Project Report
Analytics for Hospitals' Health-Care Data

1. Introduction

1.1 Project overview

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.

1.2. Purpose

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.

2. Literature survey

2.1 Existing problem

Recent Covid-19 Pandemic has raised alarms over one of the most overlooked areas to focus: Healthcare Management. While healthcare management has various use cases for using data science, patient length of stay is one critical parameter to observe and predict if one wants to improve the efficiency of the healthcare management in a hospital.

2.2. References

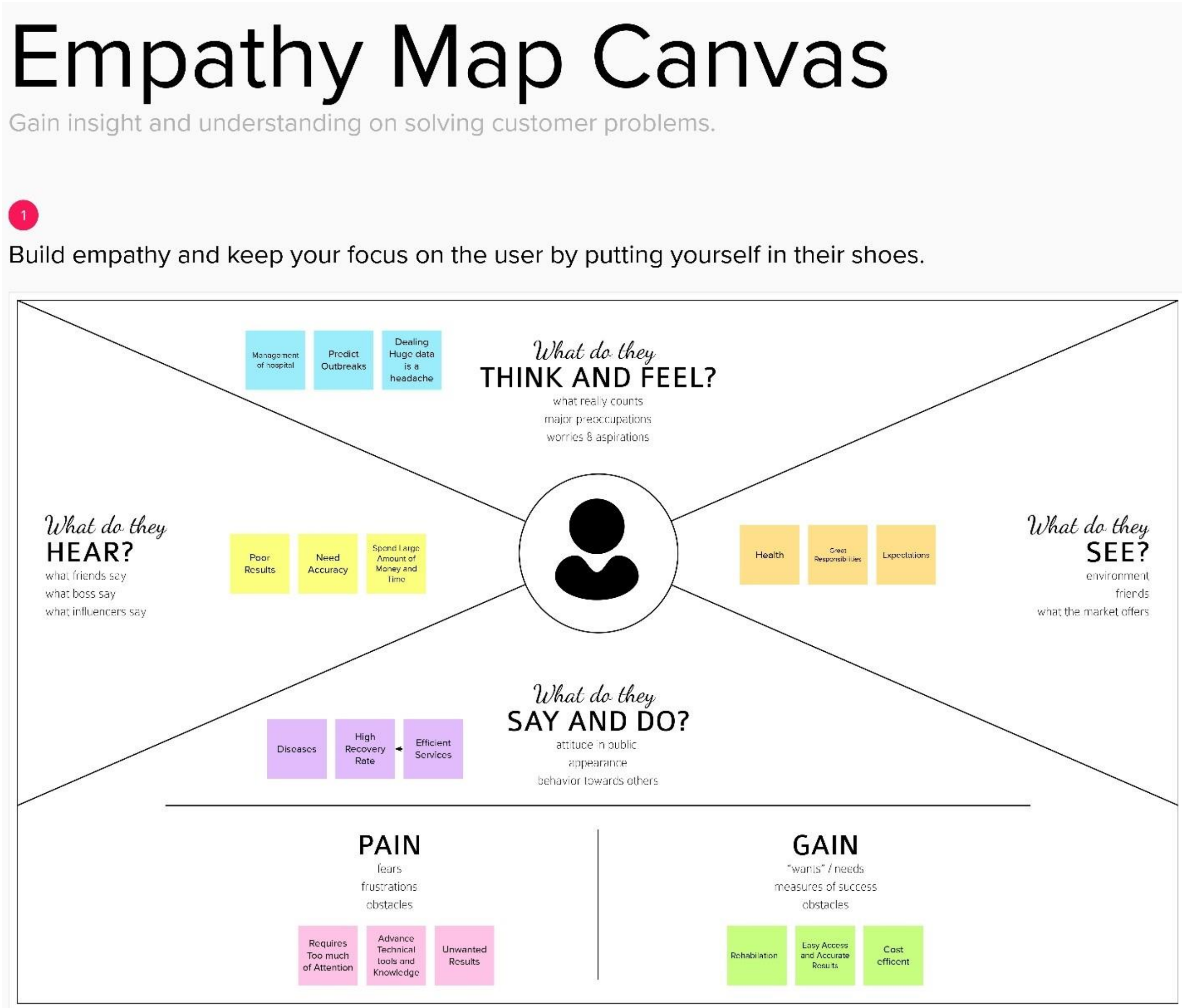
- Janatahack: Healthcare Analytics II - *Analytics Vidhya* - Link
- What Is Naive Bayes Algorithm in Machine Learning? - *Rohit Dwivedi* - Link
- Naïve Bayes for Machine Learning – From Zero to Hero - *Anand Venkataraman* - Link
- XGBoost Parameters - *XGBoost Documentation* - Link
- Predicting Heart 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

2.3. Problem statement

The task is to accurately predict the Length of Stay for each patient on case-by-case basis so that the Hospitals can use this information for optimal resource allocation and better functioning. The length of stay is divided into 11 different classes ranging from 0-10 days to more than 100 days.

3. Ideation & proposed solution

3.1 Empathy map Canvas



3.2 Ideation and Brainstorming

1

Define your problem statement

What problem are you trying to solve? Frame your problem as a How Might We statement. This will be the focus of your brainstorm.

5 minutes

Public hospitals has some main challenges such as deficient infrastructure, deficient manpower, unmanageable patient load and etc.,so peoples can be benefitted if these problems are solved adhering to certain software or some notes to maintain all.
Govt Hospitals facing data management due to lack of IT trained staffs.

Private/Small Health sectors cannot store and analyze large data set it consumes lots of money and time.

Researchers faces issues when they are dealing with large datasets as there is Depicting a diversity of opinions and experiences embedded within patient-generated information(not standard data).

Health Researchers and Students are not able to Extract useful Information's due to lack of data's made available publicly as Many hospitals are not sharing health care data being mindful with patients privacy.

Issues with system functionality, including poor user interfaces and fragmented displays, delayed care delivery. Issues with system access, system configuration, and software updates also delayed care.

2

Brainstorm

Write down any ideas that come to mind that address your problem statement.

