**BUILD A MODEL TO PREDICATE IF A PASSENGER SURVIVED THE TITANIC DISASTER**

**INTERNSHIP REPORT**

***Submitted by***

# SIRRA SANDEEP

**(ROLL No : 20R15A0513)**

***in partial fulfillment for the award of the degree of***

**BACHELOR OF TECHNOLOGY**

**in**

**COMPUTER SCIENCE AND ENGINEERING**



**GEETHANJALI COLLEGE OF ENGINEERING AND TECHNOLOGY**

Cheeryal (V), Keesara (M), Medchal Dist, Hyderabad– 501 301

(Affiliated to Jawaharlal Nehru Technological University, Hyderabad, Accredited by NAAC and NBA, New Delhi) 2019-2023

## ACKNOWLEDGEMENT

I would like to thank **Smartknower** for giving me the opportunity to do an internship within the organization. I also would like to thank all the people who worked along with me with their patience and openness. It is indeed with a great sense of pleasure and immense sense of gratitude that I acknowledge the help of these individuals.

I am highly indebted to Principal Dr. Udaya Kumar Susarla for the facilities provided to accomplish this internship. I would like to thank Head of the Department Dr.A.Sree Lakshmi for her constructive criticism throughout my internship.

.

SIRRA SANDEEP

Roll No : 20R15A0513

**GEETHANJALI COLLEGE OF ENGINEERING AND TECHNOLOGY**

Cheeryal (V), Keesara (M), Medchal Dist, Hyderabad– 501 301

(Affiliated to Jawaharlal Nehru Technological University, Hyderabad, Accredited by NAAC

and NBA, New Delhi)



## Department of Computer Science and Engineering

**CERTIFICATE**

This is to certify that the internship report titled “Build a model to predicate if a passenger survived the titanic disaster” being submitted by S.Sandeep, bearing roll number 20R15A0513, in partial fulfillment of the requirements for the award of the Degree of Bachelor of Technology in Computer Science and Engineering is a record of Bonafede work carried out under “Smartknower” guidance and supervision.

Examiner Dr. A Sree Lakshmi

Name:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (HoD) Designation:\_\_\_\_\_\_\_\_\_\_\_\_





**TABLE OF CONTENTS**

PAGE NO

|  |  |
| --- | --- |
| 1.Internship Report | 1 |
| 2.Acknowledgement | 2 |
| 3.Certificate | 3 |
| 4.Certificate of Achievement/Internship | 4-5 |
| 5.Introduction about Company | 7 |
| 6.Training Schedule | 8 |
| 7.Abstract | 9 |
| 8.Introduction | 10 |
| 9.Technologies requires | 11 |
| 10.Flow chart | 12 |
| 11.Work done/Observation | 13 – 18 |
| 12.Appendix | 19-21 |
| 13.Summary/Conclusion | ` 22 |
| 14.Applications and Future Scope | 22 |
| 15. Bibliography | 22 |

## Introduction about Company

Smart knower brings academic & industry very close for a common goal of talent creation through experiential learning & development environment. The platform enables students to move in a successful career path to reach the expectations of industry. They are building the Next Generation Talent pool with skills in emerging technologies i.e., Artificial Intelligence,

Data Science, Internet of Things(lot), Robotics, Block chain, Quantum Computing and Cyber

Security. The unique method of smart knower is project-based learning, micro-skilling and Internships helps students in building their competency & get ready for industry. They bring students, educators and employers on a common platform to fill the gap between academia & industry

