of predicting colonoscopies based on past colonoscopy numbers, usage of machine learning models, descriptive statistics and its interpretation in health care terms are explored in detail. This is achieved through in-depth interviews with the subject matter experts. Participants are selected carefully considering their expertise in the field of research topic. The list of questions is prepared and sent to the participants, and direct questions are asked, and the points noted. Further research understanding is done by doing a literature review across different themes relevant to the topic.

* 1. Data Collection:

In the next step, based on the deeper understanding of the research topic relevant data is collected from the Bowel screen database. Only the required aggregate numbers are collected from the database. Approval from the data security officer is obtained. Ethical consideration by ensuring no patient personal data is extracted from the screening database for the purpose of this project. In addition to this public data from the CSO website obtained for analysis the 2022 census data.

*Make sure necessary approvals are taken and no personal sensitive information is used.*

Colonoscopy (Bowel Screen Data)

Research Understanding

Research Understanding

Census 2022 (CSO)

*Fig 2: Chart explaining the Data Collection phase.*

* 1. Data Preparation:

Data Cleansing

Aggregated Data

Exploratory Data Analysis / Machine Learning Model Building and Evaluation

Data Extraction

Time Series Data

Feature Engineering

Iterative process to achieve better results.

*Fig 3: Chart explaining the Data Preparation phase.*

The colonoscopy data used in this research will be extracted from the Bowel Screening database (SQL Server). SQL will be framed joining multiple tables, no sensitive personal data is required for thesis and hence will not be extracted. Census data will be obtained from the CSO website. The two sets of data will be saved in a .csv file and will be extracted in the jupyter notebook to be used for further analysis. Different data frames are created from this original dataset, the gender, months will be decoded to assist in the analysis. To do a statistical analysis different level of aggregations are applied on different fields. The results will be noted and graphically represented. Mean colonoscopies, mean of census population will need to be calculated and stored in a data frame. The data preparation is an iterative process, the data will need to aggregate grouped by age, gender for descriptive statistics. Quarterly / Monthly / Weekly counts of colonoscopy will be created in each data frame and used for time series analysis. Data preparation is further done for different regression models. Feature Engineering techniques are applied to tune the model in order to obtain higher efficiency in the models.

* 1. Data Analysis:

Exploratory Data Analysis

Visualisation

Data Preparation

Descriptive Statistics

Hypothesis Testing

*Fig 4: Chart explaining the Data Analysis phase.*

Data Analysis is done in every stage of this research. The data is first validated for completeness, NULL values are handled. Exploratory Data Analysis is done were the extreme, mean values are analysed. Descriptive statistics will be done and analysed. Hypothesis testing will be done in order to confirm the outcomes of the depth interview. PYTHON libraries are used to obtain a descriptive statistic, this is further analysed. Appropriate visualisation will need to be done for completing the analysis.

* 1. Model Building and Evaluation:

4.5.1 Time Series Analysis:

Check for Trend/Seasonality

Monthly/Quarterly Predictions using traditional models or random forest regressor.

Train – Until 2020-12-31

Test – From 2021-01-01

Hyper Parameter Tuning

Evaluate the model.

Time Series Analysis

Data preparation

*Fig 5: Chart explaining the Time Series Analysis*

The time series data will be used for prediction using time series. As mentioned in the data preparation section, data frames for time series will be created for month wise, quarter wise. Train and test data are selected (<2019-12-31 Train , >2019-12-31 – Test).

Trend and Seasonality is checked, and then traditional time series models (ARMA/ARIMA/SARIMA) are applied. The models are evaluated, and the best performing models will be selected for predictions. The model building and evaluations are iterative involving necessary hyper parameter tunings and further data preparations. The train/test/predictions are all plotted for visualisation.

4.5.2 Regression Data Models:

Regression models based on the features identified in the in-depth interviews are developed. The test data will be selected (30%) will be evaluated against the train data, and best performing models will be selected. Regression models like linear regression, Random Forest, Grid search CV, neural networks will be applied. The best performing model is evaluated. Necessary hyper tuning and further data preparation will be done in an iterative manner.

Create Data sets, Quarterly /Weekly, regression models using the features identified in the research understanding phase, and confirmed using the correlation matrix.

Hyper Parameter Tuning

Evaluate the model.

Regression Models

Data preparation

*Fig 6: Chart explaining the Regression Models*

# LITERATURE REVIEW:

## Introduction:

The project aims to predict the future colonoscopy numbers by applying appropriate machine learning models. The literature review aims to evaluate and understand the existing literature on the topics of screening as well application of machine learning to predict the screening numbers. Materials that suggest the external factors that impact the numbers predicted are also reviewed.

The materials reviewed are openly available on the internet, previous research papers, articles etc are reviewed as part of the process. Eighty-one documents are reviewed as part of the process in four themes. The documents show a wide range of areas with respect to the project. This includes review of documents relating to the research done in the field of Bowel screening programme. European guidelines are also reviewed so that the guidelines with respect to the colonoscopy are included in the review. Research papers from Scottish bowel screening programme and Swedish programme are reviewed. The programme involves testing of stool in its first step, this means there is a factor of disgust to be considered while estimating the counts. In order to understand the role of fear, embarrassment existing research for emotional prediction are also reviewed.

Application of Machine Learning to predict the intake in healthcare system is analysed by reviewing research that deals with machine learning for health services researchers, and a deeper understanding of how to interpret the machine learning models are reviewed.

By the end of this review, we get a clear understanding of the bowel screening standards used across the world, usage of machine learning in healthcare intake prediction, external factors that could influence bowel screening.

Ethical consideration with respect to literature review are considering wide range of topics in eight different themes related to the research in the technical areas with respect to machine learning or statistics, even in the area of domain of colonoscopy. Each of the themes have 10 literatures reviewed, that includes different points of view considered.

## Body Of Text:

Theme based approach of the organising principle is used in this review.

In the **first** theme, the guideline for colorectal screening is reviewed. [2] has a chapter on pathology with pan-European recommendations which consider the diversity and heterogeneity of health care systems across the EU. The present paper is based on the annex to the pathology chapter which attempts to describe in greater depth some of the issues raised in the chapter in greater depth, particularly details of special interest to pathologists.[3]has a detailed guidelines for each stage of the screening, there are several chapters with guidelines to each of the process in colorectal screening. The document states there are limited evidence that suggests the suitable, for the screening to be around 55-64. It also highlights that the screening above 74 is not recommended due to comorbidity and also not recommended to age below fifty.[3] literature here details on the evidence on effectiveness of the colorectal screening . Table1 here refers to the standards for each stage in screening, chapter 5 deals with the quality assurance in endoscopy, the chapter also deals with the complexities involved in the follow-up endoscopies. The literature concludes that the recommendations have the potential to enhance the control of colorectal cancer in Europe and elsewhere through improvement in the quality and effectiveness of the screening process that extends from systematic invitation to management of screen-detected cases.[4] The literature review of this document identifies the document has different sections where the cancer screening standards are defined. The document first gives a background and the need for cancer screening. The chapter 1 focusses on the effectives of screening. Different chapters are dedicated on defining the cancer screening process, need for quality assurance. The document also highlights that the screening is done on a healthier population and hence the radiation risks should be minimised. Chapter 8 and Chapter 9 are of particular interest to this project, where the management of lesions identified and requirement of further colonoscopy within a timeframe is mentioned. As per this it is understood that the colonoscopy numbers are not equal to the number of patients deemed to be suitable for colonoscopy. There is a standards table defined to identify the performance of screening. [5] also details the guidelines for colorectal screening. [6] has studied the incidence of subsequent colorectal cancer after the adenoma is detected. The study recommends the usage of Follow-up colonoscopy examinations (Surveillance). [7] the literature here analyses the prior interaction with screening of patients diagnosed with colorectal screening and factors associated with non-screening diagnosis. The study here is the screening performance in Scotland. There are statistics provided on the cancer diagnosed at screening. The study highlights few of the cases where cancer is identified in a younger age, recommending the age limit extension to even below 50 and above 75 for healthy patients. There were significant false negative cases here which requires further investigation. There is a flow chart that identifies the cases that were invited for screening, positive FIT tests, cancer detected. [8] This literature details of the once-only flexible sigmoidoscopy in cancer detection. This trial was undertaken in 14 UK centres. Randomisation by sequential number generation was done centrally in blocks of 12, with stratification by trial centre, general practice, and household type. The primary outcomes were the incidence of colorectal cancer, including prevalent cases detected at screening, and mortality from colorectal cancer. Detailed statistics on the number of clients invited randomisation of the clients done for sigmoidoscopy. There is a flow chart explaining the randomisation. Table explaining colorectal cancer incidence and mortality in control and intervention groups. Kaplan-Meier estimates of cumulative incidence and mortality Colorectal cancer incidence, distal cancer incidence, proximal cancer incidence (E and F), and colorectal cancer mortality . Curves are truncated at 10 years of follow-up because of incomplete ascertainment of cancers in the final calendar year of the study. Findings from this large, randomised trial have shown that both incidence of and mortality from colorectal cancer are significantly reduced in people undergoing a single flexible sigmoidoscopy examination between 55 and 64 years of age. A limitation of the trial is that rather than inviting the whole population aged 55–64 years for screening, the trial used a two-stage recruitment procedure whereby eligible individuals were randomly assigned only if they responded to a questionnaire and indicated that they would be likely to attend screening. [9]The literature review presents the conflicting guidelines with respect to the colorectal screening. The study here determines that the timely treatment of colorectal screening is a quality indicator of oncological care. The study does an extensive literature review through MEDLINE, EMBASE and Cochrane databases, complemented with a web search and a survey. The results of the case study define the treatment interval. It confirms that that the interval from diagnosis to treatment could be used for prehabilitation to benefit patient recovery. [10]The literature review of the mentioned journal extracted from British journal of Cancer identifies the factors influencing time from presentation to treatment of colorectal cancer. The study understands that the stage at diagnosis and survival from cancer vary according to where people live. The aim of the study here was to determine if there is a delay according to where people live. Statistics of the patients selected for study, proximity with the GP, the differences are all noted. The study didn’t find any significant evidence to suggest people leaving far from treatment centres did not receive delayed treatment. There are several tables and charts to claim this. (Ponti et al., 2020) The literature tried to address the key issues that need to be considered while revising the current annex of the European Council recommendation (2003) on cancer screening. The literature summarises the recommendations provided the European Union along with the issues to be considered in the revised recommendation.