10 minutes

TIP

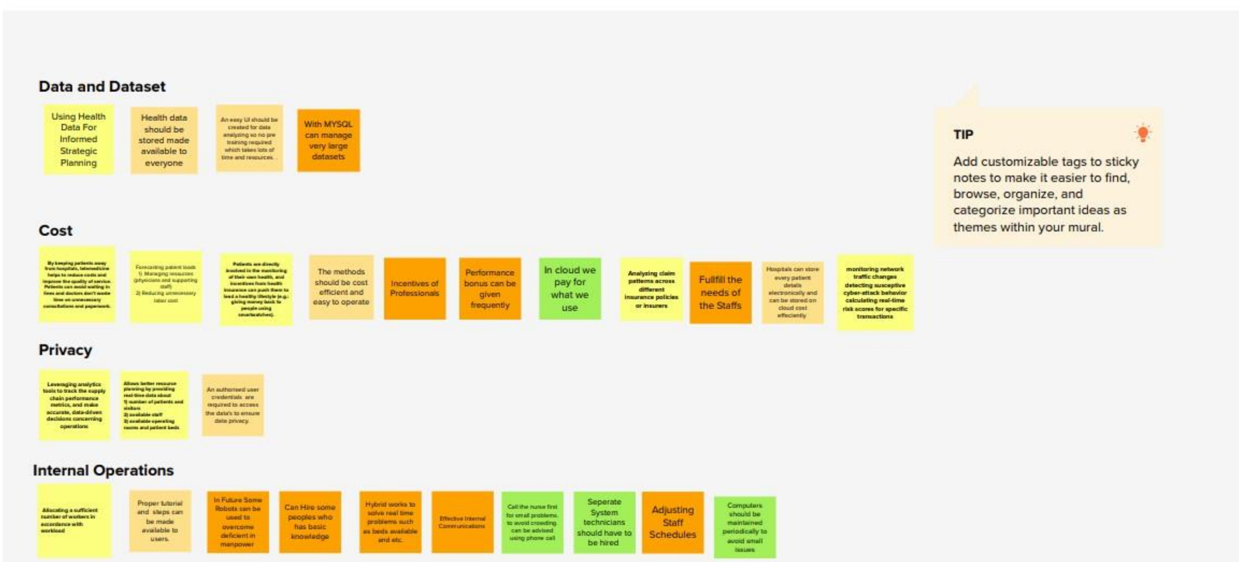
You can select a sticky note and hit the pencil [switch to sketch] icon to start drawing!

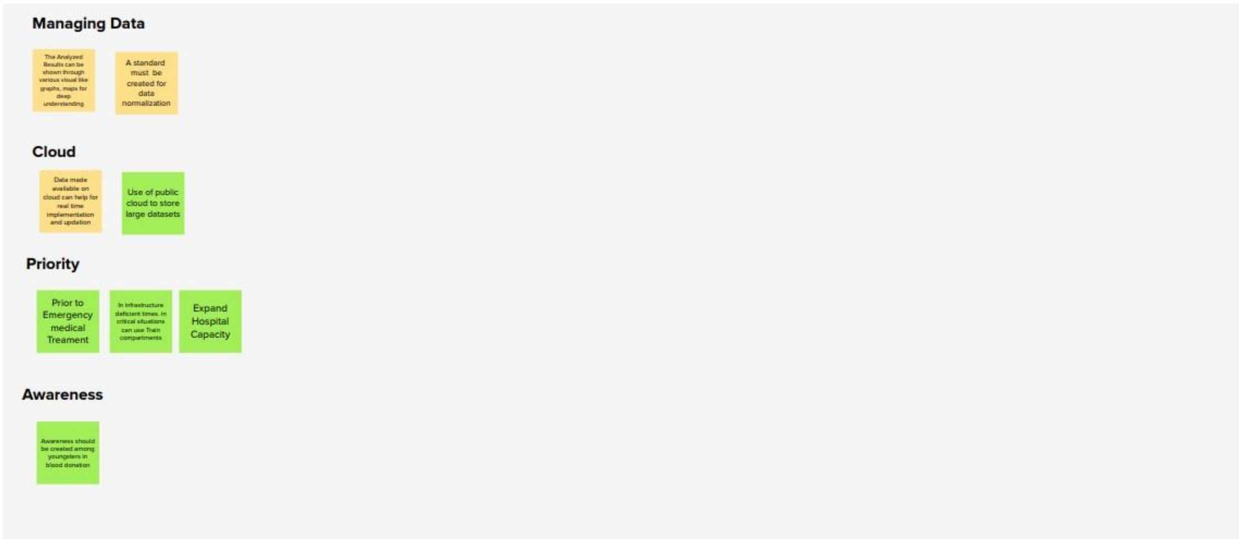
3

Group ideas

Take turns sharing your ideas while clustering similar or related notes as you go. In the last 10 minutes, give each cluster a sentence-like label. If a cluster is bigger than six sticky notes, try and see if you and break it up into smaller sub-groups.