**Methodology:** Collaborative learning in a project environment with industry interaction & mentoring.

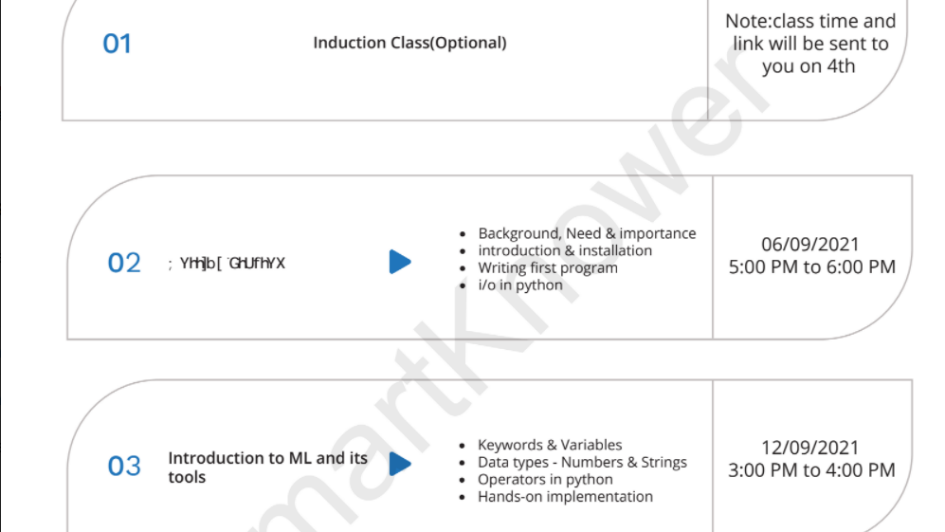
**Learn:** Project based learning environment to understand the practical application of theoretical concepts by deeply engaging the students.

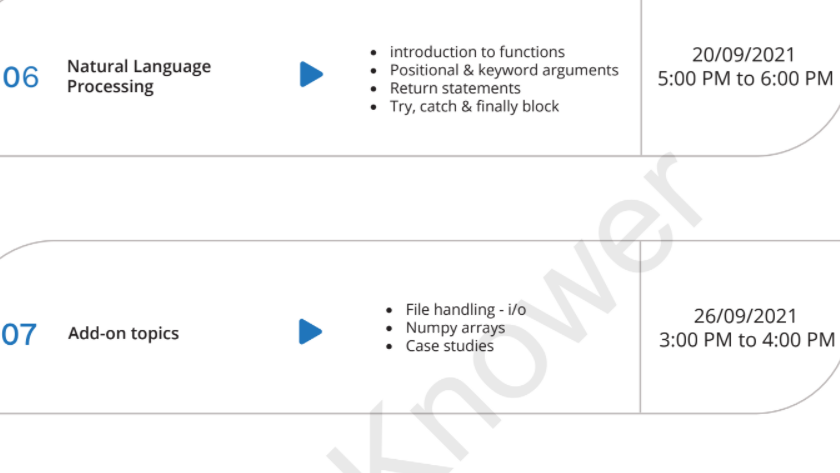
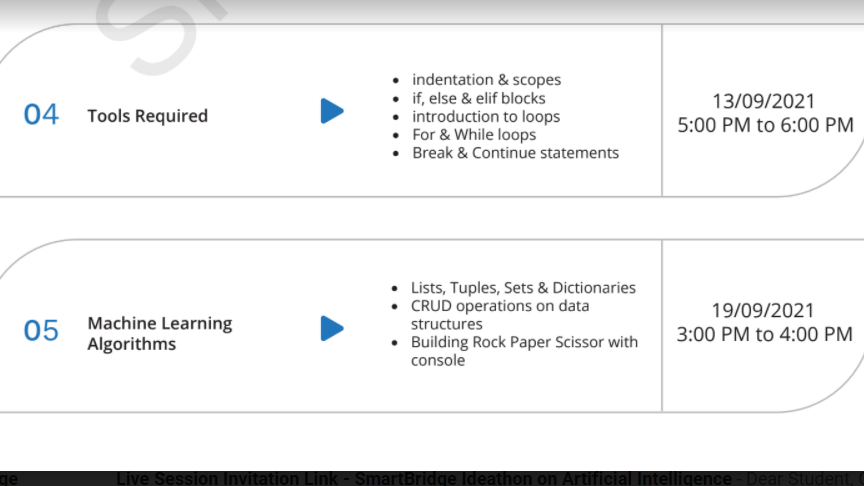
**Practice:** Practice through problem solving and project building can make a student an expert in his field of interest.

