A Project Report

on

Healthcare Analytics Using Machine Learning

Submitted in partial fulfilment of the

requirement for the award of the degree of

Bachelor of Science in Computer Science

GALGOTIAS UNIVERSITY

(Established under Galgotias University Uttar Pradesh Act No. 14 of 2011)

UnderTheSupervision of
Name of Supervisor: Dr. Sathiya Priya
Designation: Assistant Professor

Submitted By

Sarthak Dhama(20SCSE1100039)
Kunal Bhati(20SCSE1100032)

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING /
DEPARTMENT OF COMPUTER APPLICATION
GALGOTIAS UNIVERSITY, GREATER NOIDA
INDIA MAY,2023

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA

CANDIDATE'S DECLARATION

We hereby certify that the work which is being presented in the thesis/project/dissertation, entitled "**Healthcare Analytics Using Machine Learning**" in partial fulfilment of the requirements for the award of the Bachelor of Science in Computer Science submitted in the School of Computing Science and Engineering of Galgotias University, Greater Noida, is an original work carried out during the period of March,2023 to May,2023 , under the supervision of Dr. Sathiya priya, Assistant Professor , Department of Computer Science and Engineering/Computer Application and Information and Science, of School of Computing Science and Engineering , Galgotias University, Greater Noida

The matter presented in the project has not been submitted by us for the award of any other degree of this or any other places.

Sarthak Dhama(20SCSE1100039)
Kunal Bhati(20SCSE1100032)

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

Dr. Sathiya Priya

Assistant Professor

CERTIFICATE

The Final Thesis/Project/ Dissertation Viva-Voce examination of Sarthak Dhama, Kunal Bhati has been held on________________________________and his/her work is recommended for the award of Bachelor of Science in Computer Science.

Signature of Examiner(s) **Signature of Supervisor(s)**

Signature of Project Coordinator **Signature of Dean**

Date:

Place: Greater Noida

Table of Contents

List of Figures

HEALTHCARE ANALYTICS

ABSTRACT

The amount of data kept in the many public and private hospitals, clinics, and other venues of medical practise has greatly increased as medical research has advanced. This vast repository of data needs to be managed properly so that we may apply an appropriate analytic method to produce insightful and conclusive findings. Machine learning algorithms are used to analyze both structured and unstructured data in order to effectively handle such a massive volume of data. A crucial part of machine learning algorithms, predictive analytics aids users in improving and supervising their decisions. A method for efficiently sorting the voluminous and quickly growing data in the field of medical research is visual analytics. It enables us to deal with the disparate material in an orderly fashion that the human brain can easily visualize. This would then produce fresh, creative outcomes with potential. These analytics starts organized concepts in people's minds in addition to providing structured data. The technique of visual analytics will supply sorted pertinent facts relating to it as professionals examine certain anomalous occurrences. As a result, the expense of preserving massive amounts of data would be reduced. The goal of this project is to accurately predict the length of stay for each patient so that the hospitals can optimize resources and function better.

CHAPTER-1
INTRODUCTION TO HEALTHCARE ANALYTICS

Healthcare organizations are under increasing pressure to improve patient care outcomes and achieve better care. While this situation represents a challenge, it also offers organizations an opportunity to dramatically improve the quality of care by leveraging more value and insights from their data. Health care analytics refers to the analysis of data using quantitative and qualitative techniques to explore trends and patterns in the acquired data. While healthcare management uses various metrics for performance, a patient's length of stay is an important one. Being able to predict the length of stay (LOS) allows hospitals to optimize their treatment plans to reduce LOS, to reduce infection rates among patients, staff, and visitors

This Work provides the foundation for development of technology framework that makes easy to find all the relevant information regarding treatment and diseases. The tool that is built with the techniques such as Natural Language Processing (NLP) and Machine Learning (ML) has capability to find all relevant short text information regarding diseases and treatments. This work presents various Machine Learning (ML) and information for classifying short texts and relation between diseases and treatments.According to ML technique the information are shown in short texts when identifying relations between two entities such as diseases and treatment. Thus there is improvement in solutions when using a pipeline of two tasks (Hierarchical way of approaching). It is better to identify and remove the sentence that does not contain information relevant to disease or treatments. The remaining sentences can be classified according to the interest. It will be very complex to identify the exact solution if everything is done in one step by classifying sentences based on interest and also including the sentences that do not provide relevant information.Relation Extraction is a long standing research topic in Natural Language Processing. Medical information are stored in textual format among the biological data stored in Medline. Manually extracting useful information from large volume of database is a tedious work. Moreover HTML page displaying biological information contains medical information and typically unrelated materials such as navigation menus, forms, user comments, advertisement, feedback etc. The proposed work of this project extracts the useful disease related information with increased precision by using weighted bag of word representation [1] with a accuracy of 79% to 82%. The proposed approach supports in clinical decision making by providing physician with best available evidence of medical information. The challenges of information extraction. Medical subheadings and subject heading may be used to infer relationship among medical concepts. The classification algorithm used in the proposed work exhibits effectiveness, efficiency, Online learning ability.

frequent use of electronic health records and information increase the need for text mining in order to improve the quality of result for the user query. This can result in two area of real time application[7] such as Text search engine targeted with Scientific document and Text Search engine targeted with technical document. In this project we choose text mining targeted with scientific document related to Medical treatment. Medline is chosen in this project to get biomedical information because it provides answers related to patient treatment and it's the database which is most widely used by the clinicians and research scholars in medical field. More importantly it is frequently updated and the contents are proved to be accurate compared to other medical websites providing information related to human disease, health, medicines, treatment etc. With the growing number of medical thesis, research papers, research articles, researchers are faced with the difficulty of reading a lot of research papers to gain knowledge in their field of interest. Search engines like Pub Med [8] reduces this constraint by retrieving the relevant document related to the user query. Though the relevant document is retrieved, the web page displaying it may contain many non informative contents like advertisement, scroll bars, menus, citations, quick links, announcements, special credits, related searches, similar posts searched etc. This may be quite frustrating to the user when the user is in need of the information alone. In this project all the unrelated contents like advertisement etc mentioned in the above paragraph are removed and text mining is performed on the extracted document from which information or sentences related to user specified disease is extracted. From the extracted file symptoms, causes, treatment of the particular disease is filtered and displayed to the user. Thus the user gets the required information alone which saves his time and improves the quality of the result. This text mined document can be used in medical health care domain where a doctor can analyse various kinds of treatment that can be given to patient with particular medical disorder. The doctor can update the knowledge related to particular disease or its treatment methodology or the details of medicine that are in research for a particular disease. The doctor can gain idea about particular medicine that are effective for some patient but causes side effect to patient with some additional medical disorder. The patient can also use this extracted document to get clear understanding about a particular disease its symptoms, side effects, its medicines, its treatment methodologies. Understanding the effect of a given intervention on the patient's health outcome is one of the key elements in providing optimal patient care. In the proposed approach a combination of structural natural language processing with machine learning method address the general and domain specific