In the **second** theme, we will understand the process, strategies, performance, and any special considerations done in the screening programmes are worldwide. [11]. This research paper briefs about the screening programme worldwide, the pilot program implemented in Ireland (2008-2009) and its outcomes are highlighted. There is a chart on how the Americas and Asia pacific countries performed in the screening, also a mention on the countries with high cancer rate have not yet implemented screening. Literature from national/international papers reviewed here. As per the paper participation is highest amongst women, but the men showed highest percentage of FIT results being positive. [12] The literature deals with the clinical trial enrolment, patient characteristics, and survival differences in the screening registered patients. As part of the study a total of 760 mCRC patients referred for their first oncological consideration at 3 hospitals in Scandinavia covering defined populations were registered consecutively during 2003 to 2006. Clinical trial enrolment, patient characteristics, and treatment were recorded prospectively, and the follow-up was complete. chemotherapy was initiated in 61% of the patients. Approximately one-third (36%) of patients receiving chemotherapy were included in a trial. The main reason for nonparticipation was failed eligibility criteria (69%). The median survival after chemotherapy was 15.8 months for all patients, and 18 months after combination chemotherapy. Trial patients had better prognostic characteristics and significantly longer survival than non-trial patients: 21.3 months versus 15.2 months when receiving combination chemotherapy. Poor performance status was the main reason for giving best supportive care only, and the median survival was then only 2.1 months. The median survival for all 760 nonresectable mCRC patients was 10.7 months. There are detailed discussions on the clinical trials. There are charts explaining the cumulative survival over the months.[13] highlights the significance of colorectal screening . It mentions the low numbers in screening in spite of evident advantages Throughout the world there are widespread differences in CRC screening implementation status and strategy. Differences can be attributed to geographical variation in CRC incidence, economic resources, healthcare structure and infrastructure to support screening such as the ability to identify the target population at risk and cancer registry availability. This review highlights issues to consider when implementing a CRC screening programme and gives a worldwide overview of CRC burden and the current status of screening programmes, with focus on international differences. [14] reviews the efficacy in standard clinical practice of colonoscopic polypectomy in reducing colorectal cancer incidence. The literature states that the colorectal cancer is one of the leading causes of death in western countries. The literature details on the method used to select the population and samples. Age group of patients are also defined. The results identify the follow-up data as well as the cancer identified cases. The literature concludes that colonoscopic polypectomy substantially reduced the incidence of colorectal cancer in the cohort compared with that expected in the general population. These results are of particular relevance considering that those with adenomas are at increased risk of colorectal cancer and that this retrospective study was performed on data obtained in standard clinical practice. This observation strengthens the concept of effective population screening in view of the fact that adenomatous polyps are the most frequent neoplastic outcome of screening, and their removal is associated with a decrease in the incidence of colorectal cancer. Table 1 highlights the characteristics of the study cohort and the index lesions. There is another table explaining the characteristics of the six patients with colorectal cancer detected at follow up. [15]this research paper published in world journal of surgical oncology, is a study of laparoscopic versus open colectomy for colon cancer in an older population. The patients aged 65 years or older with cancer diagnosed and underwent colectomy within 6 months of diagnosis were considered for this study. Laparoscopy and colectomy patients were compared on different parameters. Laparoscopic colectomies were associated with left-sided tumours; areas with population density, income, and education level; areas in the western United States; and National Cancer Institute-designated cancer centres. Laparoscopic colectomy cases had shorter length of stay and less intensive care unit monitoring. Table 1 shows the baseline characteristics in overall cohort by surgical approach. There is a section explaining the limitations of the study. There is a chart that shows the survival times and %survived over the years. The study concludes that laparoscopic colectomy practice patterns were associated with factors which likely correlate with tertiary referral centres. Also, the study shows shorter hospitalizations for laparoscopic colectomy. [6] is an article from New England Journal of Medicine. The background of the research is to identify the long-term mortality after screening for colorectal cancer. The method of identifying the clients and approach of screening is defined clearly. There is a table that shows characteristics of the participants at Baseline and study end points. There is a figure 1. Cumulative Colorectal-Cancer Mortality shows the cumulative cancer mortality over the years. The research has a conclusion showing the percentage of mortality, percentage of mortality due to cancer. Screening reduced colorectal-cancer mortality. The study concludes that the of screening with faecal occult-blood testing on colorectal-cancer mortality persists after 30 years but does not influence all-cause mortality. The sustained reduction in colorectal-cancer mortality supports the effect of polypectomy. (Funded by the Veterans Affairs Merit Review Award Program and others.). [16], the main aim of this study was to present data from a large, contemporary, population-based cohort of patients with colonic cancer. Data on patients diagnosed between 2007 and 2011 were extracted from the Swedish Colon Cancer Registry. The data, registered prospectively in a national population of almost 10 million, included over 99 per cent of all diagnosed adenocarcinomas of the colon. analysis included 18 889 patients with 19 526 tumours (3·0 per cent had synchronous tumours). The sex distribution was fairly equal, and the median age was 74·1 (interquartile range 65–81) years. The overall and relative (cancer-specific) survival rates after 3 years were 62·7and71·4percent respectively. Some 88·0 per cent of the patients were operated on, and 83·8 per cent had tumours resected. Median blood loss during bowel resection was 200 (mean 311) ml, and the median operating time was 160 min; 5·6 per cent of the procedures were laparoscopic. Preoperative chemotherapy was administered to 2·1 per cent of patients; postoperative chemotherapy was planned in 90·1 per cent of fit patients aged less than 75 years with stage III disease. In patients operated on in an emergency setting (21·5 per cent), the preoperative evaluation was less extensive, the proportion of R0 resections was lower, and the outcomes were poorer, in both the short and long term. There is tabular representation of the result summarised. [17] studies about rectal washouts and local recurrence of cancer after anterior resection. Adenocarcinomas of the rectum shed viable cells, which have the ability to implant. Intraoperative rectal washout decreases the amount and viability of these cells, but there is no conclusive evidence of the effect of rectal washout on local recurrence after rectal cancer surgery. Data were analysed from a population-based registry of patients who had anterior resection from 1995 to 2002 and were followed for 5 years. Rectal washout was performed at the discretion of the surgeon. National inclusion of patients with rectal cancer and follow-up was near complete (approximately 97 and 98 per cent respectively). There is a flow chart explaining the methodology with clear details on the numbers in each stage and the final results. There is a table in the literature that explains the characteristics of the included patients in detail. Multivariable logistic regression analysis is performed on the data. The study concludes that there is a more favourable outcome in patients after rectal washout than without.[18] The literature here reviews the surgical treatment of colon cancer in patients aged 80 years and old. The study aimed to describe the patterns of surgery in patient aged 80 plus, and the outcomes are examined. Medicare beneficiaries aged 80 years with colon cancer who were diagnosed between 1992 and 2005 were identified from the Surveillance, Epidemiology, and End Results-Medicare database. Multivariable logistic regression analysis was used to assess factors associated with nonoperative management. Kaplan-Meier survival analysis determined 1-year overall and colon cancer-specific survival. The literature details about the variables used in the research, it also details about the patient selection process in the status. The literature mentions of the outcome variable “curative surgery”. There is a section for the statistical analysis, where the hypothesis testing is all mentioned. Figure 1 shows as Surgical treatment of patients aged 80 years with colon cancer is shown. The results show that of the selected samples 80% has undergone colectomy. Factors found to be most predictive of nonoperative management included older age, black race, more hospital admissions, use of home oxygen, use of a wheelchair, being frail, and having dementia. For both operative and nonoperative patients, the 1-year overall survival rate was lower than the colon cancer-specific survival rate (operative patients: 78% vs 89%; nonoperative patients: 58% vs 78%). The literature concludes that the majority of older patients with colon cancer undergo surgery, with improved outcomes noted compared with nonoperative management. However, many patients who are not selected for surgery die of unrelated causes, reflecting good surgical selection. Patients undergoing surgery during an urgent/emergent admission have an increased short-term mortality risk. Because the earlier detection of colon cancer may increase the percentage of older patients undergoing elective surgery, the findings of the current study may have policy implications for colon cancer screening and suggest that age should not be the only factor driving cancer screening recommendations. Cancer 2013; 119:639-47. C V 2012 American Cancer Society.