**Intern:** Collaborative working environment with industry mentorship to make young talent ready for real world challenges.

### 

## Training Schedule





## Abstract

 Data Science and analysis is playing the most significant role today covering every industry in the market.For e.g finance,e-commerce,business,education,government. Now organizations play a 360 degree role to analyzed the behavior and interest of their customers to take decisions in favour of them.Data is analyzed through programming language such as python which is one of the most versatile language and helps in doing a lot of things through it. Netflix is a pure data science project that reached at the top through analyzing every single interest of their customers. Keyword:Data Visualization, AnacondaJupyter Notebook,Exploratory Data Anlaysis,Machine Learning,Deep learning.

.

### INTRODUCTION

1. The sinking of the RMS Titanic is one of the most infamous shipwrecks in history
2. In this challenge , we ask you to complete the analysis of what sorts of people were likely to survive
3. In particular , we ask you to apply the tools of machine learning to predict which passengers survived the tragedy
4. The purpose of the this project is to document the process I went through to create my predictions for titanic survivor prediction

### TECHNOLOGIES REQUIRED

**Software Requirements:**

In order to develop this project, we need to install the following software/packages:

Tool used:

○ Anaconda Navigator

○ Jupyter notebook

To build Machine learning models you must require the following packages:

○ NumPy

○ pandas

○ sci-kit learn

To make a responsive python script you must require the following packages:

○ Requests

○ Flask

Need to download the required packages:

