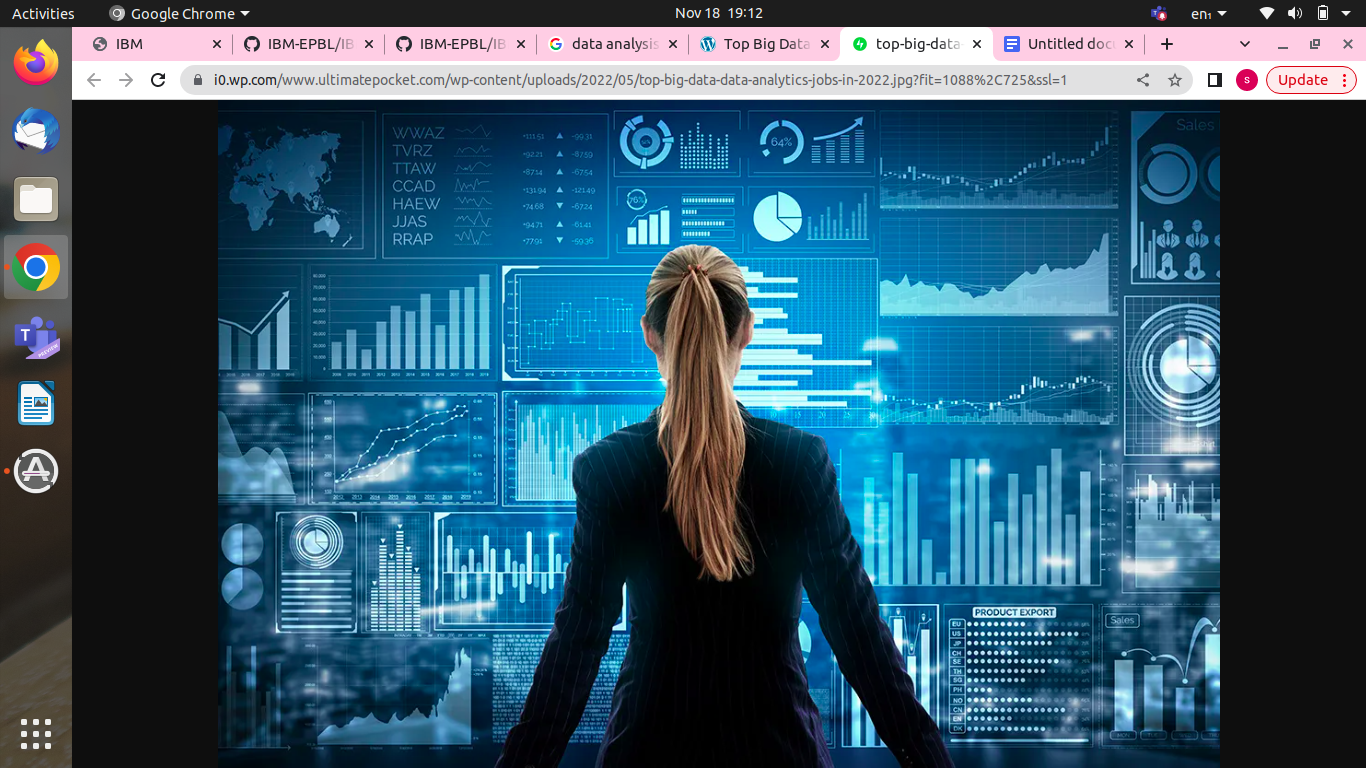
[IBM-Project-42480-1660664304](https://github.com/IBM-EPBL/IBM-Project-42480-1660664304)

**🩺Analytics for Hospitals' Health-Care Data**



**INTRODUCTION**

* 1. **Project Overview**

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.

This parameter helps hospitals to identify patients of high LOS-risk (patients who will stay longer) at the time of admission. Once identified, patients with high LOS risk can have their treatment plan optimised to minimise 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.

**1.2 Purpose**

The goal is to accurately predict the Length of Stay for each patient on a 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.