CHAPTER-2

Literature Survey

[1] In Jason D. M. Rennie, Lawrence Shih, Jaime Teevan, David R. Karger ,"Tackling The POOR Assumption Of Naïve Bayes Text Classifier" there were mentioned classification of text by using naïve bayes text classifier but use of navie bayes text classifier does not give precision 100% for output. Sometimes prediction of classifier may not be correct. [2]In T.Mouratis, S.Kotsiantis, "Increasing The Accuracy Of Discriminative Of Multinominal Bayesian Classifier In Text Classification", paper author introduced use of classifier that increased precision of output but problem in that work was at the time of classification it doesn't identify the verbs,nouns,adjectives properly so some time it nay give wrong value. [3]In B.Rosario And M.A. Hearst,"Semantic Relation In Bioscience Text" where Hidden Markov models are used for entity recognition. This includes mapping biomedical information into structural representation. It involves converting natural language text into structural format. Their work uses machine learning for information extraction. The extraction of medical abstract is obtained through text classification. Semantic lexicons of words labeled with semantic classes so associations can be drawn between words which helps in extracting the necessary sentences related to the query. In this research paper the author used sentence co-occurrence and navie bayes algorithm to extract semantic relation like Gene-Protein from Medline abstract, the precision and recall of the result obtained are shown in the graph as their experimental results but due to use of only one navie bayes algorithm it do not get good precision of output, it doesn't used bag of words to find adjective, verbs while doing classification. [4]In M.Craven, "Learning To Extract Relations From Medline" In their work the individual sentences are considered as instances that are to be processed by the navie bayes classifier. Here each instance is considered as positive training set. Alternative relation extraction are made through relational learning. Extraction of words from medline abstract has been done by using navie bayes,CNB algoritham and it also used bag of words during classification but not used natural language processing due to this performance of output degrades. [5]In Oana Frunza.et.al, "A Machine Learning Approach For Identifying Disease-Treatment Relations In Short Texts" It involves automatic extraction of relation between medical concepts. A dictionary of medical terms is used for sentence classification. The sentences are automatically parsed using semantic parser. After applying semantic extraction a set of extraction, alteration, validation rules are applied to distinguish the actual semantic relation to be extracted but problem is that due to used of only one algorithm of machine learning navie bayes may not get good precision of output.

combination of the six classification algorithm and three representation techniques. The results are shown in bar chart form. As the result of the experiment it is concluded that bag-of-representation when combined with any of six classification algorithm produces better results but it does not give disease diagnosis as well information about particular disease by parsing statement.

CHAPTER-3

Proposed System

The two tasks used in this paper are the basis for the development of information technology framework. This framework helps to identify the medical related information from abstracts. The first task deals with extraction all information regarding diseases and treatments while the task deals with extraction of related information existing between disease and treatments. The framework developed with these tasks are used by healthcare providers, people who needs to take care of their health related problems and companies that build systematic views. The future product can be provided with browser plug-in and desktop application so that it helps the user to get all information related to diseases and treatments and also the relation between those entities. It is also be useful to know more about latest discoveries related to medicine. The product can be developed and sold by companies that do research in medical care domain, Natural Language Processing (NLP), and Machine Learning (ML), and companies that develop tools like Microsoft Health Vault and Google Health. This product is valuable in e-commerce fields by showing the statistics that the information provided here are accurate and also provide all the recent discoveries related to health care. To make a product more popular it should be trust worthy so that people can buy it. It is the key factor foe any company to make product successful. When coming to health care products it should be more trust worthy since it is dealing with health related problems. Companies that wish to sell health care framework need to develop tools that automatically extract the wealth of research. For example the information provided for diseases or treatments needs to be based on recent discoveries on health care field so that people can trust. The product quality also should be taken care so that it provides dynamic
is used to analyze data and identify the pattern, such patterns can be used to make prediction which is an effective step in decision making. It can be applied to identify pattern in health care domain to find pattern observed in the symptoms of particular disease. Now the resulted file containing information related to Symptoms, Causes, and Treatment from the uploaded html file is tested for its quality. The quality of the resulted file is obtained by calculating its Precision, Recall, F-measure. The obtained result is assumed to have quality if these values are within the range of 0.0 to 1.0. The formulas for calculating these quality measures are Precision= (relevant retrieved)document/ Retrieved document. Recall =(relevant retrieved) document/ Relevant document. F-measure= mean of Precision and Recall.

CHAPTER 4
Hypothesis Generation

Understanding the problem in detail by assuming different factors that impact the outcomes of Length of Stay before any data exploration or analysis. Here the variables can be divided into two levels:

Patient-Level and Hospital-Level.

Patient-Level:

• Type of Admission – Patients can be admitted in three levels Urgent, Emergency, and Trauma. Patients admitted to urgent care are likely to stay fewer days. Whereas Trauma patients usually stay longer because they must be monitored until they are qualified to be discharged. • Severity of Illness – Severity can be classified as Minor, Moderate, and Extreme. A patient recorded as minor will stay fewer days than a patient recorded as extreme.

• Visitors with Patient – Patients with more visitors are like to stay longer in the hospital.

• Age – Infants and older Patients usually take a longer time to recover so they stay longer than younger Patients.

• Admission Deposit – Patients who are likely to deposit a high amount of money at the time of admission might have severe conditions and stay longer.