○ pip install requests

○ pip install Flask

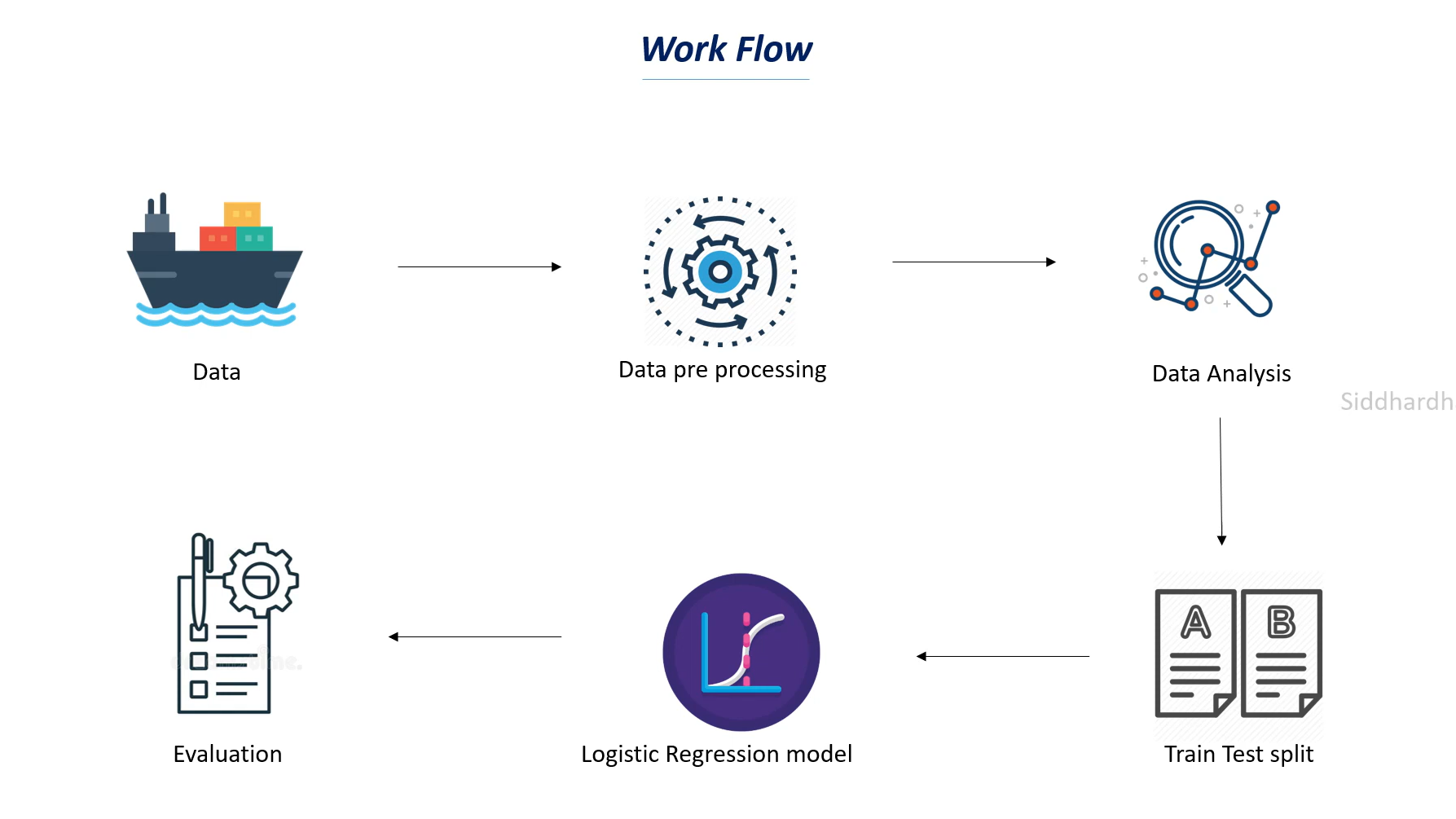
**Hardware Requirements:**

Processor: Intel® Core™ i3-2350M CPU @ 2.30GHz

Installed memory (RAM):4.00GB

System Type: 64-bit Operating System

**FLOW CHART**



**BUILD A MODEL TO PREDICATE IF A PASSENGER SURVIVED THE TITANIC DISASTER**

#### Project Explanation

**1. Data Collection:**

DS depends heavily on data, without data, it is impossible for a machine to learn. It is the most crucial aspect that makes algorithm training possible. In Data science projects, we need a training data set. It is the actual data set used to train the model for performing various actions.

**Collect The Data Set**

The dataset used for this project was obtained from Kaggle.

**2. Data Preprocessing**

We have to train the machine, but before training the machine, make sure that the collected data is cleaned.

Data Pre-processing includes the following main tasks:

**Import The Library:**

The first step is usually importing the libraries that will be needed in the program.

The required libraries to be imported to Python script are:

**Numpy**- It is an open-source numerical Python library. It contains a multi-dimensional array and matrix data structures. It can be used to perform mathematical operations on arrays such as trigonometric, statistical, and algebraic routines.

**Pandas**- It is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool, built on top of the Python programming language.

**Matplotlib**- Visualisation with python. It is a comprehensive library for creating static, animated, and interactive visualizations in Python

All the above modules can be imported into our program using the below code

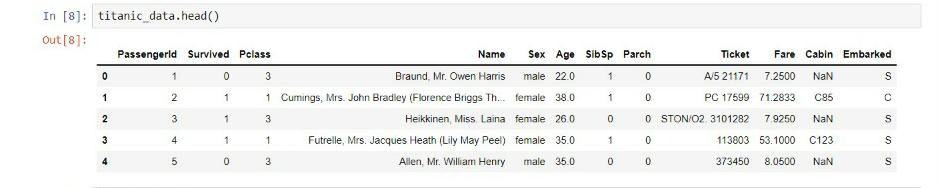


**Importing The Dataset**

You might have your data in .csv files or .excel files

Load the csv data file into pandas using the read\_csv() function. We will need to locate the directory of the csv file at first.





