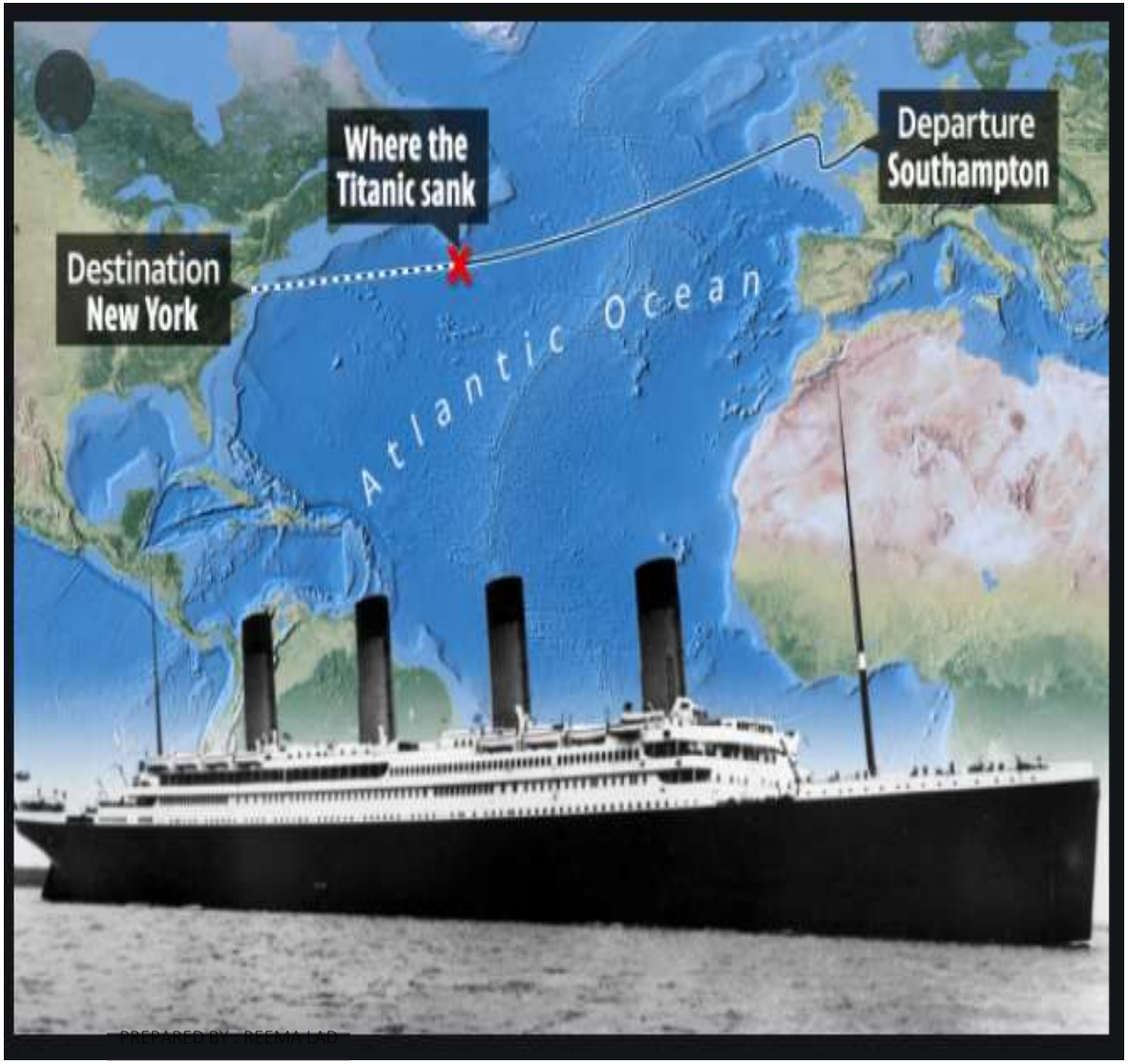
Project A : Titanic - Likely Survival

Owner: Reema Lad

IMS ID No. : IMSPRO143500
Tuesday, August 25, 2020

Project Case



Titanic : The Ship that Sinked

The sinking of the RMS Titanic is one of the most infamous shipwrecks in history.
On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. This sensational tragedy shocked the international community and led to better safety regulations for ships.