In the **third** theme, we will focus on usage of Machine Learning to predict the values in health care.[19]. The literature highlights the usage of Machine learning models in health care. The study here focusses on a machine learning based predictive model to identify acute coronary syndrome outcomes and mortality of patients in hospital. Machine learning based soft-voting ensemble classifier (SVEC) for the predictive modelling of acute coronary syndrome (ACS) outcomes such as STEMI and NSTEMI. The proposed SVEC applied 3 ML algorithms such as random forest, extra tree, and the gradient-boosting machine for predictive modelling of target and the accuracy measures were compared. The proposed predictive model outperformed other ML-based models; hence it can be used practically in hospitals for the diagnosis and prediction of heart problems so that timely detection of proper treatments can be chosen, and the occurrence of disease predicted more accurately. There are tabular data explaining the statistics and figure explaining the model outcomes. [20] The document summarises on how the different machine learning models can be interpreted. Population based interpretation and sample interpretation are both discussed. There is a graphical representation of how the interpretation is done for population and for the sample.[21] The literature review of this research document finds a deeper insight on the machine learning techniques to be used on a health care data. We see that the research was done on the insurance claims data with data focusing on age, gender, and hospitalization. The estimators to predict a binary variable and the regression data is discussed. There is a tabular comparison of the frequently used models provided. Concepts of regularization, cross validations are discussed. Brief explanations of usage of Deep Learning neural network are also provided.[22]This literature is published in the journal of economic perspectives vol.31 . The literature highlights the application of Machine learning in a wide range of applications. The literature describes the simplicity of machine learning models using technologies like PYTHON/R. Table 1 summarizes the findings of applying various procedures to this problem. The research paper also details different machine learning models, their performance metrics. The literature also does a specific focus on the application of Machine Learning in the field of economics. There is a detailed text on how the machine learning model can be implemented. The literature then concludes by stating that turning the problem to inductive, that is allowing the data to decide the rules works best in case of machine learning.[23]Literature details the trends ,perspectives and prospects of Machine Learning. The literature details that Machine learning addresses the question of how to build computers that improve automatically through experience. It is one of today’s most rapidly growing technical fields, lying at the intersection of computer science and statistics, and at the core of artificial intelligence and data science. Recent progress in machine learning has been driven both by the development of new learning algorithms and theory and by the ongoing explosion in the availability of online data and low-cost computation. The adoption of data-intensive machine-learning methods can be found throughout science, technology, and commerce, leading to more evidence-based decision-making across many walks of life, including health care, manufacturing, education, financial modelling, policing, and marketing.[24]This case study in the field of data mining highlights the importance of features and effects with dependent features The literature highlights of feature importance in machine learning models is challenging when features are dependent. Permutation feature importance (PFI) ignores such dependencies, which can cause misleading interpretations due to extrapolation. A possible remedy is more advanced conditional PFI approaches that enable the assessment of feature importance conditional on all other features. Due to this shift in perspective and in order to enable correct interpretations, it is beneficial if the conditioning is transparent and comprehensible. In this paper, it is proposed a new sampling mechanism for the conditional distribution based on permutations in conditional subgroups. As these subgroups are constructed using tree-based methods such as transformation trees, the conditioning becomes inherently interpretable. This not only provides a simple and effective estimator of conditional PFI, but also local PFI estimates within the subgroups.[25] This paper provides a review and commentary on the past, present, and future of numerical optimization algorithms in the context of machine learning applications. Through case studies on text classification and the training of deep neural networks, how optimization problems arise in machine learning and what makes them challenge are discussed. A major theme of our study is that large-scale machine learning represents a distinctive setting in which the stochastic gradient (SG) method has traditionally played a central role while conventional gradient-based nonlinear optimization techniques typically falter. Based on this viewpoint, versatile SG algorithm, discuss its practical behaviour, and highlight opportunities for designing algorithms with improved performance. This leads to a discussion about the next generation of optimization methods for large-scale machine learning, including an investigation of two main streams of research on techniques that diminish noise in the stochastic directions and methods that make use of second-order derivative approximations.[26] Post-analysis of predictive models fosters their application in practice, as domain experts want to understand the logic behind them. In epidemiology, methods explaining sophisticated models facilitate the usage of up-to-date tools, especially in the high-dimensional predictor space. Investigating how model performance varies for subjects with different conditions is one of the important parts of post-analysis. This paper presents a model-independent approach for post analysis, aiming to reveal those subjects’ conditions that lead to low or high model performance, compared to the average level on the whole sample. Conditions of interest are presented in the form of rules generated by a multi-objective evolutionary algorithm (MOGA). In this study, Lasso logistic regression (LLR) was trained to predict cardiovascular death by 2016 using the data from the 1984–1989 examination within the Kuopio Ischemic Heart Disease Risk Factor Study (KIHD), which contained 2682 subjects and 950 preselected predictors. After 50 independent runs of fivefold cross-validation, the model performance collected for each subject was used to generate rules describing “easy” and “difficult” cases. LLR with 61 selected predictors, on average, achieved 72.53% accuracy on the whole sample. However, during post-analysis, three categories of subjects were discovered: “Easy” cases with an LLR accuracy of 95.84%, “difficult” cases with an LLR accuracy of 48.11%, and the remaining cases with an LLR accuracy of 71.00%. Moreover, the rule analysis showed that medication was one of the main confusing factors that led to lower model performance. The proposed approach provides insightful information about subjects’ conditions that complicate predictive modelling.[27]This article discuss about the potential application of Machine Learning in Health Outcomes Research and some statistical cautions . Traditional analytic methods are often ill-suited to the evolving world of health care big data characterized by massive volume, complexity, and velocity. In particular, methods are needed that can estimate models efficiently using very large datasets containing healthcare utilization data, clinical data, data from personal devices, and many other sources. Although very large, such datasets can also be quite sparse (e.g., device data may only be available for a small subset of individuals), which creates problems for traditional regression models. Many machine learning methods address such limitations effectively but are still subject to the usual sources of bias that commonly arise in observational studies. Researchers using machine learning methods such as lasso or ridge regression should assess these models using conventional specification tests.[28] The study aims to compare the of standard versus enhanced models to predict future high-cost patients, especially those who move from a lower to the upper decile of per capita healthcare expenditures within 1 year—that is, ‘cost bloomers’. 6 alternative models are developed and compared with the traditional model, with 1053 traditional features. Population of Denmark to predict future high-cost patients and characterise high-cost patient subgroups. Using unseen data from a future year, each model’s prospective predictive performance by calculating the ratio of predicted high-cost patient expenditures to the actual high-cost patient expenditures in Year 2—that is, cost capture is evaluated. The best enhanced model achieved a 21% and 30% improvement in cost capture over a standard diagnosis-based model for predicting population-level high-cost patients and cost bloomers, respectively.

In combination with modern statistical learning methods for analysing large data sets, models enhanced with a large and diverse set of features led to better performance—especially for predicting future cost bloomers.