Hospital-Level:

• Ward Type – Patients allocated in ICU might stay longer than the general ward as their condition is more severe.

• Department – Patients under surgery are likely to stay longer than gynecology as their recovery time is long

CHAPTER - 5

Data Exploration

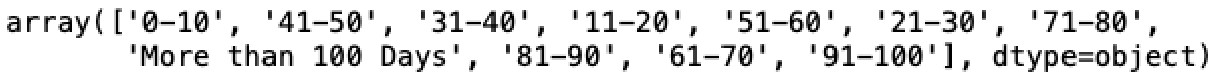
Overview of Data

The train data consist of 318438 observations for which patient length of stay can be predicted from 17 variables. The description of all variables is shown in Table 4.1.

Table 5.1 Dataset Overview

Variables	Description
case_id	Case_ID registered in Hospital
Hospital_code	Unique code for the Hospital
Hospital_type_code	Unique code for the type of Hospital
City_Code_Hospital	City Code of the Hospital
Hospital_region_code	Region Code of the Hospital
Available Extra Rooms in Hospital	Number of Extra rooms available in the Hospital
Department	Department overlooking the case
Ward_Type	Code for the Ward type
Ward_Facility_Code	Code for the Ward Facility
Bed Grade	Condition of Bed in the Ward
patientid	Unique Patient Id
City_Code_Patient	City Code for the patient
Type of Admission	Admission Type registered by the Hospital
Severity of Illness	Severity of the illness recorded at the time of admission
Visitors with Patient	Number of Visitors with the patient
Age	Age of the patient
Admission_Deposit	Deposit at the time of Admission
Stay	Patient Length of Stay

In this data, the target variable “stay” is divided into 11 different classes ranging from 0 days to more than 100 days. **Figure 5.1** shows different levels of the “Stay” variable.



```
array(['0-10', '41-50', '31-40', '11-20', '51-60', '21-30', '71-80',
       'More than 100 Days', '81-90', '61-70', '91-100'], dtype=object)
```

Data Cleaning and Preparation

In this data set, variables "City_code_patient" and "Bed Grade" have missing values. These missing values must be treated before feeding to the algorithm as they distort the model performance. So, the missing values are replaced using the "mode" imputation technique.

Since most of the variables in the dataset have ordinal data, we transformed them into levels by using a label encoder to perform further analysis on the data. Table 4.2 shows the number of distinct observations of ordinal data in the dataset.

Table 5.2 Distinct Observations of Ordinal Data

Variables	Number of distinct observations
Hospital_type_code	7
Hospital_region_code	3
Department	5
Ward_Type	6
Ward_Facility_Code	6
Type of Admission	3
Severity of Illness	3
Age	10
Stay	11

Feature Engineering

Once the data is cleaned and prepared, we grouped patientid and case_id to extract the new column “count_id_patient”. This variable contains the count of multiple admits of a patient under different case_id. Further two more columns “Hospital_region_code” and “ward_facility_code” were grouped to patientid and case_id. These two new variables “count_id_patient_hospitalCode” and “count_id_patient_wardfacilityCode” contain the count of multiple admissions in a hospital region and the count of multiple wards allocated to a patient. [Appendix - 10.1 Feature Engineering]

Before getting into analysis, the train data must be split into two parts, the first part with all the feature variables and the second part with a target variable (“Stay”). Then preprocessed into train and validation sets. So, here we are portioning the train set with 80% and validation set with 20% of the data for Naïve Bayes and XGBoost models.

FUTURE SCOPE OF MACHINE LEARNING IN HEALTHCARE ANALYTICS

1. Improve Medical Research and Clinical Trial

While conducting a clinical trial or any medical research, it takes years to get a conclusion. It demands time and money, and the chances of getting an accurate result are unsure. However, thanks to machine learning, it can be possible quicker and cleverer.

Machine Learning helps with predictive analysis for the clinical trials, based on factors like the history of customers like doctor visits and medical data. Moreover, by applying natural language processing tools, medical researchers could get worthy insights without the urge to read them all.

2. Quick Detect for Diagnosis and Disease

The most beneficial part of machine learning is that it identifies the diseases and diagnosis the illness at an early stage. It detects with more accuracy than humans with a better pace. The technology helps to predict disease in more advanced ways than it used to take. For instance, a deep learning-based prediction model can forecast breast cancer development years in advance up to five years.

The purpose is a commercially viable approach to diagnose and provide treatment in a clinical environment by automating the process as soon as possible.

3. Improve Predictive Analytics

The compliance of data science, predictive analytics, and machine learning offer possibilities to develop healthcare methods, modify clinical decision maintenance tools, and help promote patient results. The usage of machine learning in healthcare is to leverage health informatics to predict health consequences through predictive analytics, commencing to more definitive diagnosis and practice and promoting physician insights for personalized and following methods.

Machine learning also contributes more value from **predictive analytics** by interpreting data for decision-makers to unfold method ways and enhance overall healthcare business processes.

4. Track the Health records

Keeping track of records of treatments, patient visits, doctor's cases is not an easy task. Maintaining and updating health records is time-consuming and quite expensive. But because of machine learning technology, it can perform with more efficiency. It played a vital role in promoting the data entry process. Nevertheless, most of the procedure still takes a lot of time to finish because they require it to be produced manually.

Multiple institutions are now developing the future development of intelligent health records that will include technology tools from the basics up to help in the clinical treatment recommendations and diagnosis.

5. Medical Imaging Analysis

One of the significant elements of **machine learning in the healthcare industry** is medical imaging diagnosis. Machine learning and deep learning are liable for the breakthrough technology described in Computer Vision. It has gained acceptance in the initiative developed by the medical sector- which operates on image diagnostic tools for image analysis.

As machine learning becomes more convenient and as they progress in their critical role, demand to see more data specialists from various medical imagery grows as a part of this AI-driven diagnostic process.

6. Better Data crowdsourcing

Recently, the medical industry has found crowdsourcing, and now researchers and practitioners utilize the technique to obtain extensive amounts of data people upload based on their permission.

Such vital health data comes with numerous implications on how medicine will work in the future. The tool is a map based on crowdsourced information to reveal, collect, and share diabetes and insulin data in real-time.