One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, children, and the upper-class.

In this challenge, we ask you to complete the analysis of what sorts of people were likely to survive. In particular, we ask you to apply the tools of machine learning to predict which passengers survived the tragedy.

The training set should be used to build your machine learning models. For the training set, we provide the outcome (also known as the "ground truth") for each passenger. Your model will be based on "features" like passengers' gender and class.

Understanding Data And EDA

Raw Data

- Features: 12
- Observations: 891

Raw Data

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
count	891.000000	891.000000	891.000000	891	891	714.000000	891.000000	891.000000	891	891.000000	204	889
unique	NaN	NaN	NaN	891	2	NaN	NaN	NaN	681	NaN	147	3
top	NaN	NaN	NaN	Panula, Master. Eino Viljami	male	NaN	NaN	NaN	347082	NaN	G6	S
freq	NaN	NaN	NaN	1	577	NaN	NaN	NaN	7	NaN	4	644
mean	446.000000	0.383838	2.308642	NaN	NaN	29.699118	0.523008	0.381594	NaN	32.204208	NaN	NaN
std	257.353842	0.486592	0.836071	NaN	NaN	14.526497	1.102743	0.806057	NaN	49.693429	NaN	NaN
min	1.000000	0.000000	1.000000	NaN	NaN	0.420000	0.000000	0.000000	NaN	0.000000	NaN	NaN
25%	223.500000	0.000000	2.000000	NaN	NaN	20.125000	0.000000	0.000000	NaN	7.910400	NaN	NaN
50%	446.000000	0.000000	3.000000	NaN	NaN	28.000000	0.000000	0.000000	NaN	14.454200	NaN	NaN
75%	668.500000	1.000000	3.000000	NaN	NaN	38.000000	1.000000	0.000000	NaN	31.000000	NaN	NaN
max	891.000000	1.000000	3.000000	NaN	NaN	80.000000	8.000000	6.000000	NaN	512.329200	NaN	NaN

Missing Value Imputation

- 77% missing values in Cabin and 19% in Age. Instead of direct imputation will have engineered feature for the same as amount of missing data is too high

Missing Values

	No. of Missing Rows	Missing %
Cabin	687	77.10
Age	177	19.87
Embarked	2	0.22

Correlation Analysis

- Features like Pclass - Fare & Age; Age – SibSp & Parch, Fare – fairly with all other features shows considerable Correlation; which is logical too.

Correlation

Input and Output Variable

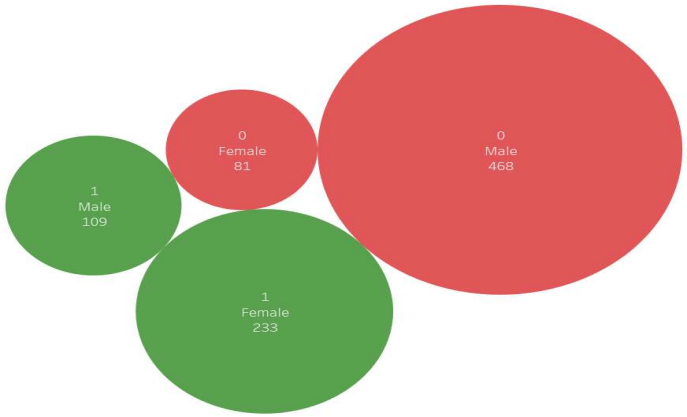
- Response Variable – **Survived**
- Predictors – Gender, Age, PClass