The **fourth** theme for the literature review is the external factors affecting the machine learning predictions.[29] The literature review of the above document understands the influence of age, gender, and deprivation on the cancer program. The data collected is between 2007 – 2014. Uptake which is measured by the number of faecal samples returned to the screening program, positivity rate, cancer detection rates are used as measures for this research. The results are well presented in few tabular formats. Table 1 shows the invitations for age and deprivation, Table 2 shows the numbers completed the screening round. There is a flow chart clearly showing the division of cancer detected. Results show from 5,308,336 screening episodes checked, younger males 50-65 is lower than women, however there is an increase in numbers for the age group 65-69. Explanations on the reasons of this is also provided in the document. [30] This literature aims to understand why people do not always engage in medical examinations that might benefit them is a public health issue which is receiving increased attention. One area of promise involves the study of medical embarrassment, although current studies are weakened in that they measure medical embarrassment in a theoretically naïve and unidimensional manner and have assumed that embarrassment is exclusively a barrier to the timely seeking of treatment. Convenience sampling was used as part of the research. Participants completed a comprehensive measure of medical embarrassment, reported on previous treatment avoidance because of embarrassment, and recorded the frequency of psychological, general, and sex-related visits across the previous 5 years. The results show that as expected, medical embarrassment was not unidimensional and appeared to have two distinct factors – bodily embarrassment and judgment concern. Bodily embarrassment generally predicted less frequent medical contact although not equally so across domains and it interacted with judgment concern in several cases, providing preliminary evidence that there are situations in which aspects of medical embarrassment may actually facilitate greater medical contact.  The data highlight the importance of considering the role of emotions other than fear in health behaviour and the means by which they may facilitate or deter the timely seeking of diagnosis and treatment.[31]The journal aims to assess the effect of gender , age and deprivation on key performance indicators in a colorectal cancer screening programme. In this literature the relevant populations were subdivided by gender into our age groups and into five deprivation categories according to the Scottish Index of Multiple Deprivation (SIMD), and key performance indicators analysed within these groups. The results are very clearly mentioned with the p scores obtained in the statistical analysis. It is then concluded in this population-based colorectal screening programme gender, age, and deprivation had marked effects on key performance indicators, and this has implications both for the evaluation of screening programmes and for strategies designed to reduce inequalities. [32] The literature review of the above research understands the process, approach, and result of the research. There is a detailed explanation of the interview approach and questionnaire details. The document concludes that medical, demographic factors, fear, embarrassment, disgust are the factors that could influence the bowel screening numbers. The literature concludes that alongside medical and demographic factors, fear, embarrassment, and disgust are worthy of consideration in colorectal cancer screening. Understanding how specific emotions impact screening decisions and behaviour an important direction for future work is and has potential to inform screening development and communications in bowel health.[33] this research paper aims to document the association between health literacy and willingness to screening. The literature explains the methods undergone the research. A [multivariate analysis](https://www.sciencedirect.com/topics/medicine-and-dentistry/multivariate-analysis) supported the hypothesis that lower health literacy would be associated with less information-seeking. Conclusion that Lowers health literacy had a direct impact on information-seeking. It was also independently associated with perceived confidence to participate in screening. [34] this literature studies the inequalities in participation in an organized national colorectal cancer screening programme. The literature explains the methodology for client selection, usage of linear regression mentioned. The results showed a 54% uptake, but showed a gradient across quintiles of deprivation, ranging from 35% in the most deprived quintile to 61% in the least deprived. Multivariate analyses confirmed an independent effect of deprivation, with stronger effects in women and older people. Statistical analysis and hypothetical testing results are provided in the table. It is then concluded that the overall uptake rates in this organized screening programme were encouraging, but nonetheless there was low uptake in the most ethnically diverse areas and a striking gradient by SES. Action to promote equality of uptake is needed to avoid widening inequalities in cancer mortality. [35] The literature studies the specific and differential prediction of health anxiety by disgust sensitivity and propensity. The literature learns that the Current models of health anxiety suggest that fear resulting from false alarms to perceived threats to one's health results in the development of hypochondriasis and related disorders. Disgust has been proposed as an effective response that may function as an etiological and maintenance mechanism in health anxiety. Moreover, the way in which an individual perceives the disgust response (disgust sensitivity) may affect health anxiety, separately from their likelihood of experiencing disgust (disgust propensity). The present study utilized multiple hierarchical regression analysis to investigate the degree to which self-reported disgust sensitivity and disgust propensity differentially predict elevated health anxiety in a sample of 620 non-treatment-seeking undergraduates. Further, this effect is tested in comparison to that of anxiety sensitivity, a construct demonstrated to be strongly related to health anxiety. Analyses indicate that disgust sensitivity, rather than disgust propensity, is primarily responsible for this relation. An additional analysis tested the specificity of disgust sensitivity relative to anxiety sensitivity. Disgust sensitivity was no longer significant after including anxiety sensitivity in the model. Suggestions for further evaluation of this relation are provided. These results suggest that although disgust sensitivity may appear related to health anxiety, this relation may be confounded by anxiety sensitivity.[36]This literature was published in the British journal of cancer (2011). The document studies the changing association between socioeconomic circumstances and the incidence of colorectal cancer. It is concluded that the deprivation was associated with higher incidence rates of male, but not female, colorectal cancer before the implementation of a national bowel screening programme. [37] learns about the impact of age, sex, and socioeconomic deprivation on outcomes in a colorectal cancer screening programme. The literature explains the statistical evidence of lower uptake of screening was associated with younger age, male sex, and deprivation. It is concluded that the individuals who are deprived are less likely to participate in screening, less likely to undergo colonoscopy and less likely to have cancer identified as a result of a positive test. Therefore, this study suggests that strategies aimed at improving participation of deprived individuals in colorectal cancer screening should be directed at all stages of the screening process and not just uptake of the test.

The **fifth** theme for the literature review is the descriptive statistics and its interpretation.[38] This ebook studies on data analysis and its interpretation. The experiment data from the operation is input to the analysis and interpretation. After collecting experimental data in the operation phase, we want to be able to draw conclusions based on this data. To be able to draw valid conclusions, we must interpret the experiment data.[39] The article discusses about the application of Analysis of Variance (ANOVA) in the Interpretation of power transformer faults. Significance of usage of ANOVA to test null hypothesis is explained and the fundamental principal of ANOVA is discussed. Case studies are discussed in this literature. There is a table showing the descriptive statistics comparison. ANOVA was proposed to evaluate the statistical significance. (Mundy et al., 2017) This literature reviews the breast cancer data and its data interpretation. The literature also explains the methods of data collection for this study. The results are explained with detailed statistics. Clinical reference point for data interpretation is provided in the conclusion.[40] This literature studies the descriptive statistics and the normality tests for statistical data. The literature understands that descriptive statistics are an important part of biomedical research which is used to describe the basic features of the data in the study. They provide simple summaries about the sample and the measures. Measures of the central tendency and dispersion are used to describe the quantitative data. For the continuous data, test of the normality is an important step for deciding the measures of central tendency and statistical methods for data analysis. When the data follow normal distribution, parametric tests otherwise nonparametric methods are used to compare the groups. There are different methods used to test the normality of data, including numerical and visual methods, and each method has its own advantages and disadvantages. In the present study, the summary measures and methods used to test the normality of the data are discussed.[41] this literature discusses on the descriptive statistics, usage and impacts on data analysis and its interpretation. Study is based in tree crown condition and is done for United states department of Agriculture. Methods of data collection, data summary are provided in detail. The detailed results with respect to the tree crown by statistical analysis is explained .[42] This is a ebook which is a handbook of data analysis and interpretation with SPSS. The book is very much detailed and explains different steps of data gathering, statistical analysis. Separate chapter is dedicated to ANOVA. This books looks great as a guide foe data analysis wanting to do statistics, regression, understand hypothesis testing and implement them, interpret the results etc. [43] This literature reviews the IBM SPSS for Statistics, use and Interpretation . This is a handbook that explains the basics of data analysis and statistics. Variables, Research problems are discussed in detail. There is a chapter dedicated to Measurement and descriptive statistics. Frequency distribution, charts and their interpretation are explained in detail. There is a chapter dedicated to data understanding and interpreting inferential statistics. Corelation, regression , NAOVA testing , explained by discussing with a sample problem .This book looks excellent for a data analysis student and also professional working on statistics, machine learning, analytics.[44] This is an older literature from Morgan , the 2019 article above was very detail and much latest. The literature, studies, examples are similar, the 2019 version is much more latest and worth using them. [45] suggests that histograms, box plots and cumulative distribution graphs are popular graphic representations for statistical distributions. The main research question that this study focuses on is how college students deal with interpretation of these statistical graphs when translating graphical representations into analytical concepts in descriptive statistics. This study is divided into two parts. The research sample included 256 college students in the first part and 187 college students in the second part. The research tools were questionnaires dealing with the interpretation of the graphs, while relating the graphs and other concepts in descriptive statistics. In spite of the benefits learners may reap from using multiple representations, the results reveal that some of the students had difficulties in relating multiple representations to the same data. Educators have to consider that only deep understanding of each and every one of the representations and their inter-relation will enable students to successfully translate one format into another. Some of the mistakes students made could be derived from the use of the intuitive rule known as the Same A-Same B. [46] learns the usefulness of descriptive statistics in the interpretation of data. The conclusion is that in data analysis it is necessary to define quantities to provide a fuller picture of distributions.