20 minutes

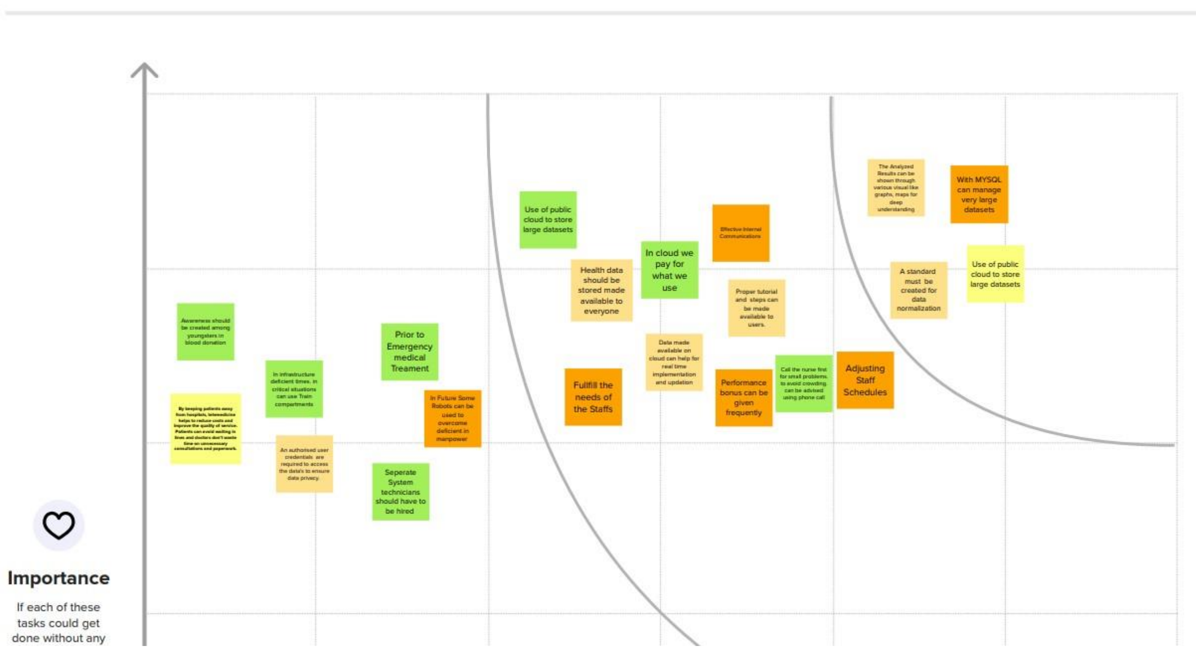


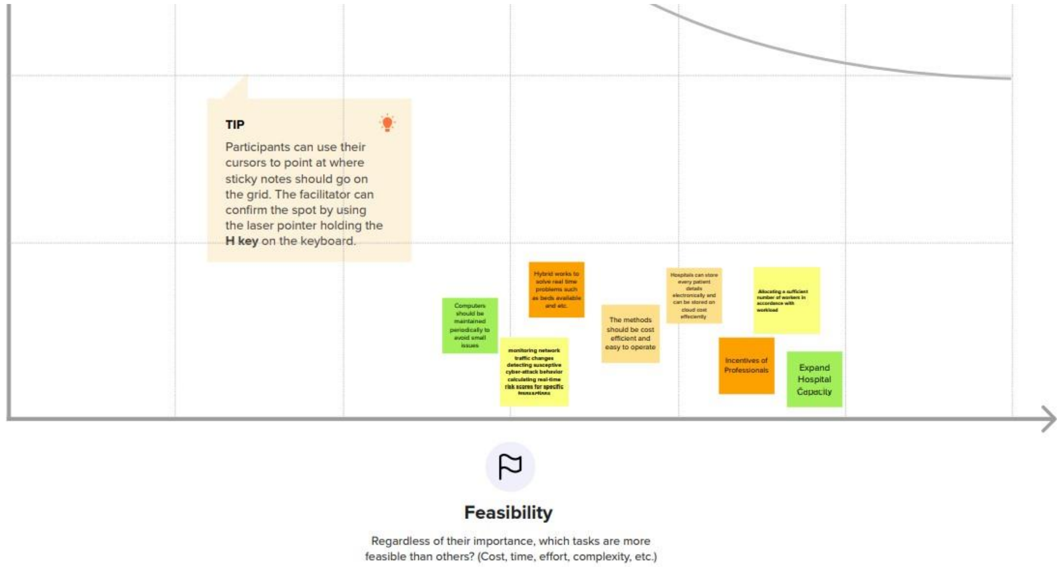


Prioritize

Your team should all be on the same page about what's important moving forward. Place your ideas on this grid to determine which ideas are important and which are feasible.

⏱ **20 minutes**





3.3 **Proposed solution**

S.No.	**Parameter**	**Description**
1.	Problem Statement (Problem to be solved)	The task is to accurately predict the Length of Stay for each patient on case-bycase basis so that the Hospitals can use this information for optimal resource allocation and better functioning. The length of stay is divided into 11 different classes ranging from 0-10 days to more than 100 days.
2.	Idea / Solution description	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 probabilityof 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.

3.	Novelty / Uniqueness	Accurate understanding of the factors associating with the LOS and progressive improvements in processing and monitoring may allow more efficient management of the LOS of inpatients
4.	Social Impact / Customer Satisfaction	A shorter LOS reduces the risk of acquiring staph infections and other healthcare-related conditions, frees up vital bed spaces, and cuts overall medical expenses
5.	Business Model (Revenue Model)	The length of stay (LOS) is an important indicator of the efficiency of hospital management. Reduction in the number of inpatient days results in decreased risk of infection and medication side effects, improvement in the quality of treatment, and increased hospital profit with more efficient bed management
6.	Scalability of the Solution	Remote patient monitoring systems enabling effective distance treatment. Patient portals that allow people to better manage their health themselves;

3.4 Problem solution fit

Define CS, fit into CC

1. CUSTOMER SEGMENT(S) CS

Who is your customer?
i.e. working parents of 0-5 y.o. kids

Hospitals,Medical professionals and hospital staffs are the customers here.

6. CUSTOMER CONSTRAINTS CC

What constraints prevent your customers from taking action or limit their choices of solutions? i.e. spending power, budget, no cash, network connection, available devices.

Limitations for my customer to buy/use my product or services are
1. Difficulty in migrating from manual process because they are used to manual process so are unable to speedily cope with the new system
2. Fear of security breach
3. High cost of software development and deployment
4. Lack of IT-friendly medical personnel
5.Huge influx of patients visiting hospitals

5. AVAILABLE SOLUTIONS AS

Which solutions are available to the customers when they face the problem or need to get the job done? What have they tried in the past? What pros & cons do these solutions have? i.e. pen and paper is an alternative to digital notetaking

The solutions available are
1. Pen and paper method in rural small health cares, which needs to be maintained, manual works, slower and time consuming process.

2. Hospital management system which contains registration. storing details.

Explore AS, differentiate

Focus on J&P, tap into BE, understand RC

2. JOBS-TO-BE-DONE / PROBLEMS J&P

Which jobs-to-be-done (or problems) do you address for your customers? There could be more than one; explore different sides.

The main jobs to be done are
1. Resource allocation
2. Improved patient care
3. Avoid errors and track every single details
4. Improve data security and retrieve ability
5. Enhanced decision making in clinics
6. Easy access to patient data
7. Schedule duties to staffs

9. PROBLEM ROOT CAUSE RC

What is the real reason that this problem exists? What is the back story behind the need to do this job?
i.e. customers have to do it because of the change in regulations.

The main causes are
1. Huge influx of patients visiting hospitals
2. Time consuming to collect, store patient data
3. Lack of security, inconsistency in data entry
4. Prone to damage and being misplaced
5. Hard to make changes, editing problems
6. Limit communication and collaboration
7. Long process to analyse and allocate jobs 8. Lots of manual work

7. BEHAVIOUR BE

What does your customer do to address the problem and get the job done?
i.e. directly related: find the right solar panel installer, calculate usage and benefits; indirectly associated: customers spend free time on volunteering work (i.e. Greenpeace)

1. The customer should quit the existing manual works and move for advanced techniques
2. Use hospital managements systems
3. Purchase products or services that stores, maintains and process the data
4. Use analytics
5. Use advanced technology to analyze and work on patients data

Focus on J&P, tap into BE, understand RC

Identify strong TR & EM

3. TRIGGERS TR

What triggers customers to act? i.e. seeing their neighbour installing solar panels, reading about a more efficient solution in the news.

The triggers for my customers are
1. Facing the existing challenges, and difficulties
2. Looking at other sectors growing
3. Advancements and growth in technology
4. Increased productivity from hospital management system
5. Increased analytics work

10. YOUR SOLUTION SL

If you are working on an existing business, write down your current solution first, fill in the canvas, and check how much it fits reality.
If you are working on a new business proposition, then keep it blank until you fill in the canvas and come up with a solution that fits within customer limitations, solves a problem and matches customer behaviour.

The answer is to accurately predict the Length of Stay(LOS) for each patient on case by case basis so that the Hospitals can use this information for optimal resource allocation and better functioning. This parameter helps hospitals to identify patients of high LOS risk at the time of admission. Once identified, patients with high LOS risk can optimise their treatment plan to minimize LOS and lower the chance of staff/visitor infection. Also, prior knowledge of LOS can aid in logistics such as room and bed allocation planning. An informative, creative dashboard can be created to present the data and utilize it for prior proper planning and resource allocation.