With the advances that are occurring in the Internet of Things field- the healthcare industry might still be on its way to finding new methods of applying the data and boost the overall performance of diagnostics.

Future of ML in Healthcare

The development in machine learning increases the performance and precision of disease apprehension to decrease the stress on doctors. **Big Data Analytics, Data Science,** and Machine Learning will remodel the future of the healthcare industry.

However, Machine learning still needs improvements, and various factors demand it to be improved.

Benefits of Machine Learning in Healthcare

Using machine learning in healthcare operations can be extremely beneficial to the company. Machine learning was made to deal with large data sets, and patient files are exactly that – many data points that need thorough analysis and organizing.

Moreover, while a healthcare professional and a machine learning algorithm will most likely achieve the same conclusion based on the same data set, using machine learning will get the results much faster, allowing to start the treatment earlier.

Another point for using machine learning techniques in healthcare is eliminating human involvement to some degree, which reduces the possibility of human error. This especially concerns process automation tasks, as tedious routine work is where humans err the most.

1. Clinical Decision Support Systems

Clinical decision support tools help analyze large volumes of data to identify a disease, decide on the next treatment stage, determine any potential problems, and overall improve patient care efficiency. CDSS is a powerful tool that helps the physician do their job efficiently and quickly, and it reduces the chances of getting the wrong diagnosis or prescribing ineffective treatment.This use of machine learning in medicine (healthcare) has been around for a while but has become more widespread in recent years. The reason behind it is the wider acceptance of the electronic health record system (EHR) and digitalization of various data points, including medical images.

2. Smart Recordkeeping

Making sure that all the patient records are updated regularly is challenging, as data entry is a monotonous task. However, it is also crucial for effective decision-making and better patient care.

One of the uses of machine learning in healthcare is using optical character recognition (OCR) technology on physicians' handwriting, making the data entry fast and seamless. This data can then be analyzed by other machine learning tools to improve decision-making and patient care.

3. Machine Learning in Medical Imaging

For the longest time, medical images, like X-rays, have been analog. This has limited the use of technology for anomaly identification, case grouping, and overall disease research. Fortunately, the digitalization of the process has led to more opportunities with these types of data analysis, including with the help of machine learning. And, according to a recent meta-analysis, machine learning algorithms do the job as well as (and, in some cases, even better) human specialists, with 87.0% sensitivity and 92.5% specificity for the deep learning algorithms and 86.4% sensitivity and 90.5% specificity for human physicians.

4. Personalized Medicine

What makes medicine such a complex and resource-heavy field is that every case has its specifics. People often have a sleuth of conditions that require simultaneous treatment. So, complex decisions must be made to construct an effective treatment plan, accounting for drug interactions and minimizing potential side effects.

How to use machine learning in healthcare to solve this problem? Well, IBM has figured that out with their Watson Oncology system that uses the patient history to produce multiple potential treatment options.

5. Predictive Approach to Treatment

When it comes to most dangerous diseases, identifying them in the early stages can raise the chances of successful treatment significantly. This also helps to identify the possibility of any potential worsening of the patient's state before it happens.

One of the cases for the importance of machine learning in healthcare is that it can be used to successfully predict some of the most dangerous diseases in at-risk patients. This includes the identification of signs of diabetes (using a Naïve Bayes algorithm), liver and kidney diseases, and oncology.

12. Infectious Disease Outbreak Prediction

COVID-19 pandemic has shown us how unprepared we were to an infectious disease outbreak of this size. It is worth mentioning that experts in the area have warned the government about the possibility of such an event for years.

Now, we have tools based on machine learning that can help to detect the signs of an epidemic early on. The algorithms analyze the satellite data, news, and social media reports, even video sources to predict whether the disease has the potential to grow out of control.

Privacy and Data Security

HIPAA and other privacy regulations ensure the security of the patient's information. Everybody should have a right to keep information about their health private. Nevertheless, a lot of healthcare data leaks are happening every day that result in up to $16 million penalties for healthcare providers. However, data is the blood of the machine learning organism. How can these points effectively coexist?

This challenge is difficult to overcome. In most cases, machine learning doesn't require a full spectrum of information on the patient (like name, email, phone number, and insurance policy number); thus, it can be effectively anonymized so that the person's identity cannot be revealed, while the precision of the ML-algorithm won't be discounted. For others, special data security approaches have to be implemented to ensure patient anonymity. If you want to learn more about security in healthcare software development – check out our special article on this: https://nix-united.com/blog/how-to-develop-a-hipaa-compliant-software/

Machine learning can be effectively used to help the elderly and people with psychological issues make decisions to improve their health. This concerns taking the right medications, creating healthy habits, and referring to the specialist whenever needed.

However, the ethical issue behind this is that people will potentially give up their autonomy and act as they are told. It limits their potential choice range to certain recommended options. So, a clear balance between the instructions from the algorithm and freedom of personal choice should be provided.

CHAPTER - 6

Modelling Strategy

Model 1 - Naïve Bayes

Naïve Bayes is a classification technique that works on the principle of Bayes theorem with an assumption of independence among the variables. Here the goal is to predict Length of Stay i.e., "Stay" column (Target Variable) and it is classified into 11 levels. We must find the probability of each patient's length of stay using feature variables, which contain the patient's condition and hospital-level information. These feature variables are ordinal and naïve Bayes is a perfect multilevel classifier.

In Bayes theorem, given a Hypothesis H and Evidence E, it states that the relation between the probability of Hypothesis P(H) before getting Evidence and probability of hypothesis after getting Evidence P(H|E)

$$\mathbf{P\ (H \mid E) = \left[\frac{P\ (E \mid H)}{P(E)} \right] P(H)}$$

When we apply Bayes Theorem to our data it represents as follows.

- P(H) is the prior probability of a patient's length of stay (LOS).
- P(E) is the probability of a feature variable.
- P(E|H) is the probability of a patient's LOS given that the features are true.
- P(H|E) is the probability of the features given that patient's LOS is true.