X & Y

	Input Variable	Feature Importance
5	Cat_Gender	32.882742
1	Cat_Age	18.046956
2	Cat_Deck	11.977153
4	Fare_Group	9.954160
0	Pclass	8.929249
3	FamilySize	8.598435
7	Cat_Family	4.908613
6	Cat_Embarked	4.702693

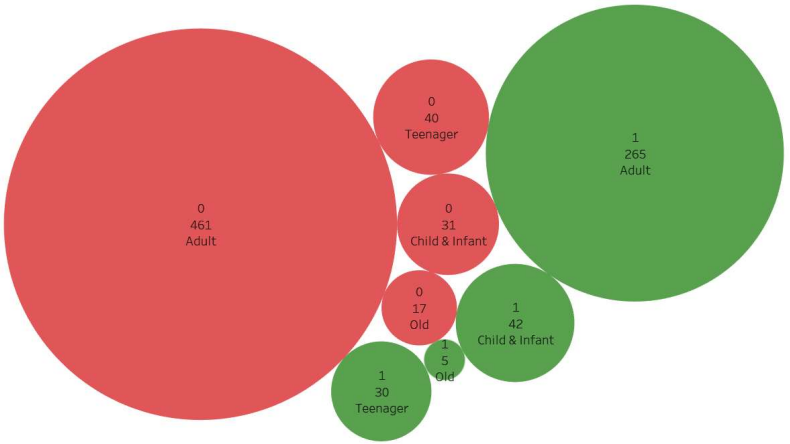
PREPARED BY : REEMA LAD