8. CHANNELS of BEHAVIOUR CH

8.1 ONLINE
What kind of actions do customers take online? Extract online channels from #7

8.2 OFFLINE
What kind of actions do customers take offline? Extract offline channels from #7 and use them for customer development

8.1 ONLINE

Customers can purchase the service/product and use it to store patients data regularly, maintain their details, create dashboards and work on it online efficiently and effectively

8.2 OFFLINE

By Using the collected data, customers can interpret, analyze, and utilize the data to allocate resources, schedule jobs to staffs, do planning for proper management of hospital

Identify strong TR & EM

4. EMOTIONS: BEFORE / AFTER EM

How do customers feel when they face a problem or a job and afterwards?
i.e. lost, insecure > confident, in control - use it in your communication strategy & design.

Before
1. Frustrated, confused
2. Inefficient time management
3. Poor resource allocation, staffing
4. Worried about huge stuff of work, workload
5. Work pressure

After
1. Secured, find it easy, efficient and reliable
2. Efficient time management
3. Better resource allocation
4. Less manual work 5. Need to develop technical knowledge

4. Requirements analysis

4.1 Functional requirements

FR No.	Functional Requirement (Epic)	Sub Requirement (Story / Sub-Task)

FR1	User Registration	Registration through Form Registration through Gmail Registration through LinkedIN
FR2	User Confirmation	Confirmation via Email Confirmation via OTP
FR3	Operability	Share patient data and make it interoperable among the management
FR4	Accuracy	The dashboard will be able to predict length of stay based on multiple combinations based on input sources with a n accuracy of upto 85%
FR5	Compliance	The product is to be used within the hospital so any form of data need not be hidden
FR6	Productivity	The dashboard is believed to improve the predictions of Length of Stay and thereby creating a scenario of providing better solution

4.2.Nonfunctional requirements

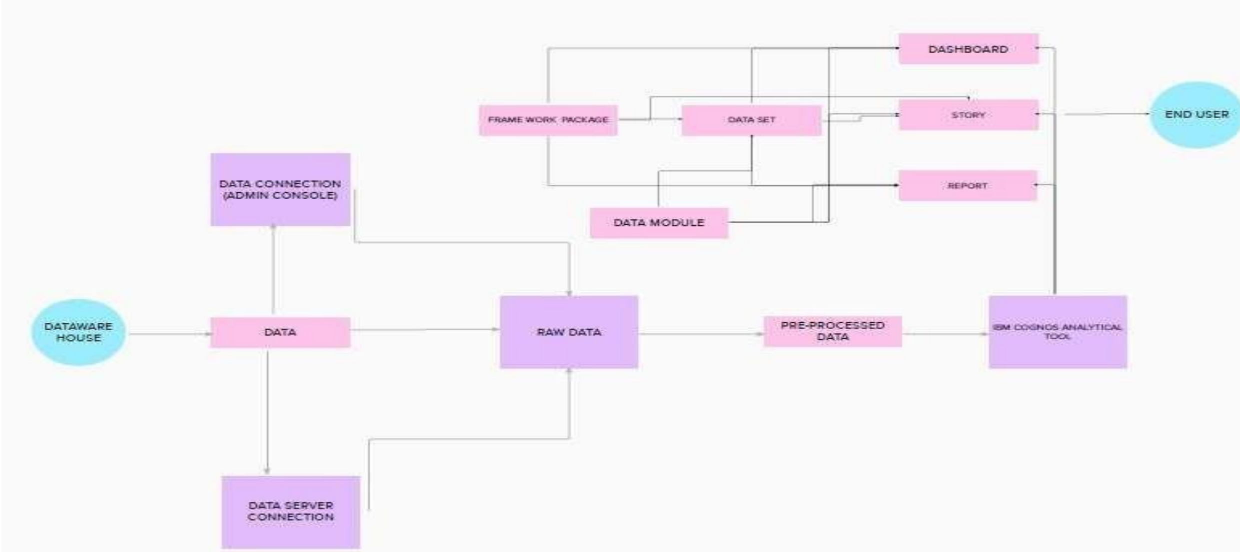
FR No.	**Non-Functional Requirement**	**Description**
NF R-1	**Usability**	This Dashboards are designed to offer a comprehensive overview of patient's LOS, and do so through the use of data visualization tools like charts and graphs.
NF R-2	**Security**	General industry level security shall be provided
NF R-3	**Reliability**	This dashboard will be consistent and reliable to the users and helps the user to use in effective, efficient and reliable manner.
NF R-4	**Performance**	The dashboard reduces the time needed for analysing data and has an automated system for that which improves the performance
NF R-5	**Availability**	The dashboard can available to meet user's demand in timely manner and it is also helps to provide necessary information to the user's dataset

NF R-6	**Scalability**	It is a multi-tenant system which is capable of rimming on lower-level systems as well.

5. PROJECT DESIGN

5.1 Data Flow Diagrams

A Data Flow Diagram (DFD) is a traditional visual representation of the information flows within a system. A neat and clear DFD can depict the right amount of the system requirement graphically. It shows how data enters and leaves the system, what changes the information, and where data is stored.