In the sixth **theme**, we will discuss the topic of feature engineering in machine learning. [47]. In this literature feature engineering and selection for predictive models are discussed. There are details on the literature review done as part of the studies , the chapters on different literatures are explained in detail.[48] This book deals with feature engineering for machine learning and data analytics . There is an overview of Feature Engineering, machine learning and data analytics. Feature Engineering for Text Data, feature extraction and learning for Visual Data is also part of the study in this book. All the chapters have the detailed graphs to explain the concepts. Overall the literature is very detailed on the concepts of feature engineering. [49] . The literature highlights the place of feature engineering in the machine learning workflow. Feature Engineering is to be applied during data cleansing and transformation. The book highlights how the feature engineering can solve the deficiency problem.[50] This a book explaining details of feature engineering, its introduction, applications, usage etc. [51] discussed on the machine learning based framework to identify type 2 diabetes. The goal of this work was to develop a semiautomated framework based on machine learning as a pilot study to liberalize filtering criteria to improve recall rate with a keeping of low false positive rate. The study made use of different machine learning models including k-Nearest-Neighbors, Naïve Bayes, Decision Tree, Random Forest, Support Vector Machine and Logistic Regression. The top performing ML algorithms are applied on the engineered features. Results indicate that the framework achieved high identification performances with feature engineering inputs.[52] integrates machine-learning algorithms into the descriptor-based design approach for rapid screening of transition-metal catalysts. By engineering numerical representation of surface metal atoms using easily accessible features such as the local electronegativity and the effective coordination number that are dependent on the surroundings of an adsorption site, together with the intrinsic properties of active metal atoms including the electronegativity, ionic potential, and electron affinity, the machine-learning model optimized with ∼250 *ab initio* adsorption energies on bimetallic alloys can capture complex, non-linear adsorbate/substrate interactions with the root mean squared errors (RMSE) ∼0.12 eV. Compared with the traditional high-throughput computational and experimental trial-and-error approach, the machine-learning [chemisorption](https://www.sciencedirect.com/topics/chemical-engineering/chemisorption) models have great potential in accelerating the discovery of interesting catalytic materials. As the complexity of catalyst structures increases, new features will be needed to learn underlying correlations and avoid introducing significant errors on top of the average DFT prediction errors expected with standard semi-local generalized gradient approximation (GGA) functionals. (Upadhyay et al., 2021) This literature learns that most of the machine learning techniques finetune the hyper-parameters to improve the detection rate, our approach focuses on selecting the most promising features of the dataset using Gradient Boosting Feature Selection (GBFS) before applying the classification algorithm, a combination which improves not only the detection rate but also the execution speed. GBFS uses the Weighted Feature Importance (WFI) extraction technique to reduce the complexity of classifiers. Various decision-tree based machine learning techniques are evaluated after obtaining the most promising features of the power grid dataset through a GBFS module and show that this approach optimizes the False Positive Rate (FPR) and the execution time. [53] understands that deep learning is an emerging transformative tool in diagnostic medicine, yet limited access and the interpretability of learned parameters hinders widespread adoption. Here a diverse repository of 838,644 histopathologic images is generated and used to optimize and discretize learned representations into 512-dimensional feature vectors. Importantly, we show that individual machine-engineered features correlate with salient human-derived morphologic constructs and ontological relationships. Deciphering the overlap between human and machine reasoning may aid in eliminating biases and improving automation and accountability for artificial intelligence-assisted medicine.[54] this literature makes use of feature engineering, explains the importance of feature engineering at data cleansing stage so that the state of the art tools can provide great results. This demonstrates that the approach significantly enhances bacteriophage discovery, and this provides a new starting point for exploring new biologies. [55] This literature on the art of feature engineering: Essentials of Machine Learning, explains in detail. This a book published in University of Cambridge press. The contents are very detailed and have charts/graphs to back a concept and interpret. Overall, a great book for machine learners.

In the **seventh** theme, we will discuss on the corelation between the features in Machine Learning.[56] In this paper, it is proposed that a hybrid machine learning system based on Genetic Algorithm (GA) and Support Vector Machines (SVM) for stock market prediction. A variety of indicators from the technical analysis field of study are used as input features. The correlation between stock prices of different companies to forecast the price of a stock, making use of technical indicators of highly correlated stocks, not only the stock to be predicted is used here. The genetic algorithm is used to select the set of most informative input features from among all the technical indicators. The results show that the hybrid GA-SVM system outperforms the stand alone SVM system.[57] Correlation analysis is an extensively used technique that identifies interesting relationships in data. These relationships help us realize the relevance of attributes with respect to the target class to be predicted. This study has exploited correlation analysis and machine learning-based approaches to identify relevant attributes in the dataset which have a significant impact on classifying a patient’s mental health status. For mental health situations, correlation analysis has been performed in Weka, which involves a dataset of depressive disorder symptoms and situations based on weather conditions, as well as emotion classification based on physiological sensor readings. Pearson’s product moment correlation and other different classification algorithms have been utilized for this analysis. The results show interesting correlations in weather attributes for bipolar patients, as well as in features extracted from physiological data for emotional states.[58] As per this, Feature selection is an effective strategy to reduce dimensionality, remove irrelevant data and increase learning accuracy. The curse of dimensionality of data poses a severe challenge to many existing feature selection methods with respect to efficiency and effectiveness. In this paper, we use three feature selection algorithms namely Fast Correlation Based Feature Selection (FCBF), a variation of FCBF called Fast Correlation Based Feature Selection # (FCBF#) and Fast Correlation Based Feature Selection in Pieces (FCFBiP). The three feature selections are compared, and experimental results prove that the FCFBiP is efficient compared to FCBF and FCBF#.[59] A new model for associative memory, based on a correlation matrix, is suggested. In this model information is accumulated on memory elements as products of component data. Two classes of memories are discussed: a complete correlation matrix memory (CCMM), and randomly organized incomplete correlation matrix memories (ICMM). The data recalled from the latter are stochastic variables, but the fidelity of recall is shown to have a deterministic limit if the number of memory elements grows without limits. A special case of correlation matrix memories is the auto-associative memory in which any part of the memorized information can be used as a key. The memories are selective with respect to accumulated data. The ICMM exhibits adaptive improvement under certain circumstances. It is also suggested that correlation matrix memories could be applied for the classification of data.[60] This paper describes a fast, correlation-based filter algorithm that can be applied to continuous-discrete problem. More the feature engineering data dimensionality can be reduced. Decision and model trees built from the pre-processed data and are often significantly smaller.[61]This article studies on how to select the feature for machine learning. There isa detailed description and methodologies of learning , insights on how it would impact the statistics as well trained model is part of the literature.[62] This is a very old study regarding machine learning using a higher order co-relation matrix. The document is very detail and explains each and every process of statistics, feature selection in detail. Though it is good to consider this document for the basics on feature selection and statistics, there are many advancements in the field of machine learning, and much more latest literatures should be reviewed.[63] showed considering correlation within predictors is crucial in making valid inferences using variable importance measures (VIMs) from three MLAs: random forest (RF), conditional inference forest (CIF) and Monte Carlo logic regression (MCLR). Using a case–control illustration, we showed that the RF VIMs—even permutation-based—were less able to detect association than other algorithms at effect sizes encountered in complex disease studies. This reduction occurred when ‘causal’ predictors were correlated with other predictors and was sharpest when RF tree building used the Gini index. Indeed, RF Gini VIMs are biased under correlation, dependent on predictor correlation strength/number and over-trained to random fluctuations in data when tree terminal node size was small. Permutation-based VIM distributions were less variable for correlated predictors and are unbiased, thus may be preferred when predictors are correlated. MLAs are a powerful tool for high-dimensional data analysis, but well-considered use of algorithms is necessary to draw valid conclusions.[64] this literature reviews the implementation of reinforcement learning for supervised machine learning algorithm. [65] This document highlights that the selected subset of features must be continuous, a sequential feature selection method is highlighted and used in several machine learning algorithm.

In the **eighth** phase, Index and Surveillance are explored in detail. [66] Objective here is to predict the risk of advanced colorectal neoplasia on the second surveillance colonoscopy could help tailor surveillance. Mean age and other statistical counts per gender were reviewed. The index stratifies the risk of advanced adenoma on the second surveillance colonoscopy. [67] reviews the clinical significance of the First Surveillance Colonoscopy after Endoscopic early colorectal cancer removal. The literature states there are significant cases of lesions missed to be removed during Index colonoscopy hence necessitating Surveillance .[68] This is a contrary view on the previous article, the literature details well on the methodology of the research. It is concluded that Post polypectomy colonoscopy intervals can be extended beyond 5 years in patients with nonadvanced adenomas. Our findings also support a rescreening interval of 5 to 10 years in patients with a negative index colonoscopy. Patients with an index advanced adenoma are at highest risk for recurrent advanced adenoma and should have repeat colonoscopy before a 5 year interval.(Leddin et al., 2013)the guidelines for surveillance colonoscopy is explained here. [70]This literature reviews index colonoscopy and the related risk factors for post-colonoscopy colorectal cancers.[71] the guidelines for post polypectomy colonoscopy surveillance . Limitations of the research are clearly defined in this research. Methods, statistics, graphs are explained in the literature clearly. [72] aims to evaluate the appropriateness of post-polypectomy surveillance colonoscopy on a community wide basis and identify factors associated with it. It is concluded here that in community practice, post-polypectomy surveillance colonoscopy is often performed earlier than recommended, especially in low-risk subjects. Interventions to improve adherence to guidelines and to reduce unnecessary examinations are needed.[73] This literature concludes that the prospectives data including long-term outcomes from the Japan Polyp Study, which is a multicentre randomized control trial, would be useful to establish the Asia–Pacific consensus in the near future.(Schreuders et al., 2013) aims to learn the appropriateness of surveillance colonoscopy intervals after polypectomy. The research here concludes that only a minority of surveillance colonoscopies were performed according to guideline recommendations. Deviation from the guidelines did not improve the adenoma detection rate. Interventions aimed at improving adherence to surveillance guidelines are needed. [75] This study is significant for this research to identify the utilization of surveillance colonoscopy on a community wide basis. It is concluded that there is substantial over utilization of surveillance colonoscopy among low-risk subjects and underutilization among subjects with advanced adenoma. Interventions to better align surveillance colonoscopy use with risk for advanced lesions is needed.