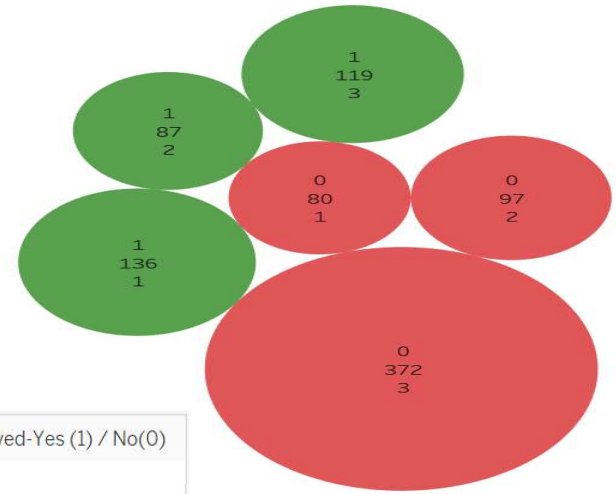
Data Statistics



Gender Wise Survival

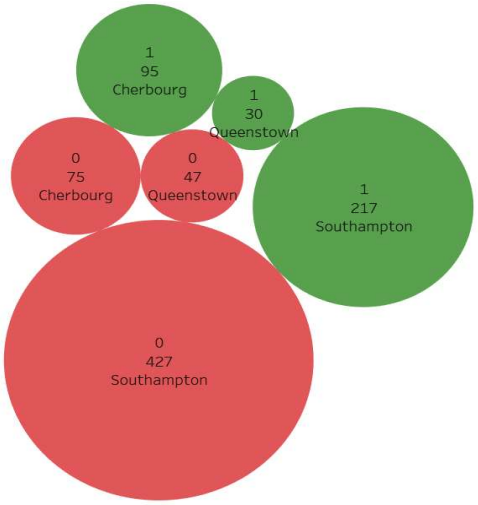
Age Group Wise Survival





Pclass Wise Survival

Embarkment Wise Survival



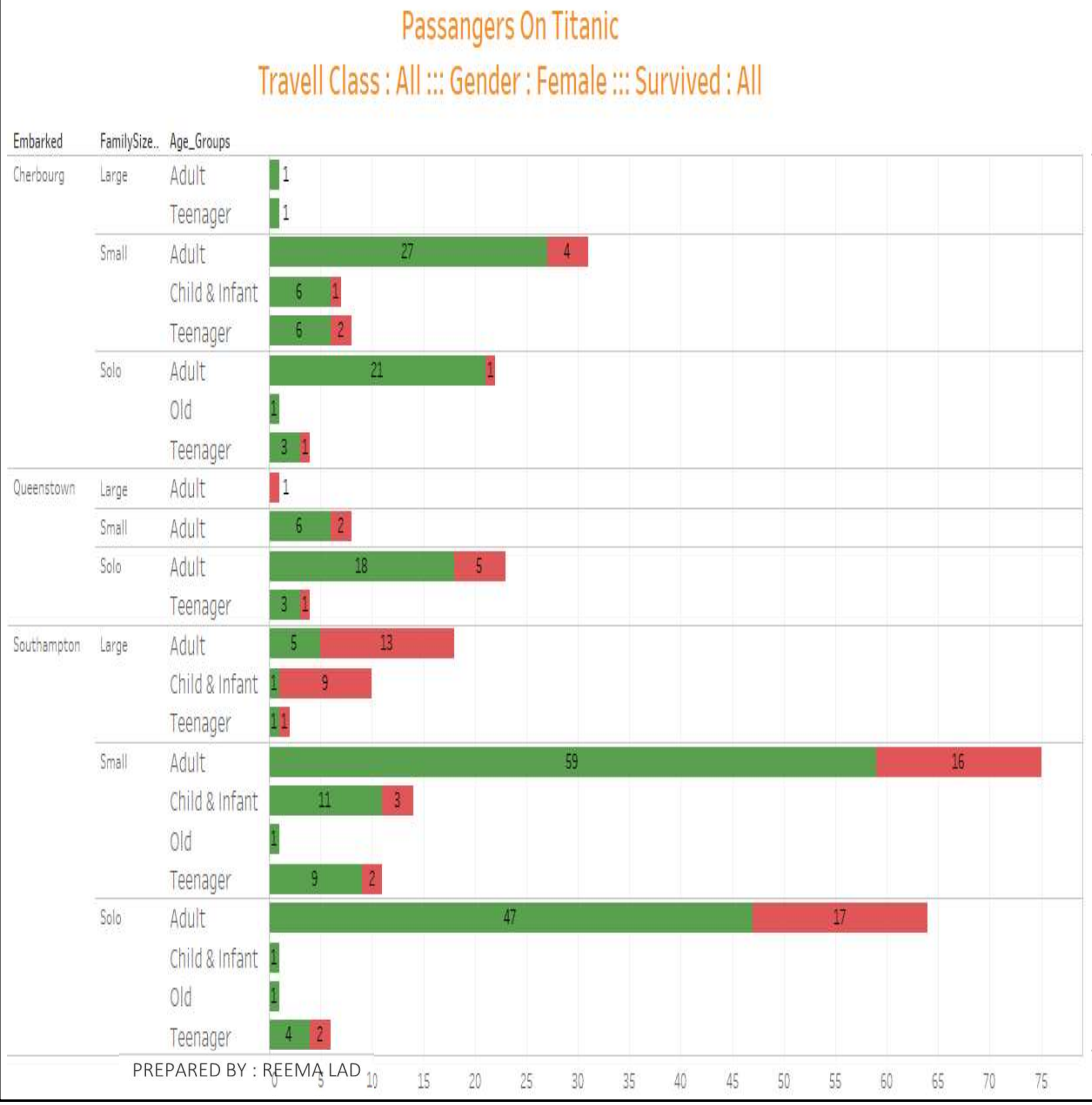
Survived-Yes (1) / No(0)