5.2 Solution & Technical Architecture

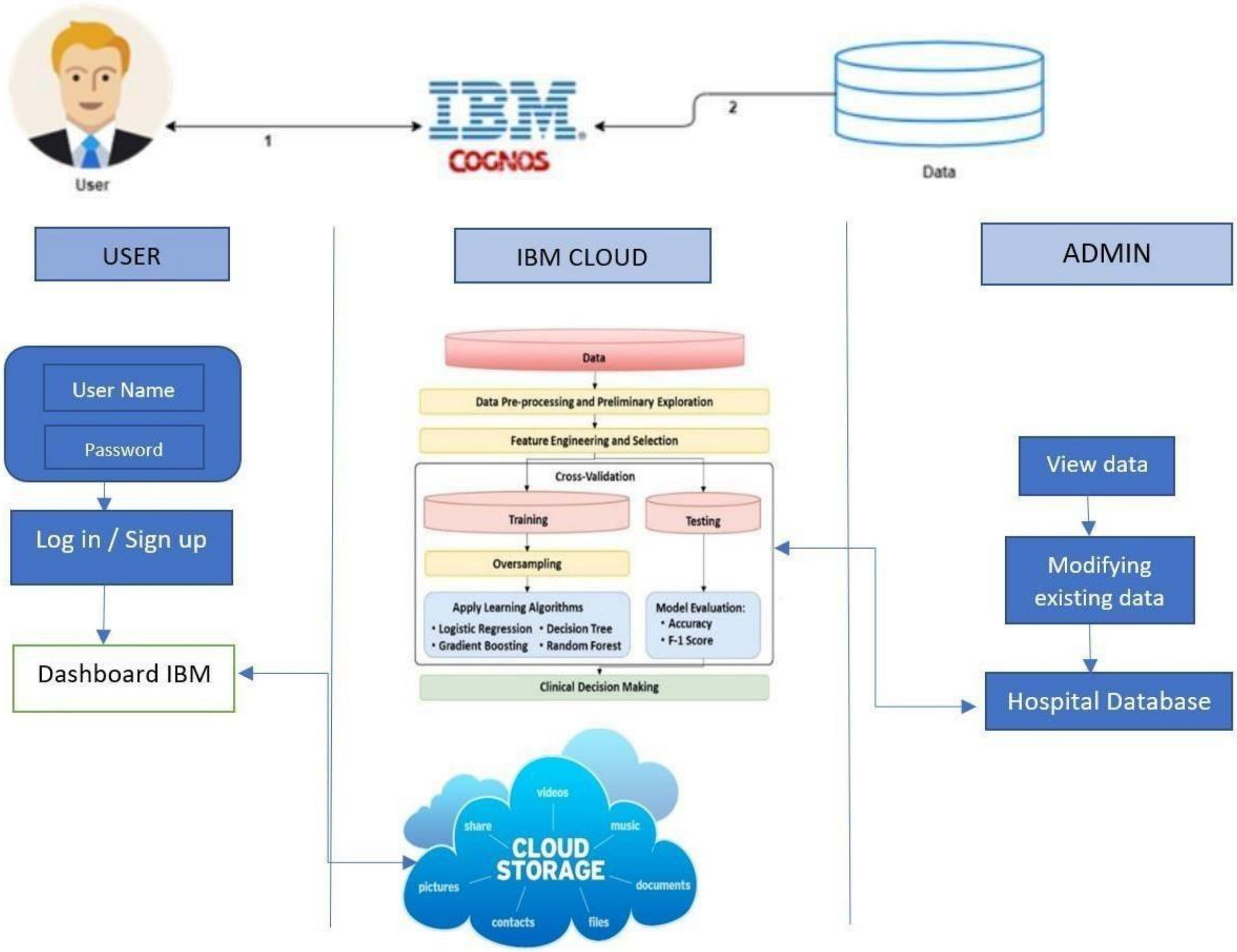


Table1: Components & Technologies:

S. No	Component	Description	Technology
1.	User Interface	How user interacts with application e.g., Web UI, Mobile App,	HTML, CSS, JavaScript / Angular Js / React Js etc…
2.	Application Logic-1	Chatbot etc. Logging in as a patient / user in the application	Python
3.	Application Logic-2	Logging in as an admin in the application	IBM Watson Assistant
5.	Database	All the data about patients such as disease, address and	MySQL, NoSQL, etc.

6.	Cloud Database	etc. IBM Watson cloud is used for storage, Cloud	IBM DB2, IBM Cloud ant etc.
7.	External API-1	Purpose of External API used in the application	Aadhar API, etc.
8.	Machine Learning Model	Purpose of Machine Learning Model	Regression Model, etc.
9.	Infrastructure (Server / Cloud)	Application Deployment on Local System / Cloud Local Server Confi gurati on,	Local, Cloud Foundry, Kubernetes, etc.

Table-2: Application Characteristics: Cloud

S. No	Characteristics	Description	Technology
1.	Open-Source Frameworks	List the open-source frameworks used	Python
2.	Security Implementations	List all the security / access controls implemented, use of	Encryption.
3.	Scalable Architecture	firewalls etc. Justify the scalability of architecture (3 –	Can supports higher workloads
4.	Availability	tier, MicroJustify the availability of application (e.g. use of load balancers,	Highly available
5.	Performance	distributed servers Design consideration for the performance of the application (number of requests per sec, use of	It performs good uses various tools and ideas in a scientific manner to meet the desired outcomes

Cache, use of

5.3 User Stories

User Type	Functional Requirement (Epic)	User Story Number	User Story / Task	Acceptance criteria	Priority	Release
Customer	Dashboard	USN 1	As a user, I can upload the datasets to the dashboard	I can access various operations	Medium	Sprint-4
	View	USN 2	As a user, I can view the patient details	I can view the visual data and the result after the prediction	Medium	Sprint-3
Admin	Analyse	USN 3	As an admin, I will analyse the given dataset	I can analyse the dataset	High	Sprint-2
User Type	Functional Requirement (Epic)	User Story Number	User Story / Task	Acceptance criteria	Priority	Release
	Predict	USN 4	As an admin, I will predict the length of stay	I can predict the length of stay	High	Sprint-1

6. Project planning & scheduling

6.1. Sprint Planning & Estimation

Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority	Team Members
Sprint-1	Data Collection	USN-1	The User needs a complete data about the patients admitted in the hospital and a dataset should be prepared.	2	Medium	Dinesh, Santhosh, Azhageshwaran
Sprint-1	Data Exploration	USN- 2	As a user, I need nicely visualized dashboard of number of beds occupied and number of free beds in hospital.	4	High	Selvakumar, Azhageshwaran, Santhosh
Sprint-2	Track of patient visit of Hospital	USN-3	Tracking a patient Health care over years of visit and Screening of data they have in hospital.	2	Medium	Dinesh, Selvakumar, Azhageshwaran
Sprint-2	Dashboard	USN - 4	As a user, I want the interactive dashboard to analyse the data. Have the data in terms of Graph.	4	High	Selvakumar, Dinesh, Santhosh
Sprint3	Detailed EHR's of patient	USN-5	Provided greater details in the EHR's of individual patient with clear idea of	2	Medium	Azhageshwaran