**2. LITERATURE SURVEY**

2.1 Existing problem

Lack of information

Lack of portability of EHRs

cause of inaccurate data is manual errors made during data entry

2.2 References

[Literature survey](https://github.com/IBM-EPBL/IBM-Project-42480-1660664304/blob/main/pre-development/Ideation%20phase/Liteture%20survey.pdf)

2.3 Problem Statement Definition

Human heart is the principal part of the human body. Basically, it regulates blood flow throughout our body. Any irregularity to the heart can cause distress in other parts of the body. Any sort of disturbance to normal functioning of the heart can be classified as a Heart disease. In today's contemporary world, heart disease is one of the primary reasons for occurrence of most deaths. Heart disease may occur due to unhealthy lifestyle, smoking, alcohol and high intake of fat which may cause hypertension .The whole accuracy in management of a disease lies on the proper time of detection of that disease. Descriptive analytics is the process of using current and historical data to identify trends and relationships,that is using python coding .The proposed work makes an attempt to detect these heart diseases at an early stage to avoid disastrous consequences.

**3. IDEATION & PROPOSED SOLUTION**

**3.1** Empathy Map Canvas

[Empathy Map](https://github.com/IBM-EPBL/IBM-Project-42480-1660664304/blob/main/pre-development/Ideation%20phase/empathy%20map.pdf)

3.2 Ideation & Brainstorming

[Brainstorming](https://github.com/IBM-EPBL/IBM-Project-42480-1660664304/blob/main/pre-development/Ideation%20phase/brainstroming.pdf)

3.3 Proposed Solution

[Proposed solution](https://github.com/IBM-EPBL/IBM-Project-42480-1660664304/blob/main/pre-development/Project%20phase%201/proposed%20solution.pdf)

3.4 Problem Solution fit

[solution fit](https://github.com/IBM-EPBL/IBM-Project-42480-1660664304/blob/main/pre-development/Project%20phase%201/Problem_solution_fit%20Sample%20Template%20(1)%20(1).pdf)

**4. REQUIREMENT ANALYSIS**

4.1 Functional requirement

[Functional requirement](https://github.com/IBM-EPBL/IBM-Project-42480-1660664304/blob/main/pre-development/project%20phase%202/Soultion%20Requirements.pdf)

4.2 Non-Functional requirements

[Non-functional requirements](https://github.com/IBM-EPBL/IBM-Project-42480-1660664304/blob/main/pre-development/project%20phase%202/Soultion%20Requirements.pdf)

**5. PROJECT DESIGN**

5.1 Data Flow Diagrams

[Data flow](https://github.com/IBM-EPBL/IBM-Project-42480-1660664304/blob/main/pre-development/project%20phase%202/Data%20Flow%20Diagram%20%26%20User%20Stories.pdf)

5.2 Solution & Technical Architecture

[Technical Architecture](https://github.com/IBM-EPBL/IBM-Project-42480-1660664304/blob/main/pre-development/project%20phase%202/Technology%20Artitecture.pdf)

[solution requirements](https://github.com/IBM-EPBL/IBM-Project-42480-1660664304/blob/main/pre-development/Project%20phase%201/Solution%20Architeture.pdf)

5.3 User Stories

[user stories](https://github.com/IBM-EPBL/IBM-Project-42480-1660664304/blob/main/pre-development/project%20phase%202/Data%20Flow%20Diagram%20%26%20User%20Stories.pdf)

6. **PROJECT PLANNING & SCHEDULING**

6.1 Sprint Planning & Estimation

[Journey map](https://github.com/IBM-EPBL/IBM-Project-42480-1660664304/blob/main/pre-development/project%20phase%202/customer%20journey%20map.pdf)

6.2 Sprint Delivery Schedule

[Delivery plan](https://github.com/IBM-EPBL/IBM-Project-42480-1660664304/blob/main/pre-development/project%20planning/_Sprint%20Delivery.pdf)

6.3 Reports from JIRA

[JIRA](https://github.com/IBM-EPBL/IBM-Project-42480-1660664304/blob/main/pre-development/project%20planning/_Sprint%20Delivery.pdf)

**7. CODING & SOLUTIONING (Explain the features added in the project along with code)**

7.1 Feature 1

TECHNICAL COMPONENTS USED

[IBM Cloud](https://en.wikipedia.org/wiki/IBM_Cloud)

[IBM Cognos Analytics](https://en.wikipedia.org/wiki/IBM_Cognos_Analytics)