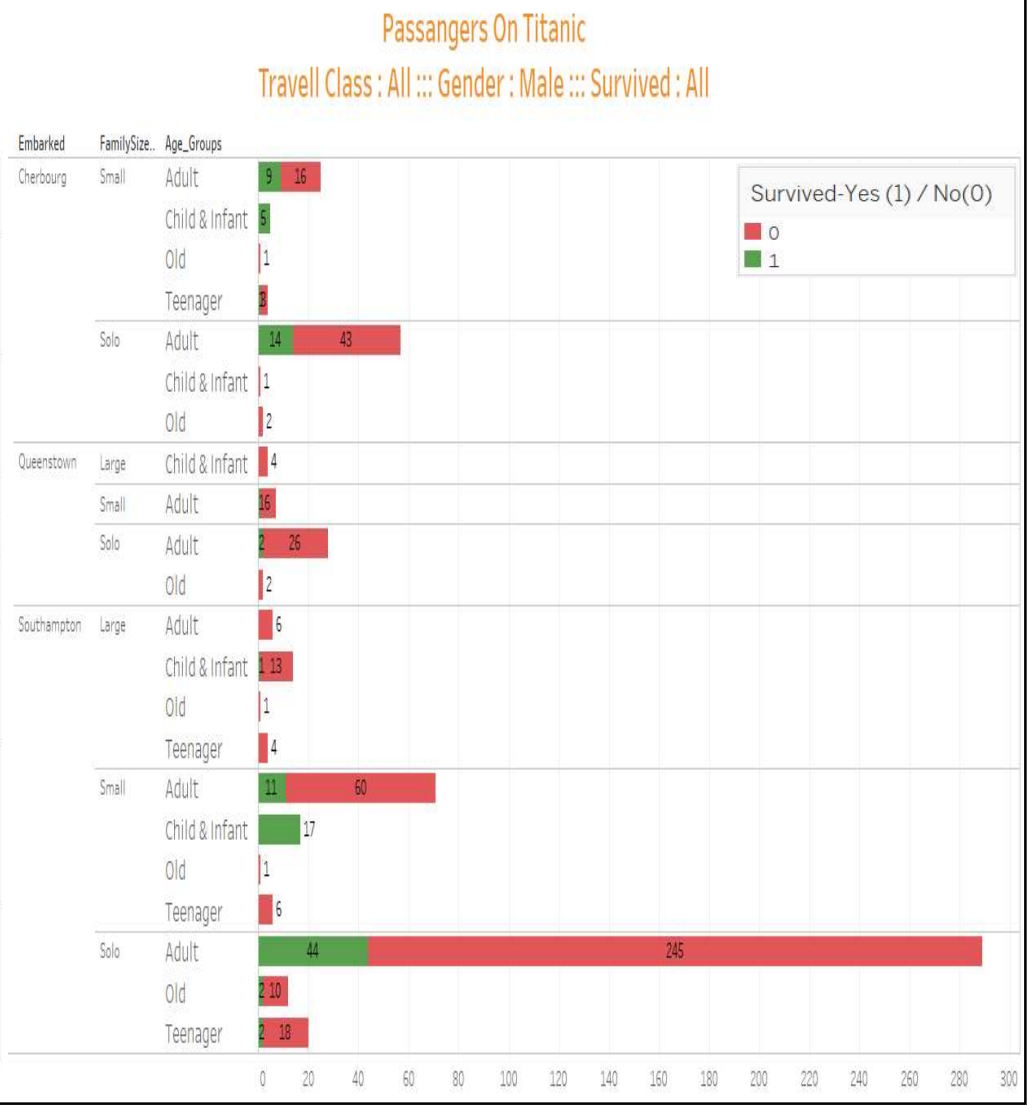
0

1

PREPARED BY : REEMA LAD

Data Statistics





PREPARED BY : REEMA LAD

Approach For Accurate Prediction

Feature Engineering

- Generated new features basis Age – which gave us 'Age Groups'. This also handled the missing values problem.