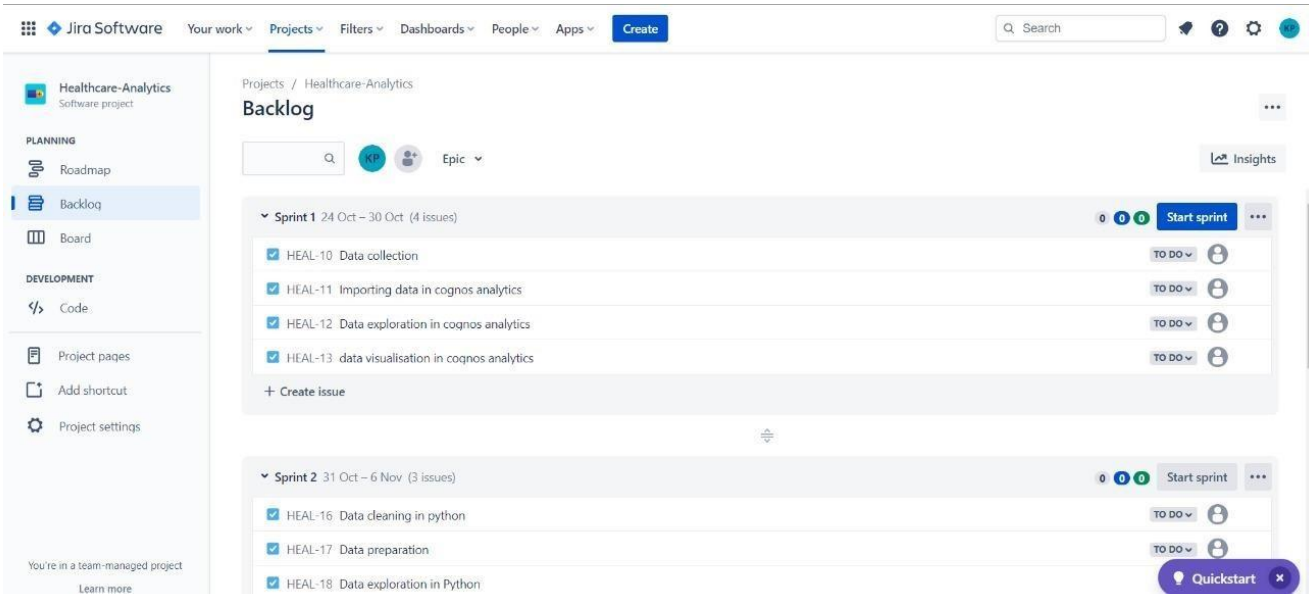
Sprint-3	Story Creation	USN-6	what to do. As a user, I need the story animation of the data set with	4	High	Dinesh, Selvakumar
Sprint4	Predict LOS	USN-7	insights As a user, I want the flawless system to predict the length of stay of the patients	4	High	Dinesh, Azhageshwaran, Santhosh
Sprint4	Using ML algorithm for Prediction	USN-8	As a user, I need prior knowledge of LOS can aid in logistics such as room and bed allocation planning.	4	High	Azhageshwaran, Selvakumar, Santhosh

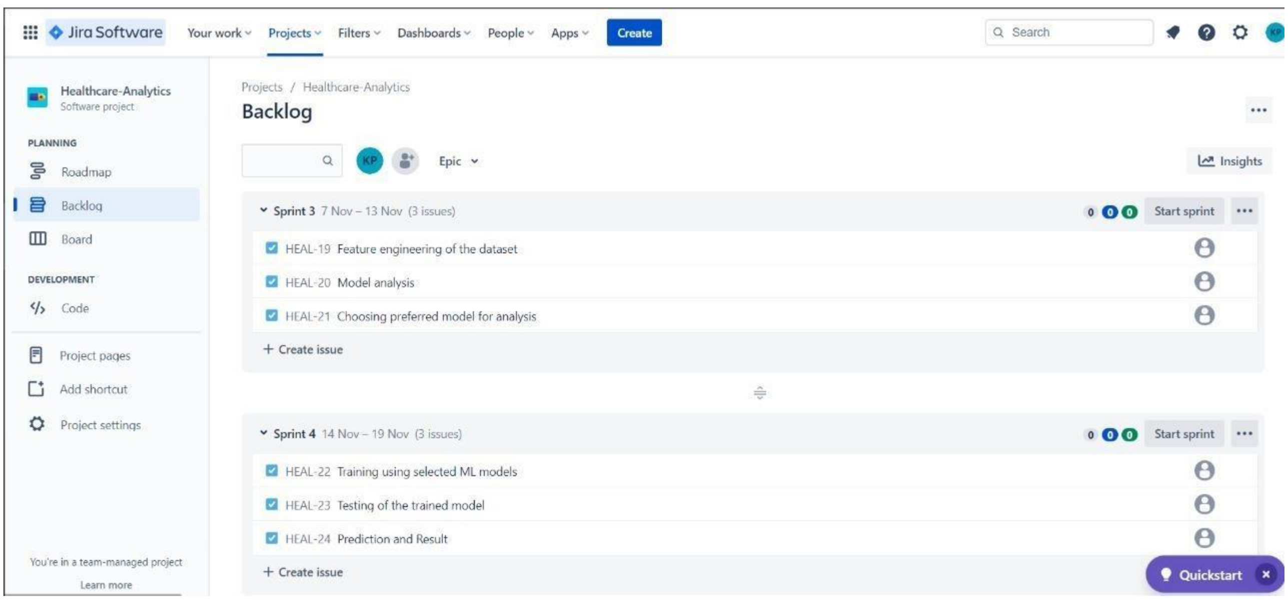
6.2.Sprint Delivery Schedule

Sprint	Total Story Points	Duration	Sprint Start Date	Sprint End Date (Planned)	Story Points Completed (as on Planned End Date)	Sprint Release Date (Actual)
Sprint-1	20	6 Days	24 Oct 2022	29 Oct 2022	20	29 Oct 2022
Sprint-2	20	6 Days	31 Oct 2022	05 Nov 2022	20	05 Nov 2022
Sprint-3	20	6 Days	07 Nov 2022	12 Nov 2022	20	12 Nov 2022
Sprint-4	20	6 Days	14 Nov 2022	19 Nov 2022	20	19 Nov 2022