Model is trained using Gaussian Naïve Bayes classifier, partitioned train data is fed to the model in array format then the trained model is validated using validation data. This model gives an accuracy score of **34.55%** after validating. [Appendix – 10.2.1 Naïve Bayes]

Model 2 – XGBoost

Boosting is a sequential technique that works on the principle of an ensemble. At any instant T, the model outcomes are weighed based on the outcomes of the previous instant (T -1). It combines the set of weak learners and improves prediction accuracy. Tree ensemble is a set of classification and regression trees. Trees are grown one after another, and they try to reduce the misclassification rate. The final prediction score of the model is calculated by summing up each and individual score.

Before feeding train data to the XGB Classifier model, booster parameters must be tuned. Tunning the model can prevent overfitting and can yield higher accuracy. In this

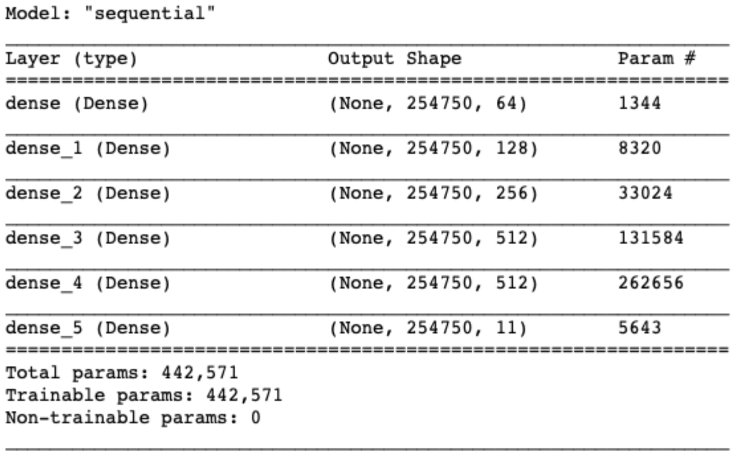
XGBoost model, we have used the following parameters for tunning,

- **learning_rate = 0.1** - step size shrinkage used to prevent overfitting. After each boosting step, we can directly get the weights of new features, and eta shrinks the feature weights to make the boosting process more conservative.
- **max_depth = 4** – Maximum depth of the tree. This value describes the complexity of the model. Increasing its value results in overfitting.
- **n_estimators = 800** – Number of gradient boosting trees or rounds. Each new tree attempts to model and correct for the errors made by the sequence of previous trees. Increasing the number of trees can yield higher accuracy but the model reaches a point of diminishing returns quickly.
- **objective = 'multi:softmax'** – this parameter sets XGBoost to do multiclass classification using the softmax objective because the target variable has 11 Levels.
- **reg_alpha = 0.5 -** L1 regularization term on weights. Increasing this value will make the model more conservative.
- **reg_lambda = 1.5** - L2 regularization term on weights and is smoother than L1 regularization. Increasing this value will model more conservative.
- **max_depth = 4** – Maximum depth of the tree. This value describes the complexity of the model. Increasing its value results in overfitting.
- **min_child_weight = 2 -** Minimum sum of instance weight needed in a child.

Once the model was trained and validated, it yields an accuracy score of **43.04%**. When compared to the Naïve Bayes model that's an 8.5% improvement. [Appendix – 10.2.2 XGBoost]

Model 3 – Neural Networks

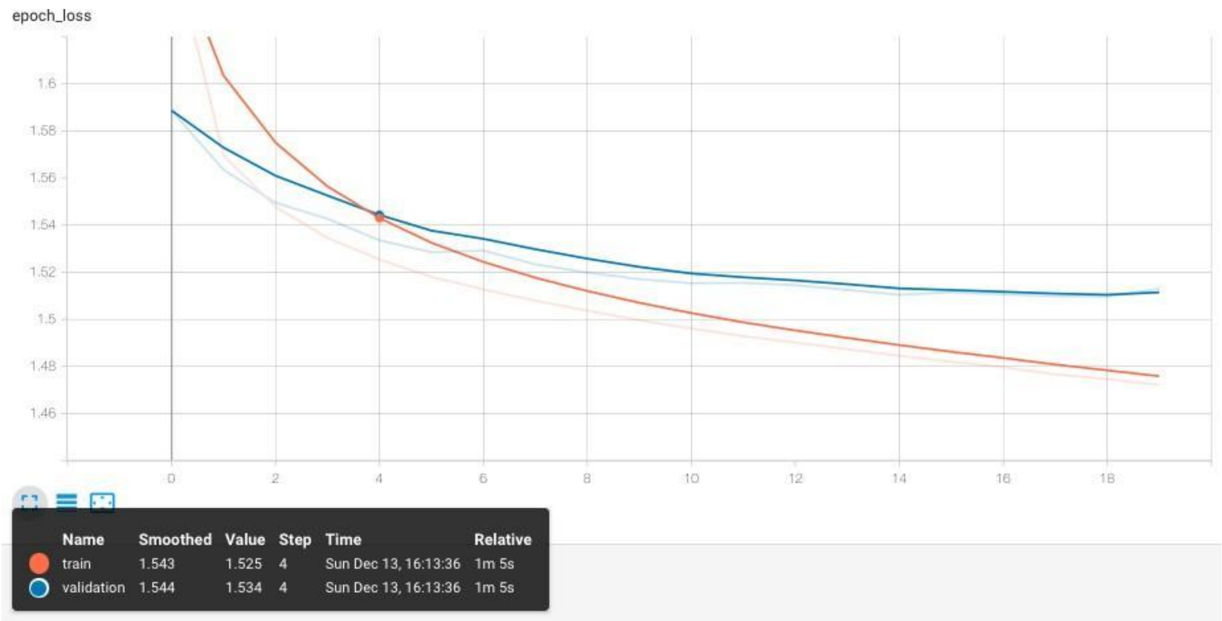
Neural Networks are built of simple elements called neurons, which take in a real value, multiply it by weight, and run it through a non-linear activation function. The process records one at a time and learns by comparing their classification of the record with the known actual classification of the record. The errors from the initial classification of the first record are fed back into the network and used to modify the network's algorithm for further iterations.



```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #   
=================================================================
dense (Dense)                (None, 254750, 64)        1344      
_________________________________________________________________
dense_1 (Dense)              (None, 254750, 128)       8320      
_________________________________________________________________
dense_2 (Dense)              (None, 254750, 256)       33024     
_________________________________________________________________
dense_3 (Dense)              (None, 254750, 512)       131584    
_________________________________________________________________
dense_4 (Dense)              (None, 254750, 512)       262656    
_________________________________________________________________
dense_5 (Dense)              (None, 254750, 11)        5643      
=================================================================
Total params: 442,571
Trainable params: 442,571
Non-trainable params: 0
_________________________________________________________________
```