[Data Analysis with Python](https://en.wikipedia.org/wiki/Data_analysis)

7.2 Feature 2

[google colab](https://colab.research.google.com/)

[Jupyter notebook](https://jupyter.org/)

7.3 Database Schema

[Kaggle](https://www.kaggle.com/)

**8. RESULTS**

8.1 Performance Metrics

NLP in healthcare media can accurately give voice to the unstructured data of the healthcare universe, giving incredible insight into understanding quality, improving methods, and better results for patients. NLP models have helped leading hospitals within India and abroad, overhaul their patient and staff experience through use cases like automation of appointment booking, feedback collection, optimization of internal processes like medical coding and data assessment as well as data entry.

**9. ADVANTAGES & DISADVANTAGES**

| **ADVANTAGE** | **DISADVANTAGE** |
| --- | --- |
| To successfully identify and implement big data solutions and benefit from the value that big data can bring, the government needs to devote time, allocate budget and resources to visioning and planning.  The problem is not the lack of data but the lack of information that can be used to support decision-making,  planning and state The problem is not the lack of data but the lack of information that can be used to support decision-making, planning and strategy | The problem is not the lack of data but the lack of information that can be used to support decision making, planning and strategy |
| This offers huge advantage  that had not been previously  possible for a more personalised approach to treating T2D that  will be safer and more beneficial for the patient as it will minimise side effects and offer faster, more effective treatment. It will also provide economic advantages  to the healthcare system. | There is need to Building and training  the model on larger databases to  increase the prediction accuracy and  develop more robust prediction models  are achieve effectively. |
| Research and prediction of  disease. Automation of  hospital administrative processes Early detection of disease. Prevention of unnecessary doctor's visits. Discovery of new drugs. More accurate calculation of health  insurance rates. More effective sharing of patient  data | Lack of standardisation  in toxicology and  coding practices among medical examiners and coroners can lead to misclassification of  cause of death, poor identification of types  of opioids involved in overdoses, and  undercounting of  intentional poisonings**.** |
| The paper has listed  some data analytics  tools and techniques  that have been used to improve healthcare performance in  many areas such as: medical operations, reports,  decision making, and prediction and prevention system | The problem is how to  handle this with older  people who are less  attached and  hard to convince to  adopt new healthcare technologies and tools,  as they consider this as  a medical care  issue involving medical staff and excluding their role in the medical care process. |
| Real-world data also helps researchers who are interested in less common conditions that aren't as likely to be studied in clinical trials. With access to thousands of patients' data,  lack of clinical trials becomes less of a barrier for researchers interested in rare diseases. | Limitations of RWE  studies can include low internal validity, lack of  quality control  surrounding data  collection and  susceptibility to  multiple sources of bias for comparing outcomes**.** |
| The advancement of technology  and other factors are compelling healthcare providers to adopt  advanced communication and collaboration systems across their settings. | The big question in front  of these healthcare  organisations is how to  crunch these numbers  and extract meaningful knowledge from health  Big Data, identify and  develop new decision  models and how to  manage Big Data |
| One advantage of Cox models is that there is no re-training needed if we change the time of interest (from 30 days to 90 days) | Adding claims data for  a partial set of patient |
| Machine learning presents  enormous opportunity within the healthcare industry to reduce inefficiency and costs while increasing the quality and accuracy of patient care | The business people,  it's often a challenge just  to communicate the  clinical side in a way that  doesn't overwhelm them. But it is a little bit of an art. |
| Big data is characterised as  extremely large data sets that  can be analysed  computationally to find patterns, trends, and associations,  visualisation, querying,  information privacy and  predictive analytics on  large wide spread collection  of data | There is a lack of  portability of EHRs to  all over the country or  world for better  treatment anywhere  anytime without  carrying past treatment  record of individual |
| We highlighted the shortcoming of the existing Big Data analytics tools in dealing with the evolution of data. The proposed Import Big Data storage is a promising solution for dealing the heterogeneous health data. | In terms of better query performance and  scalability in distributed systems. The proposed prototype will compare the scalability of the proposed framework with the other platform. |

**10. CONCLUSION**

The healthcare ecosystem is plagued with a number of challenges including prescription.Additionally, healthcare spending, waste, abuse and fraud are at an all-time high.