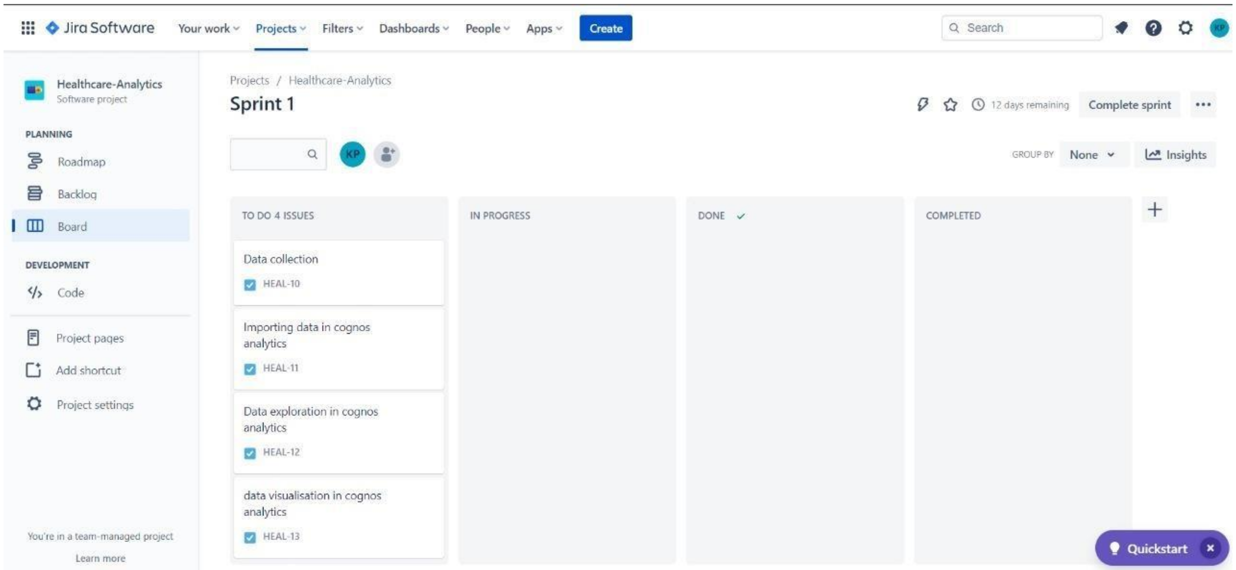
6.3. Reports from JIRA

Jira Sprints

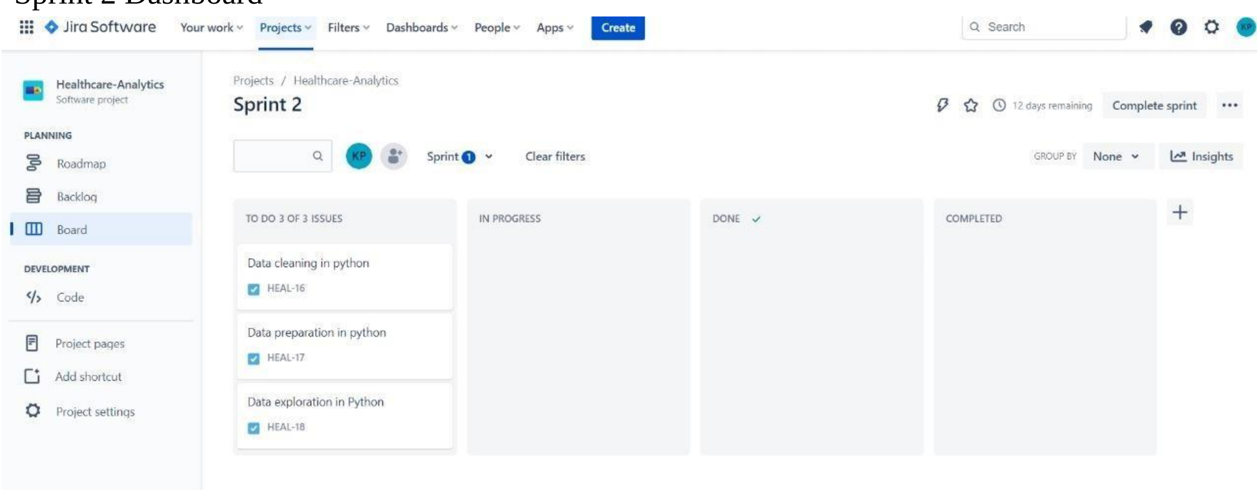




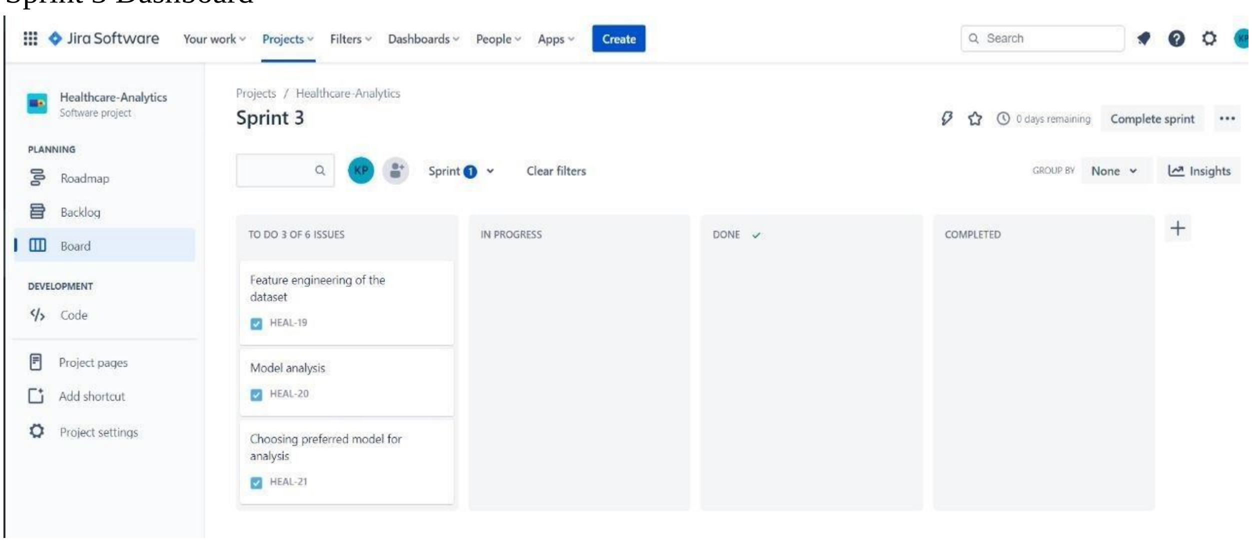
Sprint 1 Dashboard



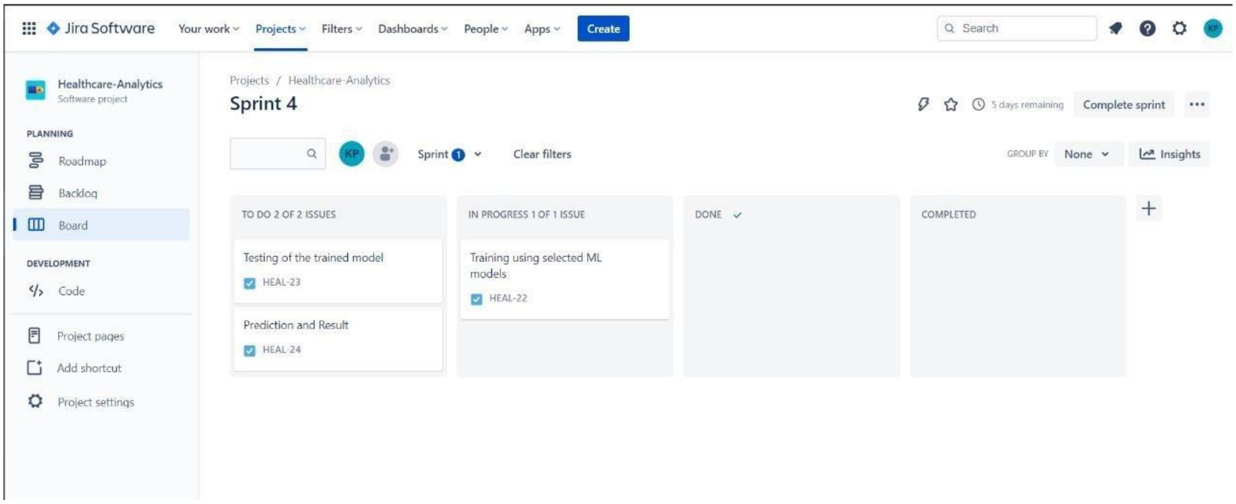
Sprint 2 Dashboard



Sprint 3 Dashboard



Sprint 4 Dashboard



7. Coding & solutioning ML Models Naive Bayes Model

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)

$$P(H \mid E) = \left[\boldsymbol{P(E \mid H) / P(E)} \right] P(H)$$

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.