Conclusion:

Literature review for this proposal involved reviewing 81 different sources across 8 different phases. The phases were equally divided into technical as well as the domain. The domain review for colonoscopy was done to take into consideration factors influencing this research, there were no conflicts with regards to the factors influencing the colonoscopy, key considerations while predicting the colonoscopy numbers. Research paper/books/articles from across the world were reviewed and the understandings confirmed. Review of technical papers with regards to Machine learning models, corelation, feature engineering, Statistics and Hypothetical testing were reviewed. The range of years included one paper of 1986 which had a very depth basic introduction to machine learning, as well as the latest one of 2023, with advanced technologies like neural networks, there weren’t any conflicts observed. The details learnt in literature review will be implemented in sampling as well as in the primary research.

# Implementation:

* 1. Primary Research:

Depth interview is the primary research methodology used in this research.

* + 1. Method:

Study Setting: The study took place in the National Screening Service, Ireland premises. The meeting was conducted in Microsoft teams.

Participant Sampling: All the participants are over the age of 18. The participants selected are the subject matter experts in the field of Bowel Screening. Two of the participants are involved in the decision-making process of the programme and has visibility on the studies done in the field of colonoscopy which will be crucial in this research. One of the participants is the deputy programme manager in Bowel Screen, the second participant heads the programme evaluation unit in National Screening Service, the third participant is a research officer who has done the research on the latest census numbers and its impact on the Bowel screen programme register. The sampling applied is that of judgement and convenience sampling. The experts are available and approachable in the office premises.

Procedure: The chart explaining the steps followed in depth interview is above (chart 1). Once the participants are identified, the participants are approached via email briefing my requirement to them. An Information leaflet briefing the research, the objectives, purpose of the in-depth interview is shared with the participants. Customised questionnaire for each of the participants is created. A meeting invite for each of the participant is sent across. The meeting was done in Microsoft Teams. An email consent of each participant is obtained to record and transcribe the meeting, and to use the meeting notes in this research. The meeting with each of the participant was for 30-40 minutes. In the beginning of the interview the research topic, objectives are all explained to the participants.

In the first meeting with the Research Officer, on the 4th of August 2023, the details of the census 2022 are discussed. There were three questions asked and replied in detail for thirty minutes. In the first question an attempt to understand the 2022 census data research was asked. The participant responded in detail on the research done on this. The second question was to understand the impact of census data on the existing Bowel screening register, this is an attempt to see if the census numbers are registers are directly proportional. The participant explained in detail and redirected me to the report that she has produced which showed the variations of census for each age in the census with the register were explained. The third question was regarding the discussions done with the Bowel Screen team on the impacts of the census, the participants has answered that the discussions are yet to take place, however she has redirected me to the Program Evaluation Unit (PEU) lead on understanding the impact.

The second interview was with the PEU lead on the 8th of August 2023. First, the research topic was explained. There were six questions targeted at the participant. The first question was to understand if the age has an impact on the colonoscopy participants. The participant has replied saying that the study was done for the uptake rate to the programme. I was redirected to the programme report to see the variations for each age range. The second question was again directed to understand the age impact on the colonoscopy, if there is a hesitancy in younger age. The third question was to understand if both men and women are equally probable to undergo colonoscopy, the participant has said there is no significant difference in the number of eligible colonoscopy clients, however there are studies indicated men participate less in the overall programme. The third question was if there are any additional more features that should be considered for the colonoscopy predictions, for which the participant suggested that the deprived zone could be a factor. Further questions were to understand the past research done in the colonoscopy predictions in the past, the answer was there any study done for Bowel screen programme in Ireland, but there are studies available for Scotland which is comparable with Ireland.

The third interview was with the deputy programme manager for Bowel Screen in National Screening Service on 9th of August 2023. There were five questions to the participant. The first question was if there is a hesitancy at younger age to participate in the programme, the participant responded stating that the round reports suggest there are no impacts, with regards to the age range extension how the younger age participates is yet to be identified, in order to better manage only one age at a time is extended. The second question was to see if the probability of men and women undergoing colonoscopy the same, participant has redirected me to the DCU research on this topic, where the results suggested male participation in the programme is less compared to female. The third question was about additional features apart from the gender and age to consider for accurate colonoscopy predictions, she said that the colonoscopy numbers are based on the department of social protection provided numbers and not really the census numbers, which means the census-based prediction need not be very accurate. The fourth question was if there were any research done in the past regarding the population extension, the participant provided references to the NHS/Scotland research. The fifth question was regarding the practical challenges in predicting the colonoscopy numbers based on the census data. For this the participant replied that currently the participants are called from the department of social protection data and not the census data, this means there should be deeper analysis on how these two datasets compare with each other.

**5.1.2 Analysis and Results:**

The interview transcribes of the three interviews are analysed in detail. The questions specific to the census research were answered very elaborately by the research officer. She has recommended that I refer to her research work which states the difference with the bowel screen database and the actual census numbers. There were at least 25% variations in the younger range groups. The deputy program manager has suggested this could be because they might not be the department of social welfare database. The significance of the gender/ age while predicting colonoscopy couldn’t be confirmed, however since there are variations in the uptake rate, we will check further in the below sections using the corelation matrix and apply the models and see if there are impacts. We will also do a hypothesis testing to confirm the impacts of these features. The Program Evaluation Unit lead has directed me to the literature for the round reports for Bowel Screen for Ireland, and the studies done by Scotland to validate the impacts of the features, the round reports review didn’t suggest any impact. The deputy program manager was emphasising on the use of department of social protection data as opposed to the census data, however the census data showed that there is a substantial difference between them, and since the prediction is used for planning the colonoscopy numbers, we will use the census data which could be more accurate.

* 1. Data Extract:

There are two sets of data extracted as part of the research. One is the actual screening colonoscopy counts from the Bowel Screen database. The second is the 2022 census data which is publicly available in the CSO database. The data from the Bowel screen is available in the SQL server database, SQL is written to extract the data. No personal sensitive data will be extracted from the database. The Id’s columns are generated for the purpose of analysis, and no link to these ID’s will be available in the bowel screen database.

**SQL Used:**

A white background with black text

Description automatically generated

The second set of data is extracted from the CSO website (<https://ws.cso.ie/public/api.restful/PxStat.Data.Cube_API.ReadDataset/FY006B/XLSX/2007/en>)

* 1. Data preparation:

The steps from Data preparation through Model building and evaluation including data analysis are all done in PYTHON, most of the times making use of the standard PYTHON libraries.

Data preparation is done as an iterative process in order to achieve reliable model parameters. In the first step, upon viewing the sample records we find that the records for 2012,2013,2020 are not complete. As confirmed in the depth interviews, this is expected because 2012-2013 is the year the programme was launched, 2020 was a COVID year and hence these three years have low colonoscopy counts. These years will impact the predictions and hence removed these years from the dataset. In order to do an exploratory data analysis, data-frames aggregated at clinic id, gender, age levels are created. The procedure date is split into date, month, year fields so that to create a pivot data on year to build a machine learning model. The lower age range had NULL values; they are replaced with zero’s because the colonoscopy data is considered only from 60-69. The gender (M or F). Mean colonoscopy numbers for each gender and age combination are calculated. Below is how the dataset would look like:

A screenshot of a graph

Description automatically generated

*Fig: Sample colonoscopy count dataset aggregated on gender and age pivoted for year*

In the next series of steps, the census data is modified to fit the research. The data is again grouped by age and gender, and the mean is calculated for this data frame.

A table with numbers on it

Description automatically generated

Since, the colonoscopy data is available for only the age range 60-69, as they are the eligible age range population, and since we need the age range 55-69 for prediction of colonoscopies after the age range extension, Percentage of colonoscopies per census population for each age range are calculated and is applied on the census population between the age of 55-60.

A computer screen shot of a computer

Description automatically generated

In the next series of steps, the data preparations are done for the time series analysis.

Three different columns Quarterdate, Weekdate, monthdate are added. Period Index object in the pandas library is used.

A screenshot of a calendar

Description automatically generated

As mentioned, earlier the entire data preparation is an iterative exercise.