There are several initiatives at the federal, state, and local levels to combat these challenges. All of these initiatives require large data storage, analytics and deep insights, in real time to accelerate decision making for preventive measures. Architecture to help unlock the potential for gaining deep insights into heart disease prediction using AI/ML capabilities.

**11. FUTURE SCOPE**

NLP in healthcare media can accurately give voice to the unstructured data of the healthcare universe, giving incredible insight into understanding quality, improving methods, and better results for patients.

**12. APPENDIX**

Source Code

import sklearn

import numpy as np

import pandas as pd

import plotly as plot

import plotly.express as px

import plotly.graph\_objs as go

import cufflinks as cf

import matplotlib.pyplot as plt

import seaborn as sns

import os

from sklearn.metrics import accuracy\_score

import plotly.offline as pyo

from plotly.offline import init\_notebook\_mode,plot,iplot

pyo.init\_notebook\_mode(connected=True)

cf.go\_offline()

heart=pd.read\_csv(r'E:\DS\Heart-Disease\heart.csv')

heart

for i in range(len(info)):

print(heart.columns[i]+":\t\t\t"+info[i])

heart['target']

heart.groupby('target').size()

heart.groupby('target').sum()

heart.shape

heart.size

heart.describe()

heart.info()

heart['target'].unique()

heart.hist(figsize=(14,14))

plt.show()

plt.bar(x=heart['sex'],height=heart['age'])

plt.show()

sns.barplot(x="fbs", y="target", data=heart)

plt.show()

sns.barplot(heart["cp"],heart['target'])

sns.barplot(heart["sex"],heart['target'])

px.bar(heart,heart['sex'],heart['target'])

sns.distplot(heart["thal"])

sns.distplot(heart["chol"])

sns.pairplot(heart,hue='target')

numeric\_columns=['trestbps','chol','thalach','age','oldpeak']

heart['target']

y = heart["target"]

sns.countplot(y)

target\_temp = heart.target.value\_counts()

print(target\_temp)

# create a correlation heatmap

sns.heatmap(heart[numeric\_columns].corr(),annot=True, cmap='terrain', linewidths=0.1)

fig=plt.gcf()

fig.set\_size\_inches(8,6)

plt.show()

# create four distplots

plt.figure(figsize=(12,10))

plt.subplot(221)

sns.distplot(heart[heart['target']==0].age)

plt.title('Age of patients without heart disease')

plt.subplot(222)

sns.distplot(heart[heart['target']==1].age)

plt.title('Age of patients with heart disease')

plt.subplot(223)

sns.distplot(heart[heart['target']==0].thalach )

plt.title('Max heart rate of patients without heart disease')

plt.subplot(224)

sns.distplot(heart[heart['target']==1].thalach )

plt.title('Max heart rate of patients with heart disease')

plt.show()

plt.figure(figsize=(13,6))

plt.subplot(121)

sns.violinplot(x="target", y="thalach", data=heart, inner=None)

sns.swarmplot(x="target", y="thalach", data=heart, color='w', alpha=0.5)

plt.subplot(122)

sns.swarmplot(x="target", y="thalach", data=heart)

plt.show()

# create pairplot and two barplots

plt.figure(figsize=(16,6))

plt.subplot(131)

sns.pointplot(x="sex", y="target", hue='cp', data=heart)

plt.legend(['male = 1', 'female = 0'])

plt.subplot(132)

sns.barplot(x="exang", y="target", data=heart)

plt.legend(['yes = 1', 'no = 0'])

plt.subplot(133)

sns.countplot(x="slope", hue='target', data=heart)

plt.show()

heart['target'].value\_counts()

heart['target'].sum()

heart['target'].unique()

heart.isnull()

X,y=heart.loc[:,:'thal'],heart.loc[:,'target']

X

X.shape

y.shape

from sklearn.model\_selection import train\_test\_split

from sklearn.preprocessing import StandardScaler

X=heart.drop(['target'],axis=1)

X

X\_train,X\_test,y\_train,y\_test=train\_test\_split(X,y,random\_state=10,test\_size=0.3,shuffle=True)