Cleaned Data & Feature Finalization

- 80% - 20% Train & Test Split
- With Predictors Pclass, Age, Gender and Response Variable Survived

Model Building

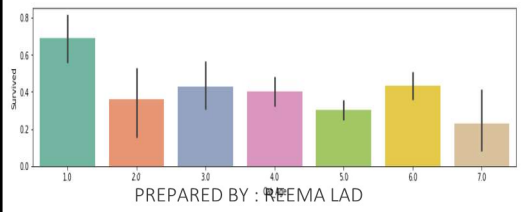
- Logistic Regression
- Gaussian Naive Bayes
- Decision Tree Classifier
- Random Forest Classifier
- KNN Classifier
- Gradient Boost
- Ada Boost

Recommendation

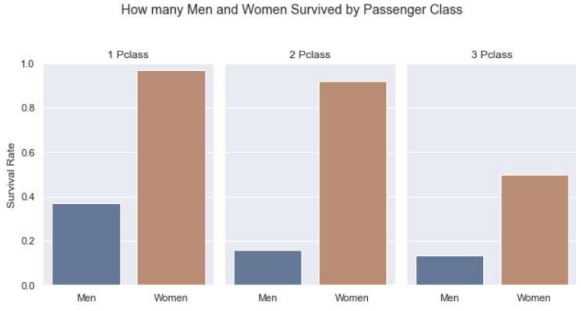
- Models with best accuracy is Decision Tree considering both the Test and Train accuracy score

Feature Engineering

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 9 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Survived      891 non-null    int64
 1   Pclass        891 non-null    int64
 2   Cat_Age       891 non-null    float64
 3   Cat_Deck      891 non-null    int64
 4   FamilySize    891 non-null    int64
 5   Fare_Group    891 non-null    float64
 6   Cat_Gender    891 non-null    int32
 7   Cat_Embarked  891 non-null    int32
 8   Cat_Family    891 non-null    int32
```



Feature Finalization




```
X_Train Info
<class 'pandas.core.frame.DataFrame'>
Int64Index: 712 entries, 151 to 595
Data columns (total 3 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Pclass      712 non-null    int64
 1   Cat_Age     712 non-null    float64
 2   Cat_Gender  712 non-null    int32
```

```
X_Test Info
<class 'pandas.core.frame.DataFrame'>
Int64Index: 179 entries, 727 to 829
Data columns (total 3 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Pclass      179 non-null    int64
 1   Cat_Age     179 non-null    float64
 2   Cat_Gender  179 non-null    int32
```

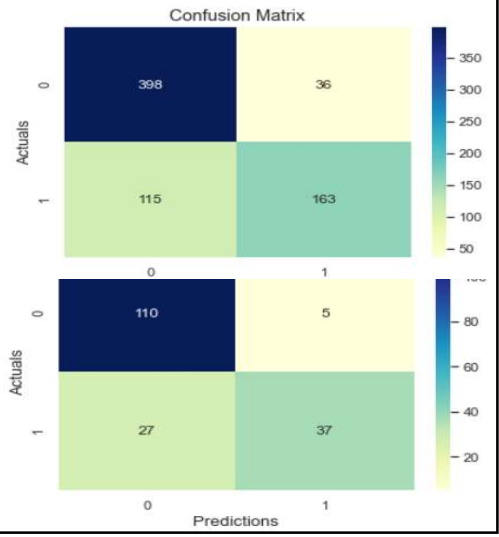
```
Y_Train Info
<class 'pandas.core.frame.DataFrame'>
Int64Index: 712 entries, 151 to 595
Data columns (total 1 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   Survived  712 non-null    int64
```