In this neural network model, there are **six** dense layers as shown in Figure 5.1, the final layer is an output layer with an activation function “**SoftMax**”. SoftMax is used here because each patient must be classified in one of the 11 levels in the Stay variable. In this model, increasing the number of neurons from each layer to the other layer, will increase the hypothetical space of the model and try to learn more patterns from the data. There are a total of **442,571** trainable parameters, this can be observed in Figure 5.1. Every layer is activated using “**relu**” activation function because it overcomes the vanishing gradient problem, allowing models to learn faster and perform better.

Before training the model, data were scaled, converted into a sparse matrix, and portioned into 80% as a train set and 20% as a test set. This neural network model was compiled using "**categorical_crossentropy"** as a function of loss because the target variable is categorical and "S**GD"** as an optimizer argument. Initially, the model was trained using portioned train data with 20 epochs and validation set argument set at 20%. In Figure 5.2 showing TensorBoard, we can observe that the model is overfitting from the 4th epoch. So, the model is retrained by setting epochs to 4.



	Name	Smoothed	Value	Step	Time	Relative
●	train	1.543	1.525	4	Sun Dec 13, 16:13:36	1m 5s
○	validation	1.544	1.534	4	Sun Dec 13, 16:13:36	1m 5s

Figure 6.1 – Epoch Loss Graph

Finally, evaluating the model with a test set yields an accuracy score of **42.05%**. Neural Networks supposedly performs better than any other models. But because of the smaller dataset, it was not able to learn more accurately than the XGBoost model. [Appendix – 10.2.3 Neural Network]

CHAPTER - 7

Prediction and Results

In the Naïve Bayes model, patients are more likely to be misclassified. This model is biased towards the duration of 21-30 days, it has classified 72,206 patients for this level (can be observed in Table 6.1).

Table 7.1 – Number of observations classified into different levels of Length of Stay from all models

Length of Stay	Predicted Observations from Naïve Bayes	Predicted Observations from XGBoost	Predicted Observations from Neural Network
0-10 Days	2598	4373	4517
11-20 Days	26827	39337	35982
21-30 Days	**72206**	58261	61911
31-40 Days	15639	12100	8678
41-50 Days	469	61	26
51-60 Days	13651	19217	21709
61-70 Days	92	16	1
71-80 Days	955	302	248
81-90 Days	296	1099	1165
91-100 Days	2	78	21
More than 100 Days	4322	2213	2799

Whereas the other two models XGBoost and Neural Networks are predicting mostly similar Length of Stay for the patient, we can see this similarity for the first five cases in Table 6.2. In Table 6.1, we can see that the observations classified by both these models are marginally similar.

Table 7.2 – Predicted Length of Stay for first five cases from different models

case_id	Length of Stay predicted from Naïve Bayes	Length of Stay predicted from XGBoost	Length of Stay predicted from Neural Networks
318439	21-30	0-10	0-10
318440	51-60	51-60	51-60
318441	21-30	21-30	21-30
318442	21-30	21-30	21-30
318443	31-40	51-60	51-60

Examining these predictions, many of the patients are staying in the hospital for 21-30 days and very few people are staying for 61-70 days. As far as the distribution of Length of Stay is concerned, 13% of the patients are discharged from the hospital within 20 days and 1% of the overall patients are staying in the hospital for more than 60 days.

CHAPTER - 8
Task and Data Sets

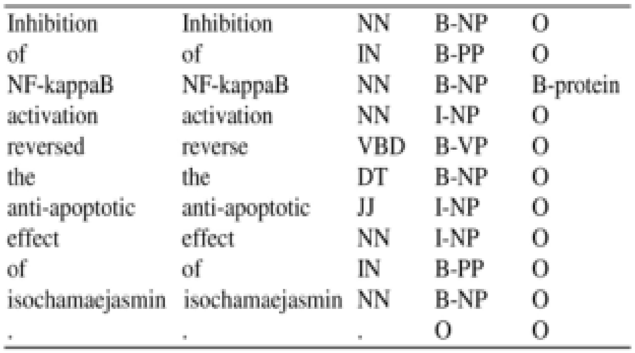
The two tasks that are undertaken in this paper provide the basis for the design of an information technology framework that is capable to identify and disseminate healthcare information. The first task identifies and extracts informative sentences on diseases and treatments topics, while the second one performs a finer grained classification of these sentences according to the semantic relations that exists between diseases and treatments.

8.1 - Bag Of Words

The bag-of-words (BOW) representation is commonly used for text classification tasks. It is a representation in which features are chosen among the words that are present in the training data. Selection techniques are used in order to identify the most suitable words as features. After the feature space is identified, each training and test instance is mapped to this feature representation by giving values to each feature for a certain instance. Two most common feature value representations for BOW representation are: binary feature values—the value of a feature can be either 0 or 1, where 1 represents the fact that the feature is present in the instance and 0 otherwise; or frequency feature values—the value of the feature is the number of times it appears in an instance, or 0 if it did not appear. Because we deal with short texts with an average of 20 words per sentence, the difference between a binary value representation and a frequency value representation is not large. In our case, we chose a frequency value representation. This has the advantage that if a feature appears more than once in a sentence, this means that it is important and the frequency value representation will capture this.

8.2 Genia Tagger

Type of representation is based on syntactic information: noun-phrases, verb-phrases, and biomedical concepts identified in the sentences. In order to extract this type of information, we used the Genia11tagger tool. The tagger analyzes English sentences and outputs the base forms, part-of-speech tags, chunk tags, and named entity tags. The tagger is specifically tuned for biomedical text such as Medline abstracts.