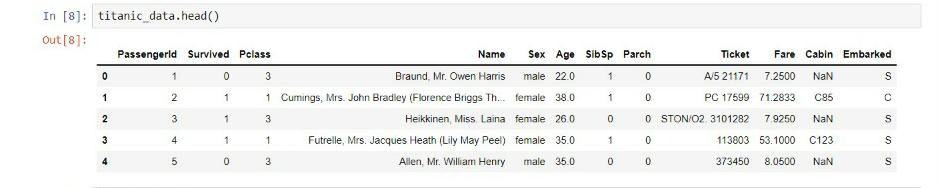
Analyze The Data head() method is used to return top n (5 by default) rows of a DataFrame or series.

tail() method is used to return bottom n (5 by default) rows of a DataFrame or series

dtypes method give the datatypes of the columns present in the dataset

**Pre Processing The Data**

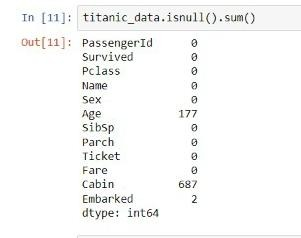
info() gives information about the data



Taking care of Missing Data

After loading the dataset, it is important to check the complete information of such as null values in a column or a row

Check for the null values. if i present then the following tasks can be done,



The 'inplace=True' argument stands for the data frame has to make changes permanent



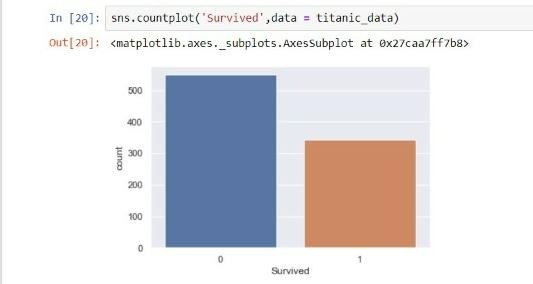
The dataset after naming convention looks like.

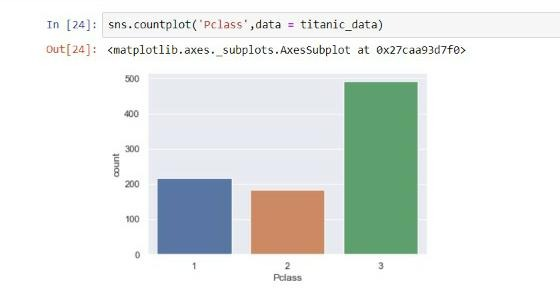
**Data Visualization:**

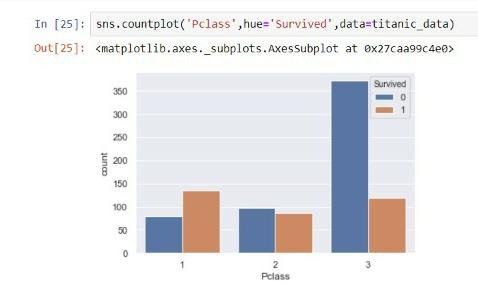
Plotting a time series helps us actually see if there is a trend, a seasonal cycle, outliers, and more.

We can plot the data easily by calling the plot() function on the DataFrame.

To visualize the dataset we need libraries called Matplotlib







**3. Model Building:**

Train , Test And Save The Model

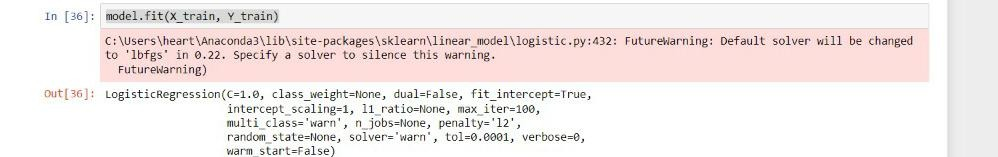
Using logistic regression :

Logical regression is a algorithm that is used to predict the probability of a categorical dependent variable. I the dependent variable is a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure).