```
Y_Test Info
<class 'pandas.core.frame.DataFrame'>
Int64Index: 179 entries, 727 to 829
Data columns (total 1 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   Survived  179 non-null    int64
```

Model Building

	Model	Train Accuracy	Test Accuracy
3	Decision Tree	78.79	82.12
4	Random Forest	78.65	82.12
6	Gradient Boosting	78.65	82.12
0	Logistic Regression - All Fea.	76.97	80.45
5	KNN	71.77	79.89
7	Ada Boost	76.97	79.89
2	Naive Bayes	77.67	78.77
1	Logistic Regression	79.07	78.21

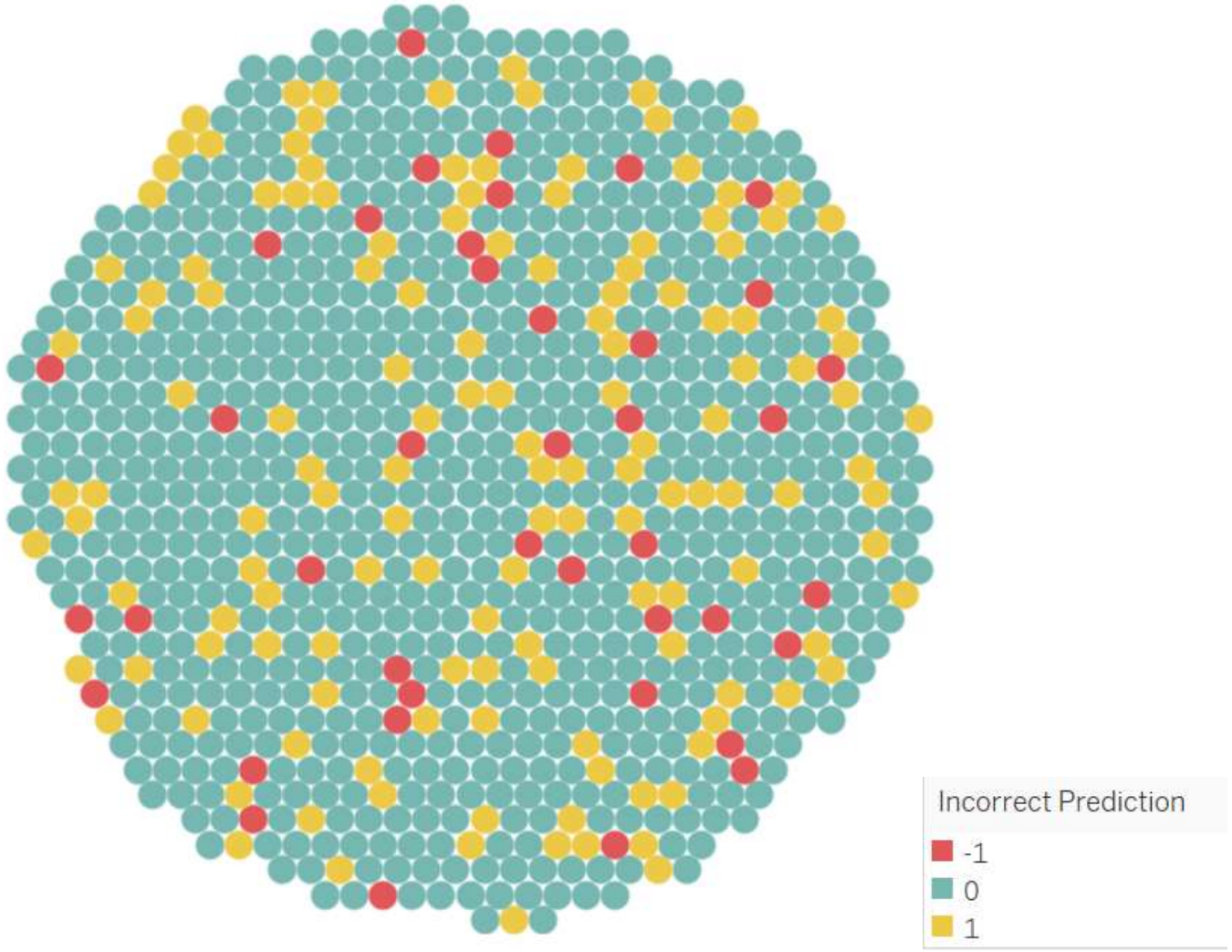
Recommendation



PREPARED BY : REEMA LAD

Error Analysis

-1:Predicted Survived & 1:Predicted Not Survived & 0:Correct Prediction



PREPARED BY : REEMA LAD

Thank You