2) XGBoost Model

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.
- 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%. This model nearly took 25 minutes to get trained but when compared to the Naïve Bayes model it gave an 8.5% improvement.

3) Neural Network Model

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.In this neural network model, there are **six** dense layers, 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. Every layer is activated using "**relu**" activation function because it overcomes the vanishing gradient problem, allowing models to learn faster and perform better.

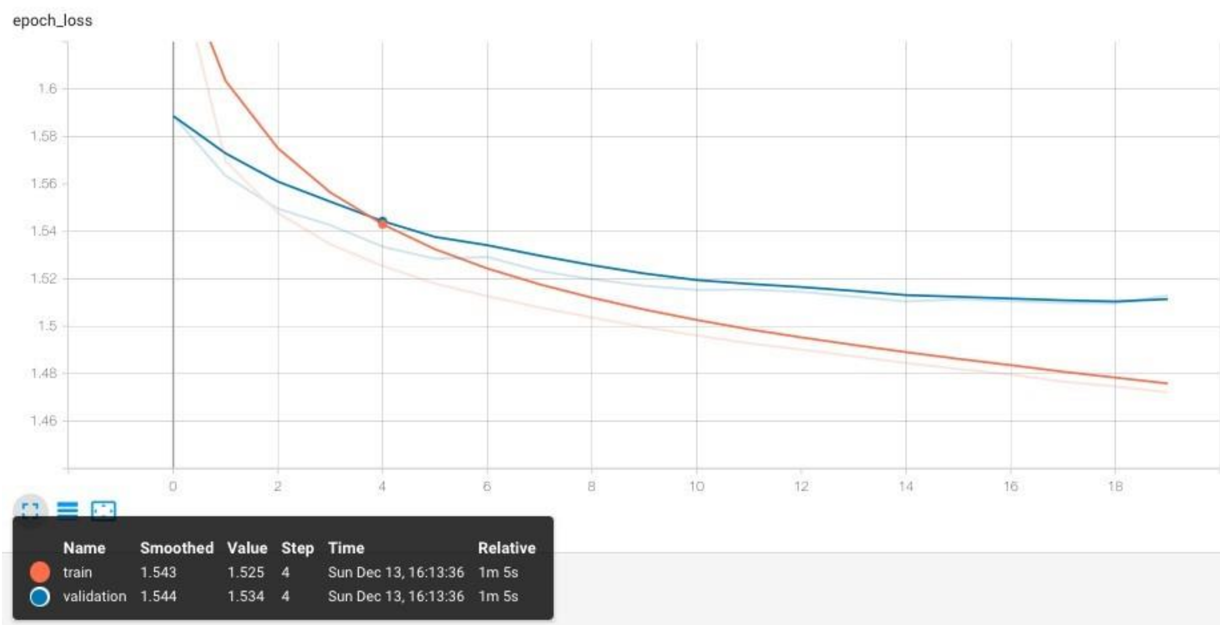
Finally, evaluating the model with a test set yields an accuracy score of **41.79%**. Neural Networks supposedly performs better than any other models. But because of the smaller dataset, it was not able to learn more accurately than the XGBoost model. It nearly took 20 minutes to train the model.

In the Naive 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. Whereas the other two models XGBoost and Neural Networks are predicting mostly similar Length of Stay for the patient

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.

9) Results

9.1 Performance metrics



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.

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

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 we can see that the observations classified by both these models are marginally similar.

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.

10) Advantages:

1.By predic5ng a pa5ent's length of stay at the 5me of admission helps hospitals to allocate resources more efficiently and manage their pa5ents more effec5vely

2.It helps hospitals in managing resources and in the development of new treatment plans

3. Effective use of hospital resources and reducing the length of stay can reduce overall national medical expenses.

11) 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.

12) 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.

Appendix:

Code:

Load the Dataset:

```
d1 = pd.read_csv('/content/drive/My Drive/Healthcare_Data/sample_sub.csv')
d2 = pd.read_csv('/content/drive/My Drive/Healthcare_Data/train_data_dictionary.csv')
test = pd.read_csv('/content/drive/My Drive/Healthcare_Data/test_data.csv')
train = pd.read_csv('/content/drive/My Drive/Healthcare_Data/train_data.csv')
```

In [163]:

```
from google.colab import drive
drive.mount('/content/drive')
```

Data Exploration

```
train.head()
train.info()
train.Stay.unique()
train.isnull().sum().sort_values(ascending = False)
train.shape
test.shape
for i in train.columns:
   print(i, ':', train[i].nunique())
for i in test.columns:
   print(i, ':', test[i].nunique())
```

Data Preparation

```
train['Bed Grade'].fillna(train['Bed Grade'].mode()[0], inplace = True)

test['Bed Grade'].fillna(test['Bed Grade'].mode()[0], inplace = True)

#Replacing NA values in  Column for both Train and Test datssets

train['City_Code_Patient'].fillna(train['City_Code_Patient'].mode()[0],
inplace = True)
```

```
test['City_Code_Patient'].fillna(test['City_Code_Patient'].mode()[0], inplace
= True)

# Label Encoding Stay column in train dataset

from sklearn.preprocessing import LabelEncoder

le = LabelEncoder()

train['Stay'] = le.fit_transform(train['Stay'].astype('str'))

train.head()
```

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')

# Droping duplicate columns

test1 = test.drop(['Stay', 'patientid', 'Hospital_region_code',
'Ward_Facility_Code'], axis =1)

train1 = train.drop(['case_id', 'patientid', 'Hospital_region_code',
'Ward_Facility_Code'], axis =1)

# Splitting train data for Naive Bayes and XGBoost

X1 = train1.drop('Stay', axis =1)

y1 = train1['Stay']

from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X1, y1, test_size =0.20,
random_state =100)
```

Naïve Bayes

```
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)
```

Neural Network

```
X = train.drop('Stay', axis =1)

y = train['Stay']

print(X.columns)

z = test.drop('Stay', axis = 1)

print(z.columns)
```

Predictions

```
pred_nb = classifier_nb.predict(test1.iloc[:,1:])

result_nb = pd.DataFrame(pred_nb, columns=['Stay'])

result_nb['case_id'] = test1['case_id']

result_nb = result_nb[['case_id', 'Stay']]

result_nb['Stay'] = result_nb['Stay'].replace({0:'0-10', 1: '11-20', 2: '21-30',
3:'31-40', 4: '41-50', 5: '51-60', 6: '61-70', 7: '71-80', 8: '81-90', 9: '91-100', 10:
'More than 100 Days'})

result_nb.head()
```

Results

```
print(result_nb.groupby('Stay')['case_id'].nunique())
```

GitHub link: https://github.com/IBM-EPBL/IBM-Project-10457-1659180600