Data preparation is extended to train and test regression data models, day wise, quarter wise, weekly aggregates. Yearly average colonoscopy by extrapolating the census data with the actual colonoscopy is 188. Average daily colonoscopy is 188/52= 4 approx., adding this to the actual colonoscopy number so that the age range extension is considered for the regression models. The corelation matrix showed there isn’t any fields that have very higher corelation. Hence, we will apply feature engineering to create additional colonoscopy fields by using shift method.

A screenshot of a computer

Description automatically generated

The NaN are removed using BFill, cannot use remove NaN since the records are very significant, also upon inspecting the records could see that the records are aligning closely with future quarters hence the bfill method is more accurate. Similar datasets are done for monthly data and weekly data.

The third step in data preparation is fine tuning the data for better results in the model. This is an iterative process, all the steps mentioned above are repeated experimentally.

Data preparation step is then extended to suit the regression models.

* 1. Data Analysis

As mentioned above, different data frames are created to have the aggregated values based on age, gender, and both. The first two objectives of this research are realised in this section.

In the first, exploratory data analysis is done in order to understand the data better, so that further data preparation and aggregation can be done. This includes checking if there are any NULL values, checking the data structure, checking the data types. Any corrections will be done again as part of the data preparation.

Next, analysing the age wise counts:

A graph of a line graph

Description automatically generated with medium confidence

It can be seen from the above graph, amongst the eligible population of 60-69, maximum participation is between 65-69, with 69 showing highest number of colonoscopies.

Visualizing the colonoscopy counts across age group for each clinic’s.

A graph of different colored lines

Description automatically generated

*Fig 7: Variation of colonoscopy for each clinic across age*

Clinic ID 4 (anonymised for non-disclosure reasons) shows the maximum colonoscopy count over 600. Most of the clinics have the count ranging from 100-300.

Plotting the clinic colonoscopy counts in a box plot for better visualisation:

A graph of different colored columns

Description automatically generated

*Fig 8: Box chart showing colonoscopy variation for each clinic.*

The above chart shows the box plot for each clinic id, clinic id 1 shows maximum interquartile range with the upper quartile at approximately 500 and lower quartile around 75. Smallest range is shown by clinic 5 and 21, further investigation on the deprived zones as mentioned in the primary research will be required on the reasons of this variation.

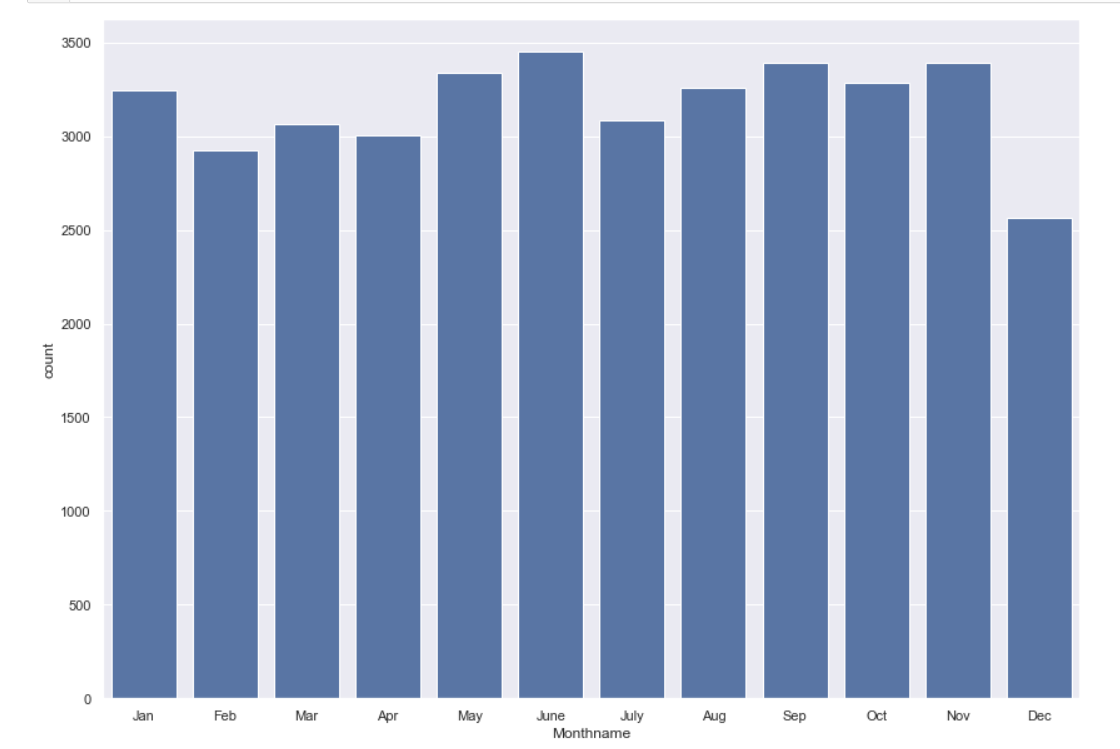
A graph showing the growth of a number of people

Description automatically generated with medium confidence

*Fig 9: Variation of colonoscopy for each gender across age*

As seen above, the male colonoscopies are more than females across the age range.

Now, visualising the colonoscopy counts for each month:



*Fig 10: Variation of colonoscopy in each month*

The highest is in June (around 3500), other months except December it ranges between 2900 to 3200), December shows less counts close to 2500, could be because of the festival season and reduced working days in December.

Checking further for each gender:

A graph of different colored bars

Description automatically generated

*Fig 11: Variation of colonoscopy in each month for each gender*

The above chart clearly shows across all the months male colonoscopy counts are substantially high compared to the females. As per the primary research the participants in the depth interview said that the screening participation of male are less compared to the female, however the colonoscopy counts show the opposite. This would need further investigation which is not in the scope of this research.

After applying the census number and projecting the colonoscopy numbers between 55-70, let us plot the age wise data for the mean colonoscopy counts:

A graph of a line

Description automatically generated with medium confidence

*Fig 12: Variation of colonoscopy across age for each gender*

The above figure shows that the colonoscopy numbers are projected to be higher for the younger age group, this is purely by using the existing colonoscopy data and calculating the younger age group using both census and colonoscopy data. More research might be required considering the psychological factors as mentioned in the depth interview, this is however not in scope of this research.

In the next steps, the feature co-relation is verified by making use of the co-relation matrix.

A screenshot of a graph

Description automatically generated

*Fig 13: corelation matrix showing impact of features on the mean colonoscopy*

The above corelation matrix shows that there is a reasonable corelation for the years 2021 and 2022 on the age, most relation exists between the previous year values.

As part of the iterative process of data preparation, to achieve better accuracy, feature engineering is applied.

A screenshot of a chart

Description automatically generated

*Fig 14: corelation matrix showing impact of features on the mean colonoscopy*

The above corelation matric is to identify the corelation between the projected column (colonoscopy) and the mean value of numbers, meanE indicates the mean estimate of eligible population for colonoscopy, mean\_l1,mean\_l2 etc are the features generated and it indicates the mean values a year back, two years back and so on. The values indicate that there is a strong co-relation between the generated mean values and the projected column. There is a small co-relation on the population. The corelation with the gender is very small. Hence, based on the data studied there is no corelation identified on the projected colonoscopy numbers with gender and age.

Hypothesis testing is the next part in our analysis, as stated in the objective, hypothesis testing is to prove the male colonoscopy for the population is less compared to the females, based on the average colonoscopy for the sample. The mean average number of colonoscopies between 60-69 is 209 as per the descriptive statistics. With 95 % Confidence Interval the values range between 158 and 259.

A screenshot of a computer code

Description automatically generated

Now, the whole population between 55-70, shows the maximum number for male is 275. Doing a T-test on one population to see if the number of colonoscopies is equal to 275 (NULL Hypothesis).

A white background with black text

Description automatically generated

P value less than 0.05 shows that the Hypothesis is rejected and the colonoscopy numbers for male is not equal to 275 and it is less than 275.

Now, considering the mean value for male between 60-69 230. Hypothesis testing to prove the colonoscopy number is greater than 230.

A screenshot of a computer code

Description automatically generated

Similar tests are done on female samples.

A graph of a graph showing the value of a number of points

Description automatically generated with medium confidence

The above plot shows that the data are normally distributed.

A screenshot of a computer code

Description automatically generated

Additional Shapiro tests done to prove normal distribution, and we can hence proceed with ANOVA.

A screenshot of a computer program

Description automatically generated

ANOVA tests show that, mean value for male and female are different. This is proved in the initial data analysis as well, median values for male and female are same as per the U-Mann Whitman test. This means there is no evidence to suggest Male colonoscopies are less compared to Female.

* 1. Model Building and Evaluation:

In this section, we will iterate through various regression models, evaluate, and identify the best models for predicting colonoscopy.

* + 1. Time Series Forecasting:

As mentioned in the data preparation, data frames for weekly, quarterly, daily colonoscopy aggregates are created for applying the time series forecasting techniques.