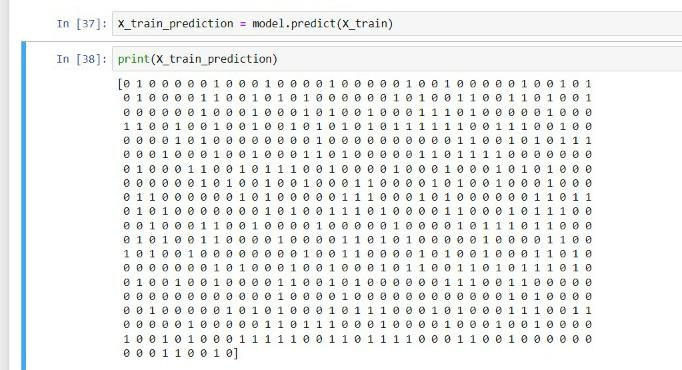


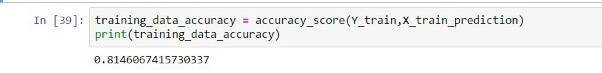
Training the model :

Training a model simply Training a model simply means learning (determining) good values for all the weights and the bias from labeled example.



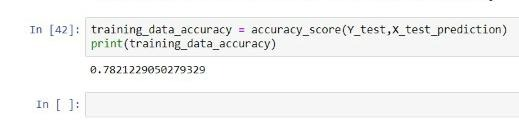
Predicting the train model





**Checking the accuracy**:

Accuracy determines that how percentage of test data is correctly classified.



#### Appendix

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

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

from sklearn.linear\_model import LogisticRegression

from sklearn.metrics import accuracy\_score

titanic\_data = pd.read\_csv('train1.csv')

titanic\_data.head()

titanic\_data.shape

titanic\_data.info()

titanic\_data.isnull().sum()

titanic\_data = titanic\_data.drop(columns = 'Cabin' , axis = 1)

titanic\_data['Age'].fillna(titanic\_data['Age'].mean() , inplace = True)

titanic\_data.isnull().sum()

titanic\_data['Embarked'].fillna(titanic\_data['Embarked'].mode()[0] , inplace = True)

titanic\_data.isnull().sum()

titanic\_data['Survived'].value\_counts()

sns.set()

sns.countplot('Survived',data = titanic\_data)

sns.countplot('Sex',hue='Survived',data=titanic\_data)

sns.countplot('Pclass',data = titanic\_data)

sns.countplot('Pclass',hue='Survived',data=titanic\_data)

titanic\_data.replace({'Sex':{'male':0,'female':1},'Embarked':{'S':0,'C':1,'Q':2}},inplace = True)

X = titanic\_data.drop(columns = ['PassengerId','Name','Ticket','Survived'],axis = 1)

Y = titanic\_data['Survived']

X\_train,X\_test,Y\_train,Y\_test = train\_test\_split(X,Y,test\_size = 0.2 , random\_state = 2)

model=LogisticRegression()

model.fit(X\_train, Y\_train)

X\_train\_prediction = model.predict(X\_train)

print(X\_train\_prediction)

training\_data\_accuracy = accuracy\_score(Y\_train,X\_train\_prediction)

print(training\_data\_accuracy)

X\_test\_prediction = model.predict(X\_test)

print(X\_test\_prediction)

training\_data\_accuracy = accuracy\_score(Y\_test,X\_test\_prediction)

print(training\_data\_accuracy)

##### 

##### Applications :

**.** data science Analysis is applied for various purposes, such as:

* Advanced Image Recognition

* Fraud and Risk Detection

* Targeted Advertising

* Airline Route Planning.

* Augmented Reality

##### CONCLUSION :

I have removed variables like passenger id , name,ticket,fare,cabin as they are not effecting the target variable much

And iam getting an accurancy of 78.21%

##### BIBILOGRAPHY :

References of previous works or websites visited/books referred for analysis about the project, solution previous findings etc.