Inhibition	Inhibition	NN	B-NP	O
of	of	IN	B-PP	O
NF-kappaB	NF-kappaB	NN	B-NP	B-protein
activation	activation	NN	I-NP	O
reversed	reverse	VBD	B-VP	O
the	the	DT	B-NP	O
anti-apoptotic	anti-apoptotic	JJ	I-NP	O
effect	effect	NN	I-NP	O
of	of	IN	B-PP	O
isochamaejasmin	isochamaejasmin	NN	B-NP	O
.	.	.	O	O

example of Genia tagger output including for each word: its base form, its part-of-speech, beginning (B), inside (I), outside (O) tags for the word, and the final tag for the phrase.

Fig.2presentsan example of the output of the Genia tagger for the sentence: “Inhibition of NF-kappaB activation reversed the anti-apoptotic effect of is ochamaejasmin.” The noun and verb-phrases identified by the tagger are features used for the second representation technique. We ran the Genia tagger on the entire data set.

CHAPTER - 9

Future Insights

- **Smart Staffing & Personnel Management:** having a large volume of quality data helps health care professionals in allocating resources efficiently. Healthcare professionals can analyze the outcomes of checkups among individuals in various demographic groups and determine what factors prevent individuals from seeking treatment.
- **Advanced Risk & Disease Management:** Healthcare institutions can offer accurate, preventive care. Effectively decreasing hospital admissions by digging into insights such as drug type, conditions, and the duration of patient visits, among many others.
- **Real-time Alerting: Clinical Decision Support (CDS):** applications in hospitals analyzes patient evidence on the spot, delivering recommendations to health professionals when they make prescriptive choices. However, to prevent unnecessary in-house procedures, physicians prefer people to stay away from hospitals
- **Enhancing Patient Engagement:** Every step they take, heart rates, sleeping habits, can be tracked for potential patients (who use smart wearables). All this information can be correlated with other trackable data to identify potential health risks.

CHAPTER - 10

Appendix – Code

10.1 Feature Engineering

```
def get_countid_enocde(train, test, cols, name): temp
  =
train.groupby(cols)['case_id'].count().reset_index().rename(columns =
{'case_id': name})
  temp2 =
test.groupby(cols)['case_id'].count().reset_index().rename(columns =
{'case_id': name})
  train = pd.merge(train, temp, how='left', on= cols) test
  = pd.merge(test,temp2, how='left', on= cols) train[name]
  = train[name].astype('float') test[name] =
  test[name].astype('float')
  train[name].fillna(np.median(temp[name]),   inplace   =   True)
  test[name].fillna(np.median(temp2[name]),   inplace   =   True)
  return train, test

train, test = get_countid_enocde(train, test, ['patientid'], name =
'count_id_patient')
train, test = get_countid_enocde(train, test,
                                  ['patientid', 'Hospital_region_code'],
name = 'count_id_patient_hospitalCode')
train, test = get_countid_enocde(train, test,
                                  ['patientid', 'Ward_Facility_Code'],
name = 'count_id_patient_wardfacilityCode')
```

10.1 Models

10.1.1 Naive Bayes Model

```
from sklearn.naive_bayes import GaussianNB
target = y_train.values
features = X_train.values classifier_nb =
GaussianNB()
model_nb = classifier_nb.fit(features, target)

prediction_nb = model_nb.predict(X_test) from
sklearn.metrics import accuracy_score
acc_score_nb = accuracy_score(prediction_nb,y_test)
print("Acurracy:", acc_score_nb*100)

Acurracy: 34.55439015199096
```

10.1.1 XGBoost Model

```
import xgboost
classifier_xgb = xgboost.XGBClassifier(max_depth=4, learning_rate=0.1,
n_estimators=800,objective='multi:softmax', reg_alpha=0.5, reg_lambda=1.5,
                                   booster='gbtree', n_jobs=4,
min_child_weight=2, base_score= 0.75)

model_xgb = classifier_xgb.fit(X_train, y_train)

prediction_xgb = model_xgb.predict(X_test) acc_score_xgb =
accuracy_score(prediction_xgb,y_test) print("Accuracy:",
acc_score_xgb*100)
```

```
Accuracy: 43.047355859816605
```

10.1.2 Neural Network

```
from keras.utils import to_categorical #Sparse
Matrix
a = to_categorical(y_train) b =
to_categorical(y_test)

model = Sequential()
model.add(Dense(64, activation='relu', input_shape = (254750, 20)))
model.add(Dense(128, activation='relu'))
model.add(Dense(256, activation='relu'))
model.add(Dense(512, activation='relu'))
model.add(Dense(512, activation='relu'))
model.add(Dense(11, activation='softmax'))

model.summary()
```

```
  Model: "sequential"

  Layer (type)                Output Shape              Param #
  ============================================================
  dense (Dense)               (None, 254750, 64)        1344

  dense_1 (Dense)             (None, 254750, 128)       8320

  dense_2 (Dense)             (None, 254750, 256)       33024

  dense_3 (Dense)             (None, 254750, 512)       131584

  dense_4 (Dense)             (None, 254750, 512)       262656

  dense 5 (Dense)             (None, 254750, 11)        5643
  ============================================================
  Total params: 442,571
  Trainable params: 442,571
  Non-trainable params: 0
```

```
model.compile(optimizer= 'SGD',
                                  loss='categorical_crossentropy',
metrics=['accuracy'])
```

CHAPTER - 11

Diagnosis Algorithm

I/P

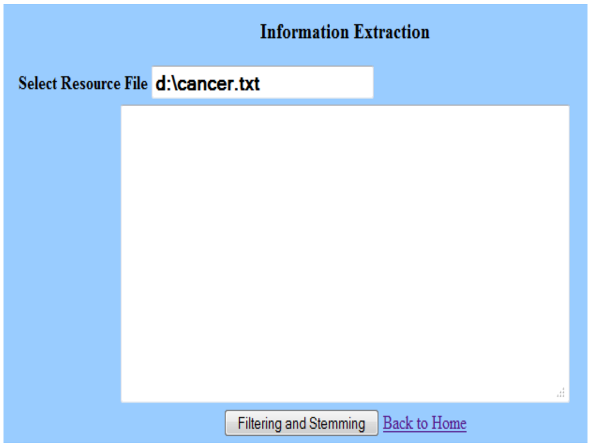
D: a Training Set. N:Number of instance O/P. F:Filtered Dataset.