ARMA:

As per the data analysis section above, we don’t see any seasonality and trend component in the data. Hence, we will use the ARMA approach. We will use the PYTHON library stats model for this purpose.

Train data selected between 2014 August through 2019, test data post 2019.

A screen shot of a computer code

Description automatically generated

For ARMA (autoregressive moving average) model we will use the syntax as above, ARIMA(p,q) where p is the training data .Then we will fit the model. For timeseries prediction we will predict using the test data at each test data point and will visualise the prediction versus the test data and evaluate the model. As mentioned earlier in the data preparation section it is an iterative approach, and the data selections are done for improved measures.

A graph showing a graph of a graph

Description automatically generated with medium confidence

*Fig 15: TRAIN vs TEST vs ARMA Predictions*

RMSE: 124

Now, doing a prediction for the weekly dataset:

RMSE:1707.

The above evaluation shows very high RMSE, this could primarily be due to the fact that the data for 2020 is manufactured for the thesis, since counts where extremely low .

Now, we will build machine learning algorithms for time series:

A screenshot of a computer program

Description automatically generated

First, applying the ForecasterAutoreg, using the Random Forest Regressor after testing the linear regressor which showed very high RSE.

A graph with blue and orange lines

Description automatically generated

The error is 17285 which is again very high.

Applying the same data for weekly data showed better MSE as 2093.

Now, applying some hyper parameter tuning, using Grid search CV.

A screenshot of a computer code

Description automatically generated

A close-up of a number

Description automatically generated

Better metrics obtained to 294.

A graph showing the time line

Description automatically generated with medium confidence

Applying the machine learning regression on the daily timeseries data:

A screenshot of a computer program

Description automatically generated

A screen shot of a computer

Description automatically generated

This is the best achieved so far at 39.

* + 1. Regression Model Building:

As explained in detail in the data preparation section and the data analysis section above, iterative approach is applied and data for the most accurate model is experimented.

Referring to Fig 14 above, corelation between the new fields created as part of feature engineering and the projected columns for colonoscopy showed high.

However, when linear regression is applied to these fields showed very less accuracy.

Training data Accuracy: 0.6828569612453668

Testing data Accuracy: -9.27782899093964

Doing a Grid Cross Validation to see of the metrics improves.

Applying Random Forest regression since the linear regression showed less accurate, applying Grid CV for accurate measure selection. Kfold is the cross validation applied.

A screenshot of a computer program

Description automatically generated

A screen shot of a computer code

Description automatically generated

Test accuracy is very less compared to the Training accuracy this results in overfitting.

K Neighbours, Decision Tree regressor were also applied and they all showed very less accuracy.

Now, again data preparation is done, and daily aggregate data is used.

Applying Grid CV with Linear Regression, K Fold cross validation on this dataset now the train score is 0.65 and test score is 0.58. This means the overfitting has reduced but the training model score is also reduced. Applying Random Forest regression, the accuracy improves however again there is over-fitting.

Training Accuracy: 0.98

Test Accuracy: 0.45

Applied, Booster Algorithm using LGBMRegressor Library in python, however no improvements seen.

We will now apply neural network and see if there are any improvements after hyper tuning.

A screenshot of a computer program

Description automatically generated

A screenshot of a computer program

Description automatically generated

In the first hidden layer, we have used 9 neurons(units). Activation function used here “relu”.

Activations functions add non-linearity in the learning. Relu activation is a default activation function since it is easier to train and test. The sigmoid and tangent activation function are like Relu in the sense they both can train complex learning, however for the sake of our analysis we prefer a simpler and widely default activation function like the Relu (Rectified linear activation function).

The first hidden layer also has a parameter, input shape. This is the number of columns in the features, this is dynamically calculated from the shape of the training data (feature).

Further dense layers used in this regression deep learning are similar to the first hidden layer, except that input shape is not required for further neuron layers. The number of hidden layers is decided based on analysing which is the combination that provides the best model in with MSE/MAE scores. The details of results obtained are provided at the end of the research with a conclusion on what is the best approach for this research.

The final layer towards the output should have only one unit with no activation function.

Before we can train our model, we need to tell Keras how we want our network to learn. We do this using the compile method, with our optimization algorithm, loss function, and one or more performance metrics.

We will use mse (mean square error loss) as a loss function, this is a default loss function for regression problems. Let us view the distribution of the target variables.

A graph showing the distribution of the colonoscopy

Description automatically generated

Since the default loss function is MSE, and the curve is nearly gaussian we will go ahead with the MSE loss function.

Optimizers are used to change the attributes of the neural network to reduce loss. The optimizer decides how much the output varies with change in input, this is called gradient. Studies have shown that ADAM (Adaptive Moment Estimation) is meant to reduce the training cost compared to another optimizer. The model can be trained faster compared to other optimiser, also the data is not 100% homogenous, ADAM does work well in sparse data as well. This is the reason ADAM is a default optimizer, there is no compelling reasons for selecting any other optimizer than ADAM. The results showed that each passing echoes improved the performance of the model when the optimiser Adam is used.

* 1. Prediction

# Results:

The in-depth interview done for the primary research was done with the subject matter experts. The variations in the census data and the data in the programme register was established. The factors affecting colonoscopies could not be established in the in-depth interviews, there were references given to the existing studies on the programme intake which proves male participation is less in the programme, however the age has no major impact. This needs to establish further in the data analysis and hypothesis testing.

Data analysis was done on the secondary data extracted based on the primary research outcomes. The colonoscopy counts had variations in the age, and age range 65-70 showed the highest total colonoscopy counts for all years. On analysing the colonoscopies there is a huge variation in the numbers in different colonoscopies. Monthly variations were analysed and there is no significance variations across the months. When the estimated colonoscopies based on the percentage of population undergone colonoscopy was analysed, younger age group showed higher colonoscopy numbers. Co-relation between different variables to assist in predictive models were analysed no corelation were observed on the colonoscopy counts with the exogeneous variables like the age and gender. However new features to represent previous year colonoscopy showed closer co-relation. Hypothesis testing was done to validate the primary research outcomes, at 95% Confidence Interval, male colonoscopy number for the entire eligible population is expected to be between 270 (maximum value in the sample) and 230 (average in the sample). ANOVA tests were done and it suggested that at 95% confidence interval male colonoscopies will be higher than the women.

Machine Learning models were developed and evaluated for predicting future colonoscopies.

Time Series Analysis:

Root Mean Square Error for each model.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Date Frequency | ARMA | Linear Regression | Random Forest Regressor | GRID CV (Random Forest Regressor) |
| Quarterly | 124 | 17285 | 294 | 294 |
| Weekly | 1707 | 293 | 39 | 329 |

The above table summarizes the performances of each time series models, with different datasets. The root mean square is the measure used for comparison. We see that the performance over all is less because the root mean square error is high. Amongst different set of experimentation done on the models, Weekly data with Random Forest Regressor showed the highest performance with minimum root mean square error compared to other models tested. The Quarterly data performed very poorly because the data is considered from 2014-2022(without 2020) which is very less. This is because as per the primary research depth interview 2020 colonoscopies are expected to be very low because of COVID.

Regression Models:

Quarterly data aggregate (30 % Test)

|  |  |  |  |
| --- | --- | --- | --- |
|  | Linear Regression | GRID CV – Linear Regression | GRID CV – Random Forest Regressor |
| Train Accuracy | 0.67 | 0.69 | 0.85 |
| Test Accuracy | 0.44 | 0.12 | 0.085 |

Weekly data aggregate (22 % Test)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Linear Regression | GRID CV – Linear Regression | GRID CV – Random Forest Regressor | LGBM |
| Train Accuracy | 0.64 | 0.64 | 0.89 | 0.7472 |
| Test Accuracy | 0.23 | 0.63 | 0.61 | 0.7474 |

Daily Data Aggregate (27 % Test)

|  |  |  |
| --- | --- | --- |
|  | GRID CV – Random Forest Regressor | LGBM |
| Train Accuracy | 0.437 | 0.3995 |
| Test Accuracy | 0.14 | 0.0153 |

The above three table compares different models tested with different level of data aggregates.

The test data is iterated and tested, and the above tables have the highest accuracy obtained after iterations. As seen above quarterly aggregated data showed poor performance because the train data was very less, and hence the test v/s train was not of comparable size.

The daily data aggregate also showed less performance, because certain days had very less counts, and the train and test data selected couldn’t really match and comparable.

The weekly aggregate was of the reasonable blend, and shower better performance compared to others. GRID CV Linear Regression showed high Training accuracy (89%), test accuracy (61%). This is a case of overfitting. Light Gradient Booster Algorithm was applied along with hyper-tuning techniques, and this resolved the over-fitting issue to obtain train accuracy as 0.7472 and Train Accuracy of 0.7474.

In order to further seek best performance, Neural network was applied on the daily data aggregate, appropriate hyperparameter techniques, usage of different layers of network, usage of correct optimizer, loss function resulted in improvement in performance during the epoch. The loss value reduced from 400 to 28.

# Conclusion: