ANALYTICS FOR HOSPITAL'S HEALTHCARE DATA

PROJECT REPORT

SUBMITTED BY

SANTHOSH G-737819CSR175
SOWBARNIKA P S-737819CSR188
VIDYAKEERTHI SU-737819CSR218
YOGAPRIYA S-737819CSR229

TEAM ID: PNT2022TMID04527

1.1 INTRODUCTION

1.2 Project overview:

The pressure on healthcare institutions to enhance patient outcomes and provide better care is expanding. Even while this situation is difficult, it also gives enterprises a chance to significantly raise the standard of care by utilizing additional information and insights from their data. Health care analytics is the term for the efficiently analysis of data to discover patterns and trends in the collected data. The average duration of stay for a patient is one of many performance measures used in healthcare management. With the help of the project Hospitals can tailor their treatment programmers to minimize length of stay (LOS) and cut down on infection rates among patients, workers, and all the people in the hospital.

1.2. Purpose

The project objective is to precisely estimate each patient's length of stay, in order to effectively utilize hospital resources.

2. LITERATURE SURVEY

2.1 Existing problem

Covid-19 recently One of the most neglected areas to concentrate on has come under scrutiny due to the pandemic: healthcare management. Patient duration of stay is a crucial statistic to monitor and forecast if one wishes to increase the effectiveness of healthcare management in a hospital, even if there are many use cases for data science in healthcare management.

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 - *JasonBrownlee* - Link
- Big Data Analytics in Healthcare That Can Save People - *Sandra Durcevic* - Link

2.3. Problem statement

The goal is to correctly anticipate the length of stay for each patient on a case-by-case basis so that hospitals may utilize this data to better allocate resources and operate. The length of stay is divided into 11 different classes ranging from 0-10 days to more than 100 days.

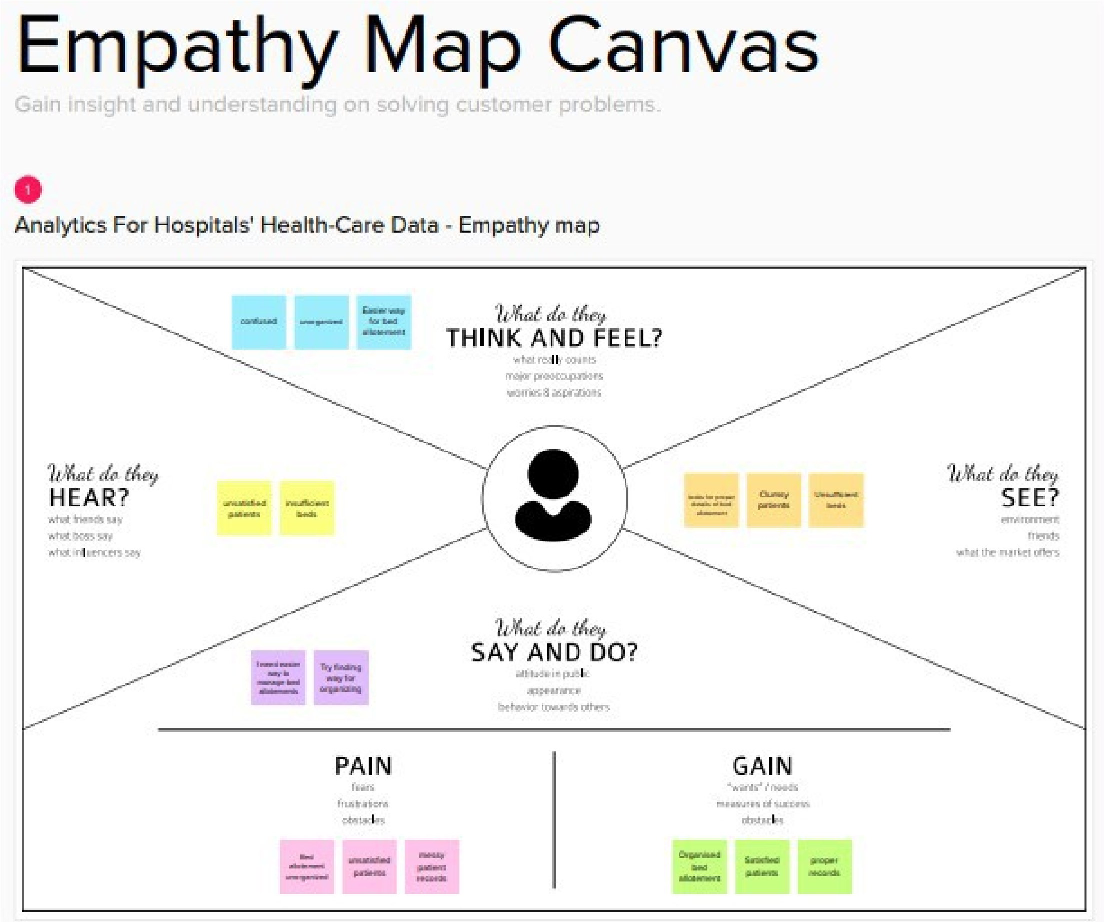
S.NO	PAPER	AUTHOR	YEAR	METHOD AND ALGORITHM	ACCURACY
1	Machine learning model for predicting the length of stay in the intensive care unit for Covid-19 patients in the eastern province of Saudi Arabia	Dina A. Alabbad, Abdullah M. Almuhaideb, Shikah J. Alsunaidi, Kawther S. Alqudhaihai, Fatimah A. Alamoudi, Maha K. Alhobaishi, Naimah A. Alaqeel, Mohammed S. Alshahrani	2022	Random Forest (RF), Gradient Boosting (GB), Extreme Gradient Boosting (XGBoost), and Ensemble models	94.16%
2	Predicting length of stay in hospitals intensive care unit using general admission features	Merhan A. Abd-Elrazek a , Ahmed A. Eltahawi b,⇑ , Mohamed H. Abd Elaziz c , Mohamed N. AbdElwhab d	2021	ML techniques used are Neural Networks(NN), Classification Tree(CT), Tree Bagges(TB), Random Forest(RF), Fuzzy Logic(FL), Support Vector Machine(SVM) , KNN, Regression Tree(RT) and Navie Bayes(NB)	92%

3	Pandemic Analytics: How Countries are Leveraging Big Data Analytics and Artifcial Intelligence to Fight COVID−19	Nishita Mehta1 · Sharvari Shukla2	2021	big data and AI techniques	90%
4	Applications of big data analytics to control COVID19 Pandemic	Shikah J. Alsunaidi 1, Abdullah M. Almuhaideb 2,*, Nehad M. Ibrahim 1, Fatema S. Shaikh 3, Kawther S. Alqudaihi 1, Fahd A. Alhaidari 2, Irfan Ullah Khan 1, Nida Aslam 1 and Mohammed S. Alshahrani	2021	artificial intelligence (AI); big data; big data analytics.	98%

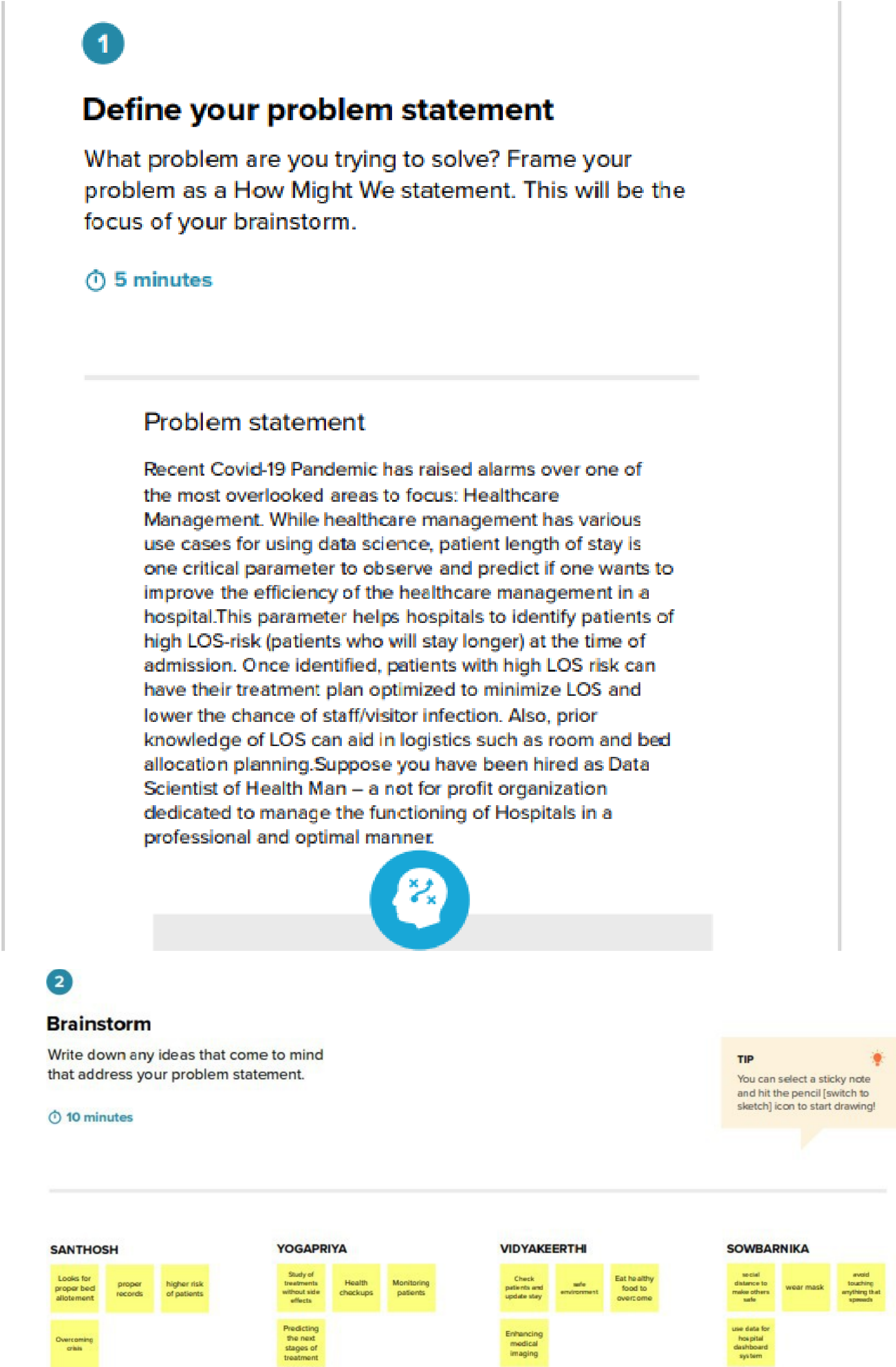
5	Data Science in Healthcare: COVID -19 and Beyond	Tim Hulsen	2020	ML ,Deep Learning,AI,NL P	95%

3. IDEATION & PROPOSED SOLUTION

3.1 Empathy map canvas



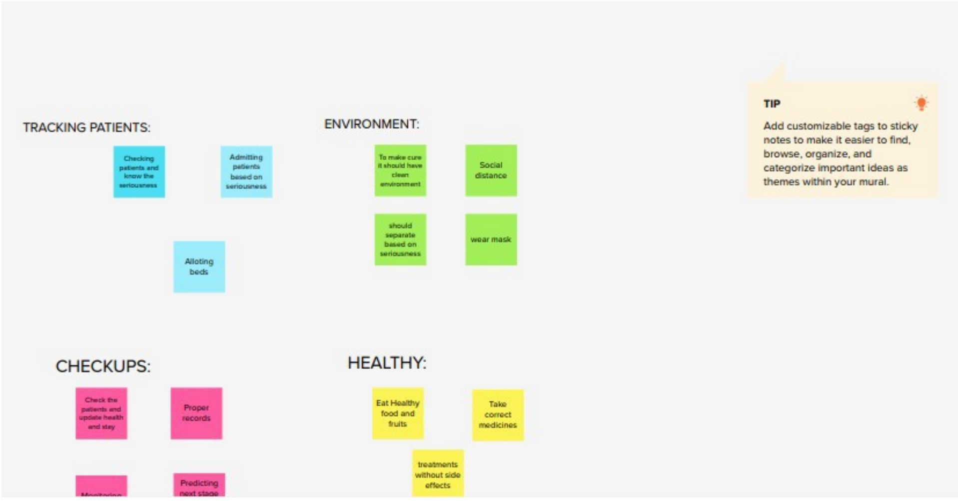
3.2 Ideation and Brainstorming



Group ideas

Take turns sharing your ideas while clustering similar or related notes as you go. Once all sticky notes have been grouped, 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.

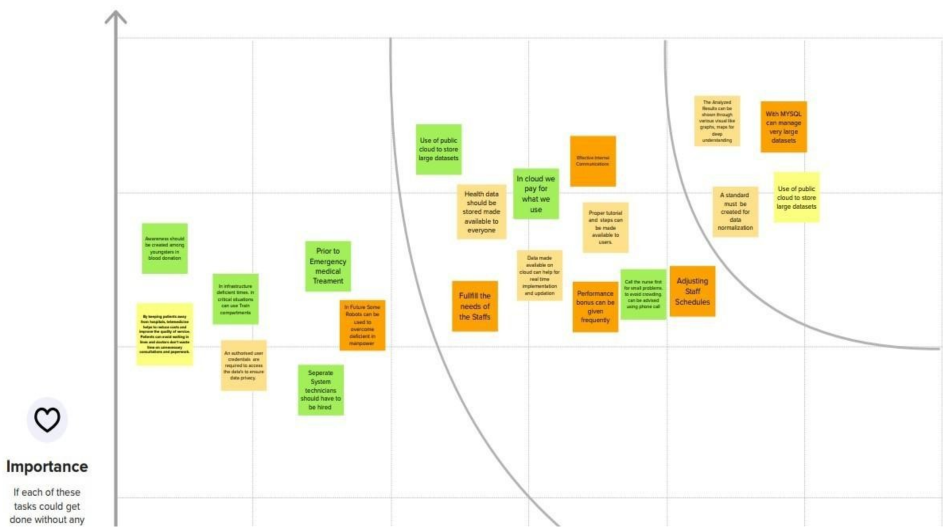
⏱ 20 minutes



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

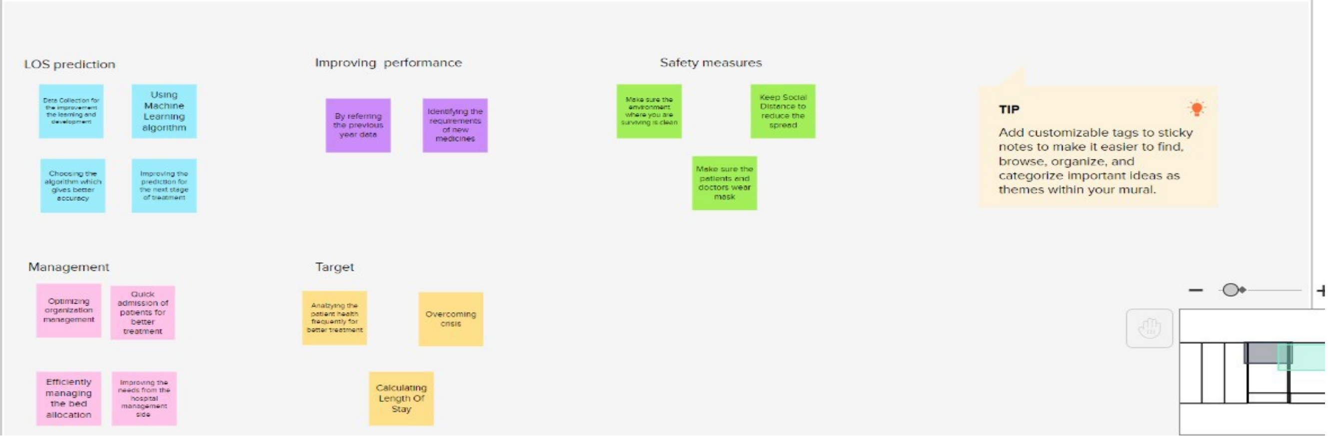


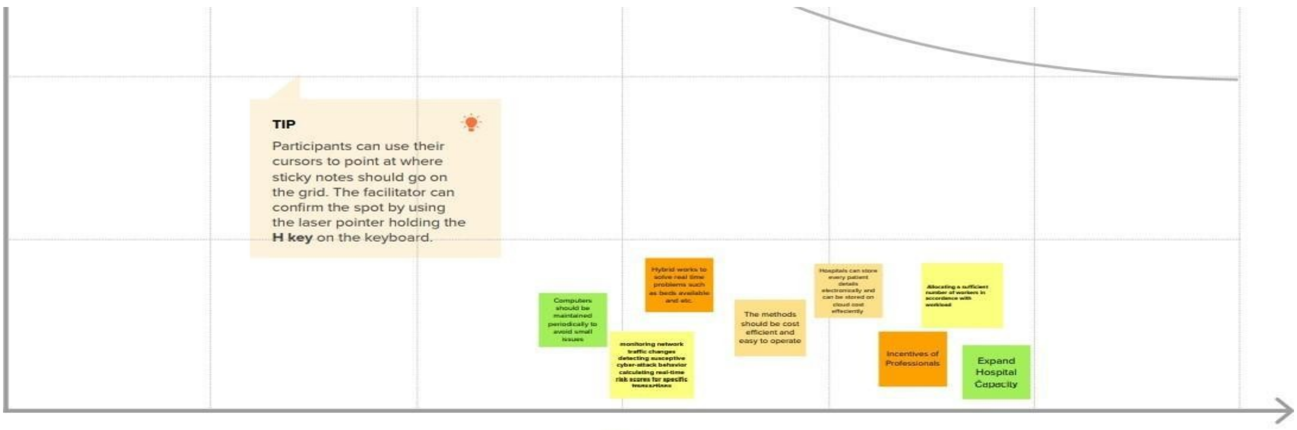
3

Group ideas

Take turns sharing your ideas while clustering similar or related notes as you go. Once all sticky notes have been grouped, 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





Feasibility

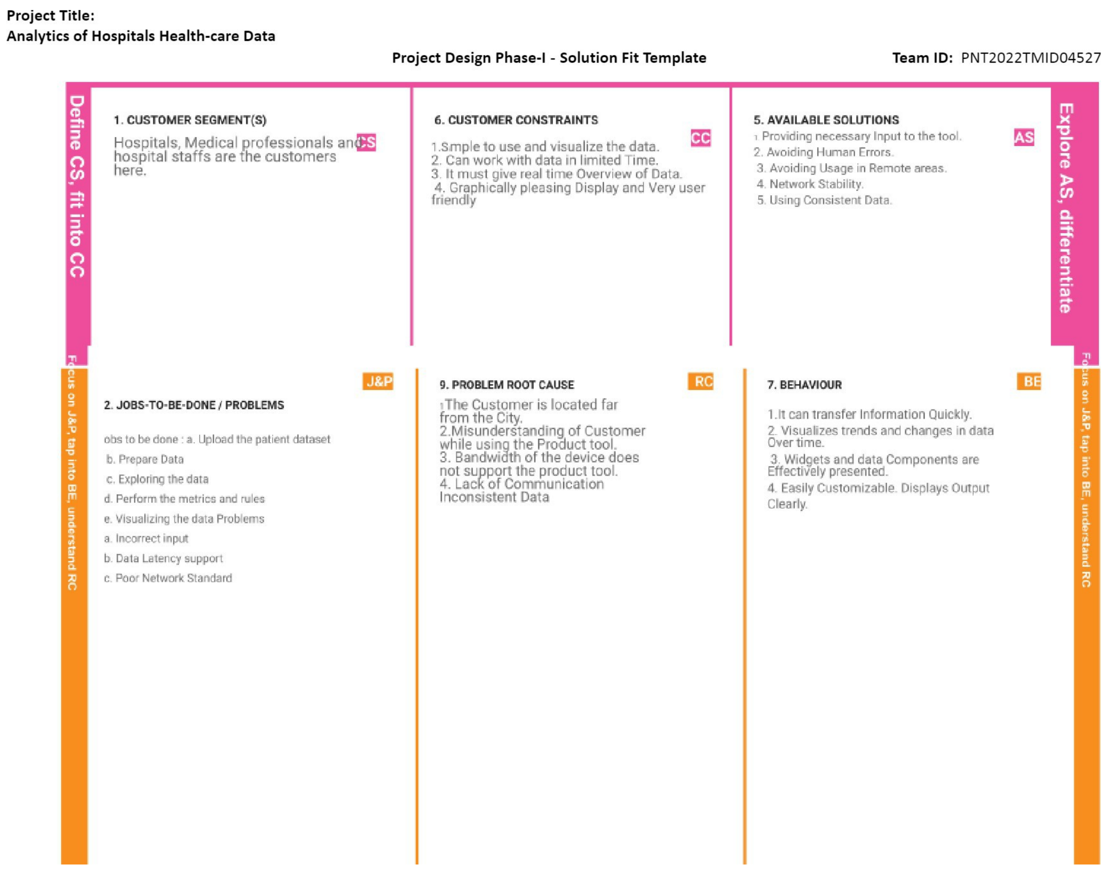
Regardless of their importance, which tasks are more feasible than others? (Cost, time, effort, complexity, etc.)

3.3Proposed solution

S.No.	Parameter	Description
1.	Problem Statement (Problem to besolved)	The task is to accurately predict the Length of Stay for each patient on case-by- case basisso that the Hospitals can use this informationfor optimal resource allocation and betterfunctioning. 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 anassumptionon 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 theLOS of inpatients
4.	Social Impact / CustomerSatisfaction	A shorter LOS reduces the risk of acquiring staph infections and other healthcare-relatedconditions, 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



Project Title:
Analytics of Hospitals Health-care Data

Project Design Phase-I - Solution Fit Template

Team ID: PNT2022TMID04527

Define CS, fit into CC

1. CUSTOMER SEGMENT(S) CS

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

6. CUSTOMER CONSTRAINTS CC

1.Smple to use and visualize the data.
2. Can work with data in limited Time.
3. It must give real time Overview of Data.
4. Graphically pleasing Display and Very user friendly

5. AVAILABLE SOLUTIONS AS

1. Providing necessary Input to the tool.
2. Avoiding Human Errors.
3. Avoiding Usage in Remote areas.
4. Network Stability.
5. Using Consistent Data.

Explore AS, differentiate

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

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

obs to be done : a. Upload the patient dataset
b. Prepare Data
c. Exploring the data
d. Perform the metrics and rules
e. Visualizing the data Problems
a. Incorrect input
b. Data Latency support
c. Poor Network Standard

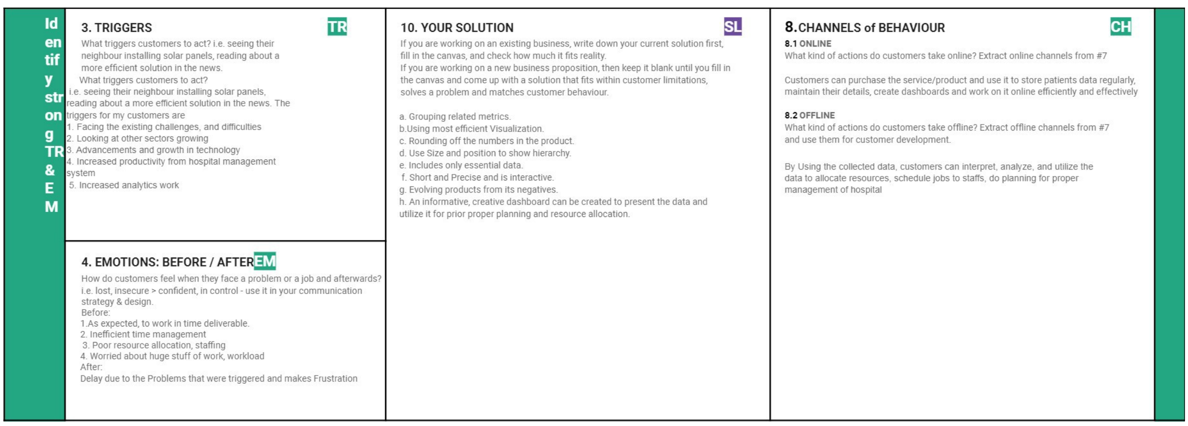
9. PROBLEM ROOT CAUSE RC

1The Customer is located far from the City.
2.Misunderstanding of Customer while using the Product tool.
3. Bandwidth of the device does not support the product tool.
4. Lack of Communication Inconsistent Data

7. BEHAVIOUR BE

1.It can transfer Information Quickly.
2. Visualizes trends and changes in data Over time.
3. Widgets and data Components are Effectively presented.
4. Easily Customizable. Displays Output Clearly.

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.
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

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.As expected, to work in time deliverable.
2. Inefficient time management
3. Poor resource allocation, staffing
4. Worried about huge stuff of work, workload
After:
Delay due to the Problems that were triggered and makes Frustration

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.

a. Grouping related metrics.
b.Using most efficient Visualization.
c. Rounding off the numbers in the product.
d. Use Size and position to show hierarchy.
e. Includes only essential data.
f. Short and Precise and is interactive.
g. Evolving products from its negatives.
h. 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

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

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

3 Requirements analysis

Functional requirements

FR No.	Functional Requirement (Epic)	Sub Requirement (Story / Sub-Task)
FR- 1	**User Registration**	Registration through Form Registration through Gmail Registration through LinkedIN
FR- 2	**User Confirmation**	Confirmation via EmailConfirmation via OTP
FR- 3	**Operability**	Share patient data and make it interoperable among themanagement
FR- 4	**Accuracy**	The dashboard will be able to predict length of stay based on multiple combinations based on input sourceswith a n accuracy of upto 85%
FR- 5	**Compliance**	The product is to be used within the hospital so any formof data need not be hidden
FR- 6	**Productivity**	The dashboard is believed to improve the predictions ofLength of Stay and thereby creating a scenario of providing better solution

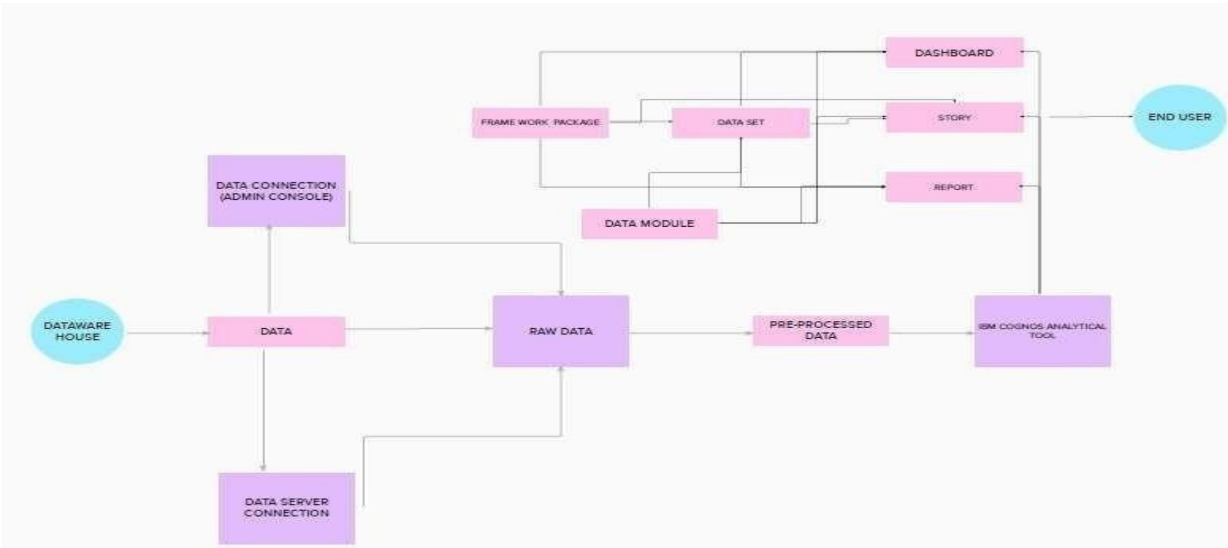
4.2.Nonfunctional requirements

Following are the non-functional requirements of the proposed solution.

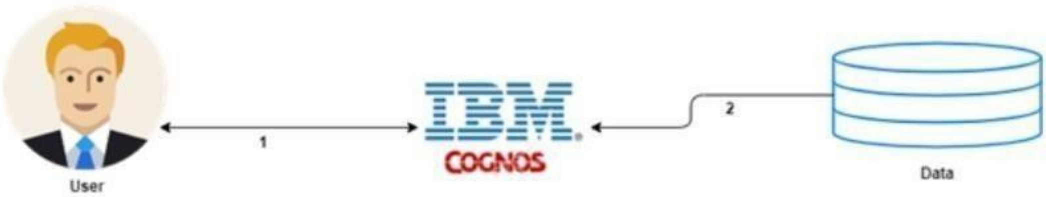
FR No.	Non-Functional Requirement	Description
NF R-1	**Usability**	This Dashboards are designed to offer a comprehensive overview of patient’s LOS, and doso through the use of data visualization tools like charts and graphs.
NF R-2	**Security**	General industry level security shall be provided
NF R-3	**Reliability**	This dashboard will be consistent and reliable to the users and helps the user to use in effective, efficientand 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.

PROJECT DESIGN
3.3 Data Flow Diagrams

The classic visual representation of how information moves through a system is a data flow diagram (DFD). The appropriate amount of the system need can be graphically represented by a clean and unambiguous DFD. It demonstrates how information enters and exits the system, what modifies the data, and where information is kept.



Solution & Technical Architecture



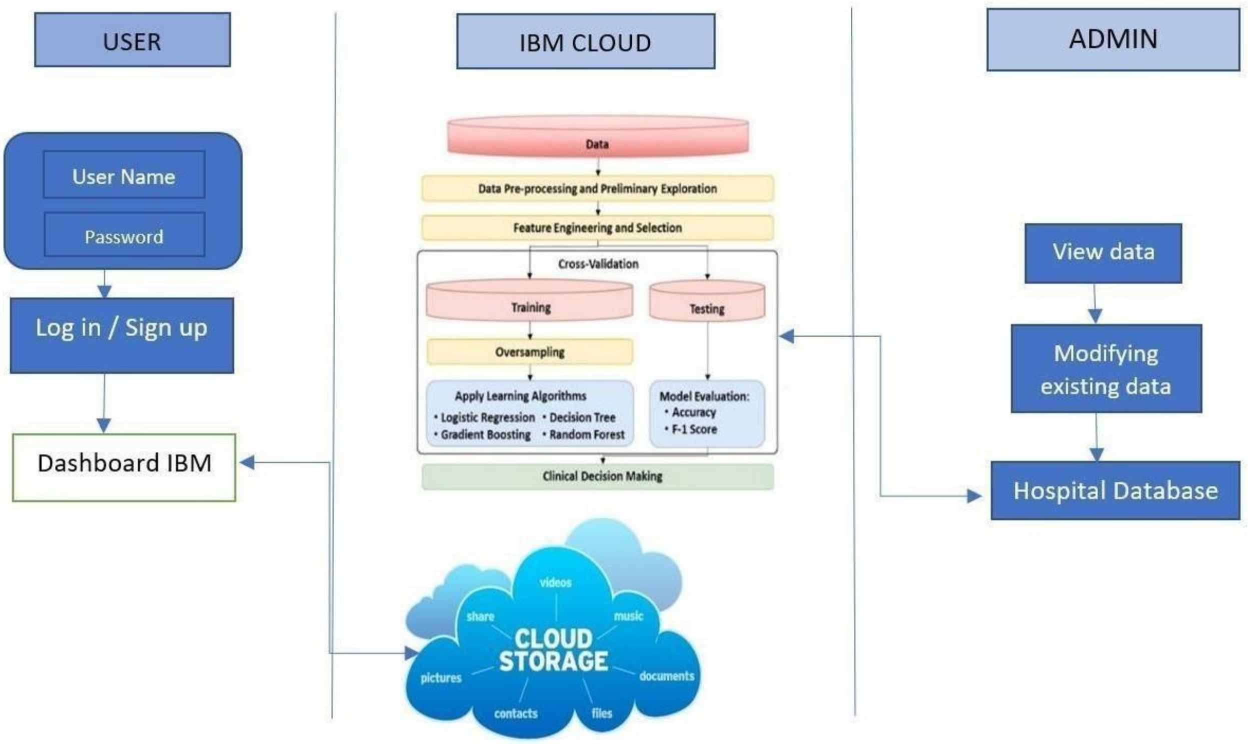


Table-1 : Components & Technologies:

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	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, addressand	MySQL, NoSQL, etc.
6.	Cloud Database	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 Confiu ration,	Local, Cloud Foundry, Kubernetes, etc.

Table-2: Application Characteristics:

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	Justify the scalability of architecture (3 –	Can supports higher workloads
4.	Availability	Justify the availability of application (e.g. use ofload balancers,	Highly available
5.	Performance	Design consideration for the performance of theapplication (number of requests per sec, use of	It performs good uses various tools and ideas in a scientific manner to meet the desired outcomes

User Stories :

Use the below template to list all the user stories for the product.

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
	Predict	USN 4	As an admin, I will predict the length of stay	I can predict the length of stay	High	Sprint-1

6.1. Sprint planning & Estimation

Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority	Team Members
Sprint-1	Analyse	USN-1	As an admin, I will analyse the given dataset (Data pre-processing)	20	High	Santhosh G
Sprint-2	Visualization	USN-2	As a user, I can select the visualization type (Creating visualization)	20	Medium	Sowbarnika P S
Sprint-3	Dashboard	USN-3	As a user, I can upload the datasets to the dashboard and view visualizations (Creating dashboard)	20	Medium	Vidyakeerthi SU
Sprint-4	Predict	USN-4	As an admin, I will predict the length of stay (Prediction)	20	High	Yogapriya S

6.1. Sprint Delivery Schedule

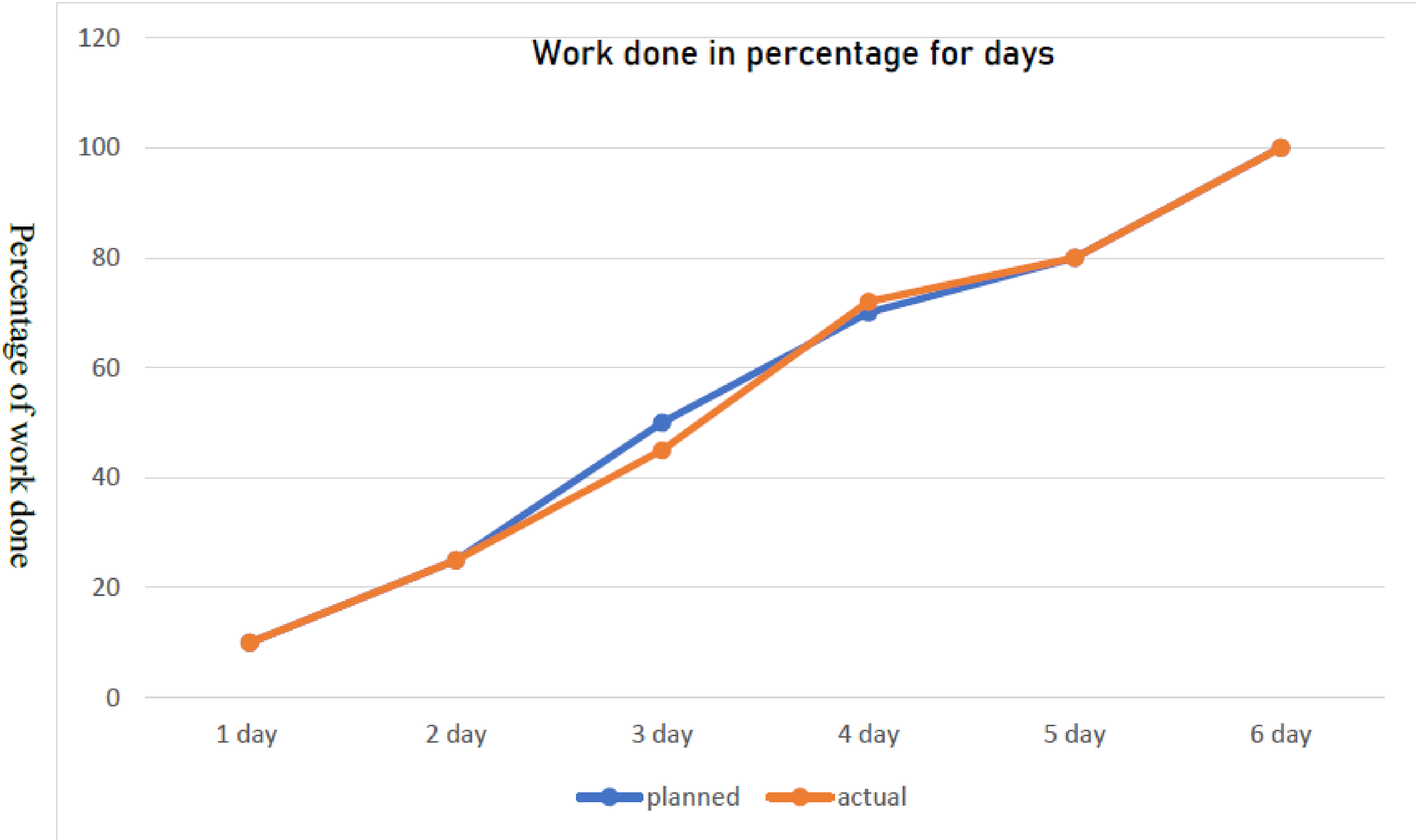
Sprint	Total Story Points	Duration	Sprint Start Date	Sprint End Date (Planned)	Story Points Completed (as on Planned End Date)	Sprint Release Date (Actual)
Sprint-1	20	6 Days	24 Oct 2022	29 Oct 2022	20	29 Oct 2022
Sprint-2	20	6 Days	31 Oct 2022	05 Nov 2022	20	05 Nov 2022
Sprint-3	20	6 Days	07 Nov 2022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

Velocity:

$$AV = \frac{sprint\ duration}{velocity} = \frac{20}{10} = 2$$

Burndown Chart:

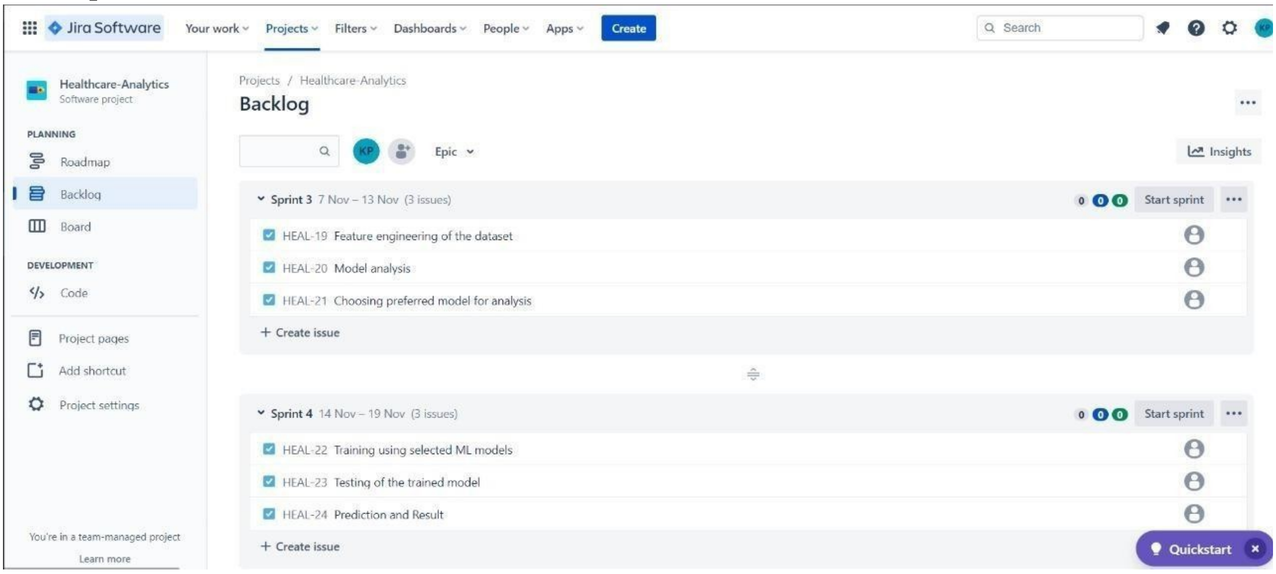
A burn down chart is a graphical representation of work left to do versus time. It is often used in agile software development methodologies such as Scrum. However, burn down charts can be applied to any project containing measurable progress over time.

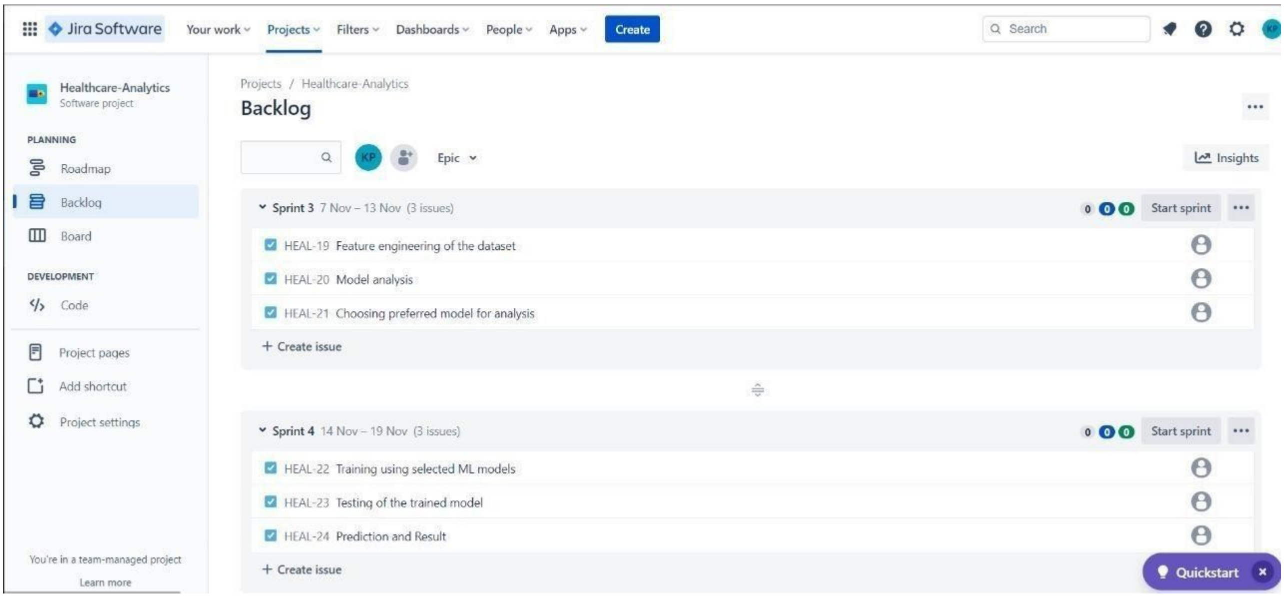


6.1. Reports from JIRA

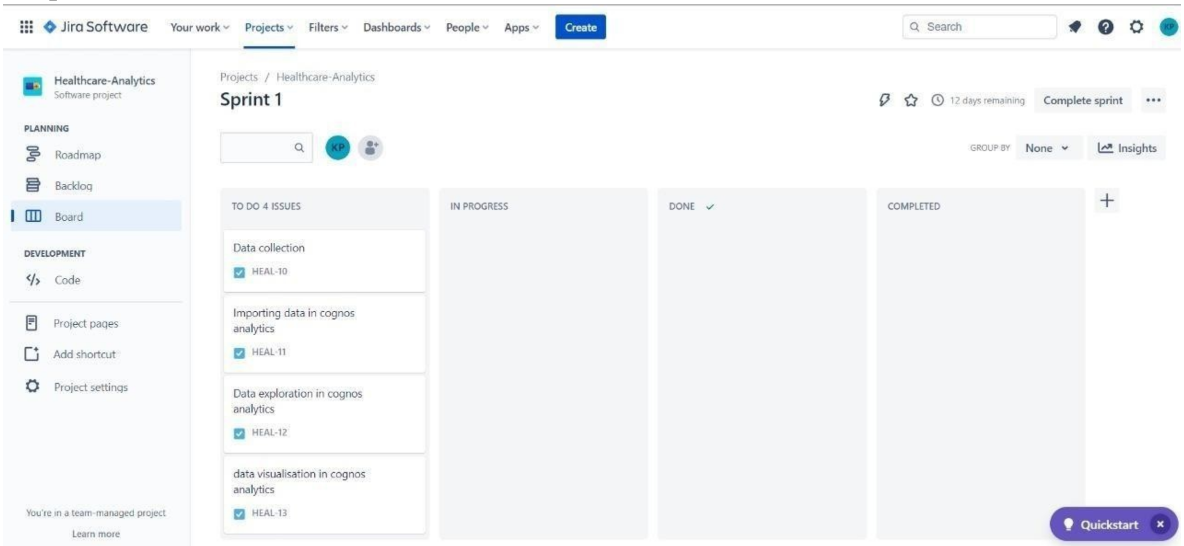
6.2. Reports from JIRA

Jira Sprint



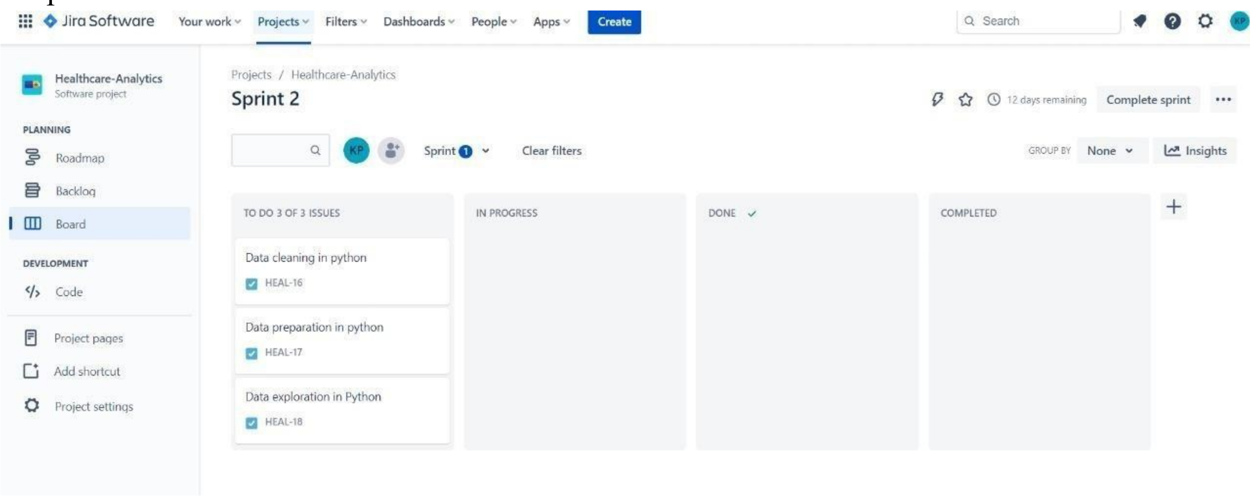


Sprint 1 Dashboard

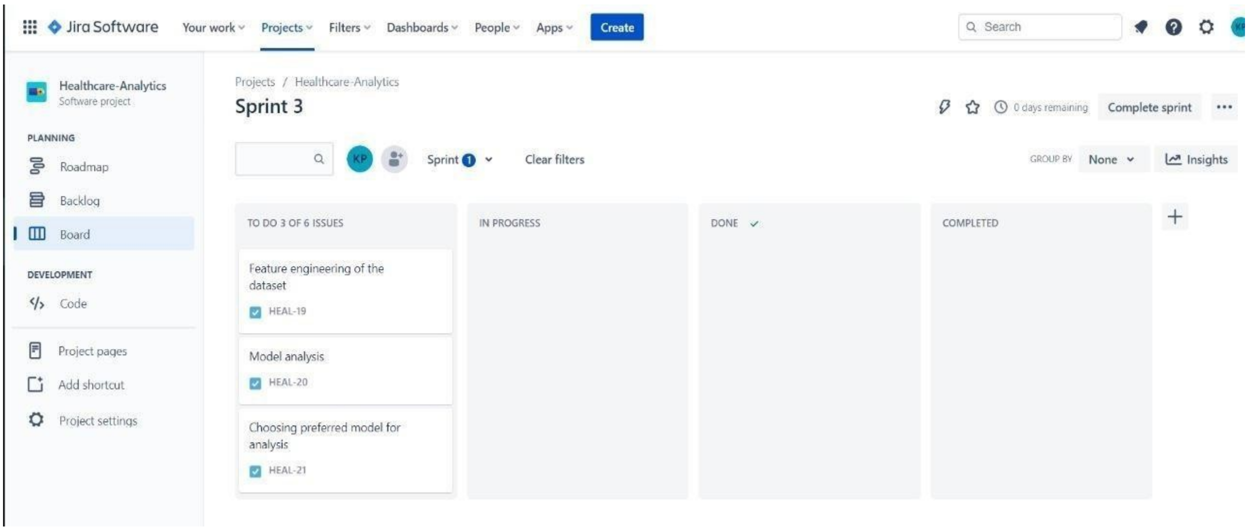


Team id: PNT2022TMID53225

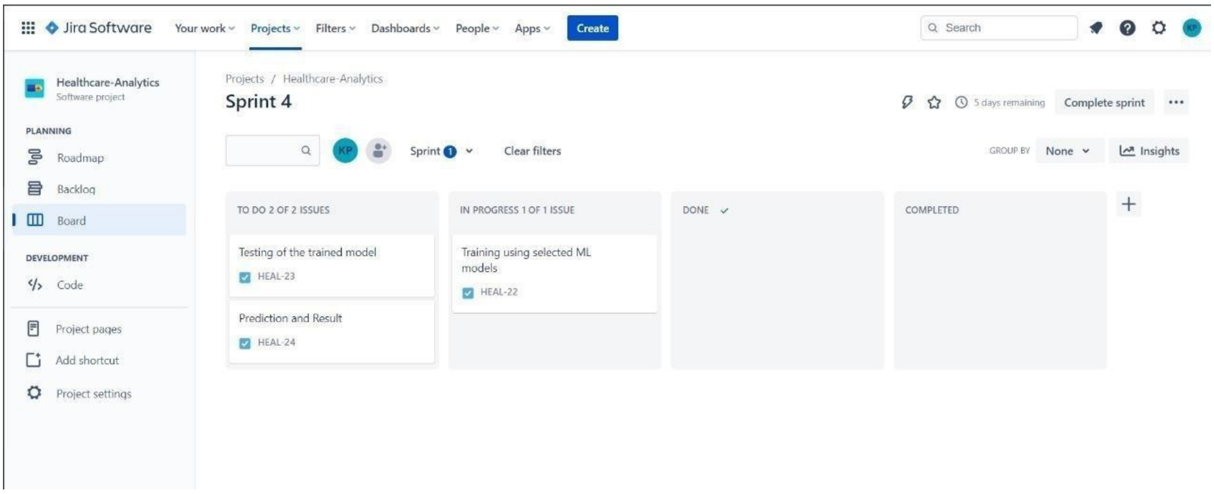
Sprint 2 Dashboard



Sprint 3 Dashboard



Sprint 4 Dashboard



3 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)

$$P(H \mid E) = [P(E \mid H) / P(E)]\, P(H)$$

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.

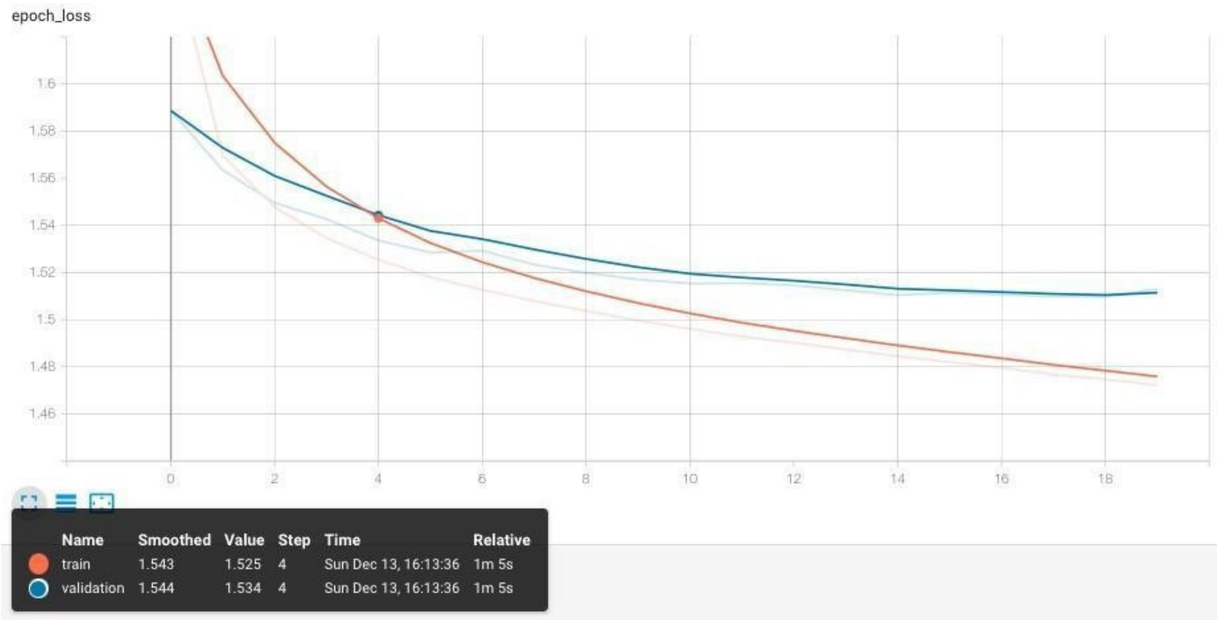
In this model, increasing the number of neurons from each layer to the other layer, will increase the hypothetical space of the model and try to learn more patterns from the data. There are a total of **442,571** trainable parameters. Every layer is activated using "**relu**" activation function because 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 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
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:

Feature engineering:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
sns.set_style("white")
plt.style.use("ggplot")
```

DATA PREPARATION:

```
import os
for dirname, _, filenames in os.walk('/content/Healthcare_Data'):
    for filename in filenames:
        print(os.path.join(dirname, filename))
train = pd.read_csv('/content/Healthcare_Data/train_data.csv')
test = pd.read_csv('/content/Healthcare_Data/test_data.csv')
dictionary =
pd.read_csv('/content/Healthcare_Data/train_data_dictionary.csv') sample =
pd.read_csv('/content/Healthcare_Data/sample_sub.csv')
dictionary
```

DATA EXPLORATION :

```
train.info()
train.tail(5)
plt.figure(figsize=(10,7))
train.Stay.value_counts().plot(kind="barh", color = ['blue'])
train.isnull().sum()
```

DATA PREPROCESSING:

```
train.dropna(inplace=True)
test.dropna(inplace=True)
# Combine test and train dataset for processing
new_set = [train, test]
from sklearn.preprocessing import LabelEncoder
for data in new_set:
    label = LabelEncoder()
    data['Department'] = label.fit_transform(data['Department'])
for dataset in new_set:
    label = LabelEncoder()
    dataset['Hospital_type_code'] = label.fit_transform(dataset['Hospital_type_code'])
```

```
dataset['Ward_Facility_Code'] =
label.fit_transform(dataset['Ward_Facility_Code']) dataset['Ward_Type'] =
label.fit_transform(dataset['Ward_Type'])
dataset['Type of Admission'] = label.fit_transform(dataset['Type of Admission'])
dataset['Severity of Illness'] = label.fit_transform(dataset['Severity of Illness'])
new_set[0]
new_set[0].Age.hist()
new_set[0].Age.unique()
age_dict = {'0-10': 0, '11-20': 1, '21-30': 2, '31-40': 3, '41-50': 4, '51-60': 5, '61-70': 6, '71-80': 7, '81-90': 8, '91-100': 9}
for dataset in new_set:
    dataset['Age'] = dataset['Age'].replace(age_dict.keys(), age_dict.values())
new_set[0].Age.hist()
columns_list = ['Type of Admission', 'Available Extra Rooms in Hospital', 'Visitors with Patient', 'Admission_Deposit']
len(columns_list)
from sklearn.preprocessing import StandardScaler
s1= StandardScaler()
for dataset in new_set:
    dataset[columns_list]= s1.fit_transform(dataset[columns_list].values)
plt.figure(figsize=(17,17))
sns.heatmap(new_set[0].corr(), annot=True, cmap='Greens')
```

DATA MODELLING

```
from sklearn.linear_model import LogisticRegression
from sklearn.svm import SVC, LinearSVC
from sklearn.ensemble import RandomForestClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.tree import DecisionTreeClassifier
train = new_set[0]
test = new_set[1]
sample
X_train = train.drop(['case_id', 'Stay','Hospital_region_code'], axis=1)
Y_train = train["Stay"]
X_test = test.drop(['case_id','Hospital_region_code'], axis=1).copy()
X_train.shape, Y_train.shape, X_test.shape
X_train = X_train.astype(int)
Y_train = Y_train.astype(int)
X_test = X_test.astype(int)

X_test.columns
# Accuracy while using KNN
knn = KNeighborsClassifier(n_neighbors = 3)
```

```
knn.fit(X_train, Y_train)
Y_pred = knn.predict(X_test)
knn_accuracy = round(knn.score(X_train, Y_train) * 100,
2) print("Accuracy of KNN ")
knn_accuracy
# Accuracy while using Decision Tree
decision_tree =
DecisionTreeClassifier()
decision_tree.fit(X_train, Y_train)
Y_pred = decision_tree.predict(X_test)
decision_tree_accuracy = round(decision_tree.score(X_train, Y_train) * 100,
2) print("Accuracy of Decision Tree ")
decision_tree_accuracy

# Accuracy which using Random Forest
random_forest = RandomForestClassifier(n_estimators=100)
random_forest.fit(X_train, Y_train)
Y_pred = random_forest.predict(X_test)
random_forest.score(X_train, Y_train)
random_forest_accuracy = round(random_forest.score(X_train, Y_train) * 100,
2) print("Accuracy of Random Forest ")
random_forest_accuracy

  sns.barplot(x= ['KNN','Decision Tree','Random Forest'],y= [knn_accuracy,
  decision_tree_accuracy, random_forest_accuracy],color = 'orange')
```

RESULT

```
LOS_predicted = pd.DataFrame({
      "case_id": test["case_id"],
      "Stay": Y_pred
})

LOS_predicted['Stay'] = LOS_predicted['Stay'].replace(stay_dict.values(), stay_dict.keys())
LOS_predicted.to_csv('LOS.csv', index = False)

LOS = pd.read_csv('/content/LOS.csv')
 LOS.info()

plt.figure(figsize=(10,5))
LOS.Stay.value_counts().plot(kind="bar", color =
['blue'])
```

GitHub link: https://github.com/IBM-EPBL/IBM-Project-17670-1659675014