O:Outlier Dataset.

```
1. Empty F & O.
2. Train(T).
3. Assign i=1
4. If Dw € T then
5. Inser Dw to F else
6. Insert Dw to O end
7. Increase I by 1,then go to step 4
8. Do it until i=N then go to step 8
9. Return F,O
```

Results And Evaluation

Fig11.1.Select abstract file to extract information in short Text.

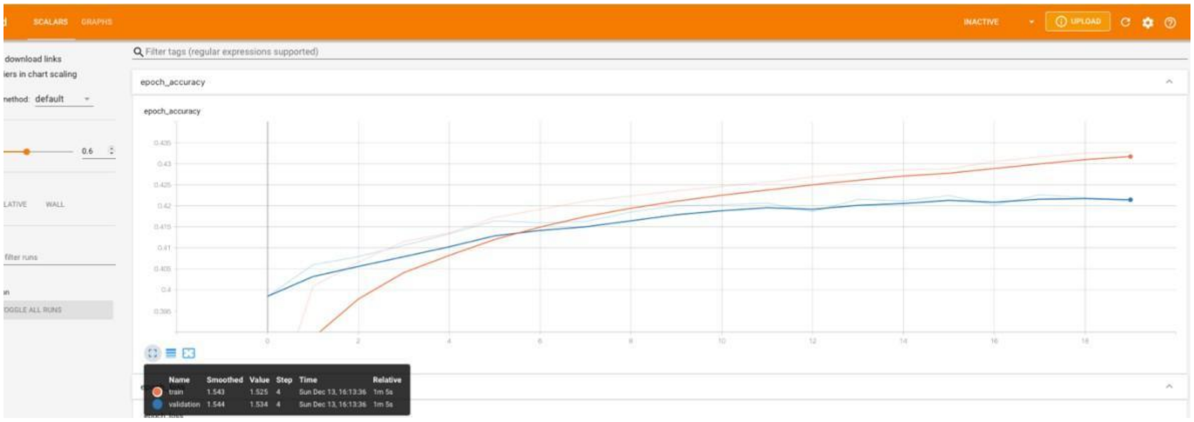


Information Extraction

Select Resource File d:\cancer.txt

```
callbacks = [tf.keras.callbacks.TensorBoard("logs_keras")]
model.fit(X_train, a, epochs=20, callbacks=callbacks, validation_split
= 0.2)

# Genrating tensorboard
!tensorboard --logdir logs_keras
```



	Name	Smoothed	Value	Step	Time	Relative
🟠	train	1.543	1.525	4	Sun Dec 13, 16:13:36	1m 5s
🔵	validation	1.544	1.534	4	Sun Dec 13, 16:13:36	1m 5s

Retraining the model with 4 epochs

```
model.fit(X_train, a, epochs=4, validation_split =
0.2) print("\n Model Evaluation")
model.evaluate(X_test,b)

Model Evaluation
1991/1991 [==============================] - 2s 1ms/step - loss: 1.5071 -
accuracy: 0.4205
```

```
[1.5071101188659668, 0.4204716682434082]
```

.

CHAPTER-12

CONCLUSION

In this project, different variables were analyzed that correlate with Length of Stay by using patient-level and hospital-level data.

By predicting a patient's length of stay at the time of admission helps hospitals to allocate resources more efficiently and manage their patients more effectively. Identifying factors that associate with LOS to predict and manage the number of days patients stay, could help hospitals in managing resources and in the development of new treatment plans. Effective use of hospital resources and reducing the length of stay can reduce overall national medical expenses.its relevant Symptoms, Cause and Treatment. Experimental result shows that the technique used in the proposed work minimizes the time and the work load of the doctors in analyzing information about certain disease and treatment in order to make decision about patient monitoring and treatment. This text mined document can be used in medical health care domain where a doctor can analyse various kinds of treatment that can be given to patient with particular medical disorder. The doctor can update the knowledge related to particular disease or its treatment methodology or the details of medicine that are in research for a particular disease. The doctor can gain idea about particular medicine that are effective for some patient but causes side effect to patient with some additional medical disorder. The patient can also use this extracted documentto get clear understanding about a particular disease its symptoms, side effects, its medicines, its treatment methodologies. This paper also present healthcare diagnosis treatment & prevention of disease, illness, injury in human.

CHAPTER -13

References

- **Janatahack: Healthcare Analytics II -** *Analytics Vidhya* - Link
- **What Is Naive Bayes Algorithm in Machine Learning?** - *Rohit Dwivedi* - Link
- **Naïve Bayes for Machine Learning – From Zero to Hero** - *Anand Venkataraman* - Link
- **XGBoost Parameters** - *XGBoost Documentation* - Link
- **Predicting Heart Failure Using Machine Learning, Part 2**- *Andrew A Borkowski* - Link
- **How to Tune the Number and Size of Decision Trees with XGBoost in Python** - *Jason Brownlee* - Link
- **Big Data Analytics in Healthcare That Can Save People** - *Sandra Durcevic* - Link
- **Learning Process of a Neural Network** – *Jordi Torres* - Link
- Oana Frunza.et.al, "**A Machine Learning Approach For Identifying Disease-Treatment Relations In Short Texts",** May 2011
- L. Hunter And K.B. Cohen**, "Biomedical Language Processing:What's Beyond Pubmed?"** Molecular Cell, Vol. 21-5, Pp. 589-594,2006.
- Jeff Pasternack, Don Roth **"Extracting Article Text From Webb With Maximum Subsequence Segmentation",** WWW 2009 MADRID.
- Abdur Rehman, Haroon.A.Babri, Mehreen saeed,**" Feature Extraction Algorithm For Classification Of Text Document",** ICCIT 2012.
- Adrian Canedo-Rodriguez, Jung Hyoun Kim,etl**.,"Efficient Text Extraction Aalgorithm Using Color Clustering For Language Translation In Mobile Phone"** , May 2012.
- Oana Frunza, Diana Inkpen, and Thomas Tran, Member, IEEE "**A Machine Learning Approach for Identifying Disease- Treatment Relations in Short Texts**" IEEE transactions on knowledge and data engineering, vol. 23, no. 6, june 2011.