X\_test

y\_test

print ("train\_set\_x shape: " + str(X\_train.shape))

print ("train\_set\_y shape: " + str(y\_train.shape))

print ("test\_set\_x shape: " + str(X\_test.shape))

print ("test\_set\_y shape: " + str(y\_test.shape))

art Disease or at leaset Corona Virus Soon...','Yes you have Heart Disease....RIP in Advance']

from sklearn.tree import DecisionTreeClassifier

dt=DecisionTreeClassifier()

dt.fit(X\_train,y\_train)

prediction=dt.predict(X\_test)

accuracy\_dt=accuracy\_score(y\_test,prediction)\*100

accuracy\_dt

print("Accuracy on training set: {:.3f}".format(dt.score(X\_train, y\_train)))

print("Accuracy on test set: {:.3f}".format(dt.score(X\_test, y\_test)))

X\_DT=np.array([[63 ,1, 3,145,233,1,0,150,0,2.3,0,0,1]])

X\_DT\_prediction=dt.predict(X\_DT)

X\_DT\_prediction[0]

print(Catagory[int(X\_DT\_prediction[0])])

print("Feature importances:\n{}".format(dt.feature\_importances\_))

def plot\_feature\_importances\_diabetes(model):

plt.figure(figsize=(8,6))

n\_features = 13

plt.barh(range(n\_features), model.feature\_importances\_, align='center')

plt.yticks(np.arange(n\_features), X)

plt.xlabel("Feature importance")

plt.ylabel("Feature")

plt.ylim(-1, n\_features)

plot\_feature\_importances\_diabetes(dt)

plt.savefig('feature\_importance')

sc=StandardScaler().fit(X\_train)

X\_train\_std=sc.transform(X\_train)

X\_test\_std=sc.transform(X\_test)

X\_test\_std

from sklearn.neighbors import KNeighborsClassifier

knn=KNeighborsClassifier(n\_neighbors=4)

knn.fit(X\_train\_std,y\_train)

prediction\_knn=knn.predict(X\_test\_std)

accuracy\_knn=accuracy\_score(y\_test,prediction\_knn)\*100

print("Accuracy on training set: {:.3f}".format(knn.score(X\_train, y\_train)))

print("Accuracy on test set: {:.3f}".format(knn.score(X\_test, y\_test)))

k\_range=range(1,26)

scores={}

scores\_list=[]

for k in k\_range:

knn=KNeighborsClassifier(n\_neighbors=k)

knn.fit(X\_train\_std,y\_train)

prediction\_knn=knn.predict(X\_test\_std)

scores[k]=accuracy\_score(y\_test,prediction\_knn)

scores\_list.append(accuracy\_score(y\_test,prediction\_knn))

scores

plt.plot(k\_range,scores\_list)

px.line(x=k\_range,y=scores\_list)

X\_knn=np.array([[63 ,1, 3,145,233,1,0,150,0,2.3,0,0,1]])

X\_knn\_std=sc.transform(X\_knn)

X\_knn\_prediction=dt.predict(X\_knn)

X\_knn\_std

(X\_knn\_prediction[0])

print(Catagory[int(X\_knn\_prediction[0])])

algorithms=['Decision Tree','KNN']

scores=[accuracy\_dt,accuracy\_knn]

sns.set(rc={'figure.figsize':(15,7)})

plt.xlabel("Algorithms")

plt.ylabel("Accuracy score")

sns.barplot(algorithms,scores)

**Project Contributors**

Batch Name : B8-2A4E

Team Id : PNT2022TMID14072

[@swathiga.g.v](https://github.com/swathigagv#:~:text=swathigagv,-Edit%20profile)

[@Tharunkumar.R](https://github.com/tharun9160#:~:text=Follow-,tharun9160,-Follow)

[@Thambu ganesh](https://github.com/544ganesh#:~:text=Follow-,544ganesh,-Follow)

@tharani

**💻GITUP & PROJECT DEMO LINK**

**Video link:**[**https://drive.google.com/file/d/12hpCbbu049t9-pwIzrcW4YQMkcc8zsJc/view?usp=share\_link**](https://drive.google.com/file/d/12hpCbbu049t9-pwIzrcW4YQMkcc8zsJc/view?usp=share_link)

**Gitup link:**[**https://github.com/IBM-EPBL/IBM-Project-42480-1660664304**](https://github.com/IBM-EPBL/IBM-Project-42480-1660664304)