Report

Covid-19 Herd Immunity Classification

IT788A Introduction to Data Science 15 ECTS

DS mini project assignment

a22sujjo

Suja John

2022-12-23

Contents

1. Introduction ..... **Error! Bookmark not defined.**
1.1 Research questions/hypotheses 5
2. Background 5
3. Data 6
3.1 Data preparation 7
4. Approach 10
5. Results 13
6. Discussion 16
6.1. Limitation and Challenges 16
7. Conclusions 16
8. Reflections on own work 17
References 18

ABSTRACT

Herd immunity classification for Covid-19 using total vaccination and population data of each country. Here I took 60 percentage of total population of a country has vaccinated then the country will be declared or classified as herd immunity. I did the analysis using a real and live database, that is updated on daily basis by Covid-19 data repository by the Center for Systems Science and Engineering (CSSE) at John Hopkins University. As it is a real database, it required a lot of data preprocessing. Using this database, I am going to classify whether the country has formed a herd immunity or not. The total vaccination and the population of the country are the key parameters used for this classification. Here binary classification is carried out. So for that Logistic Regression classifier, Naïve Bayes, Decision Tree classifiers are used. Logistic Regression is used to be one of the most widely used methods in data mining and especially in binary data classification. This paper is focused mainly on an overview of the most important aspect of Logistic Regression, Naïve Bayes, Decision Tree and Machine Learning perspective.

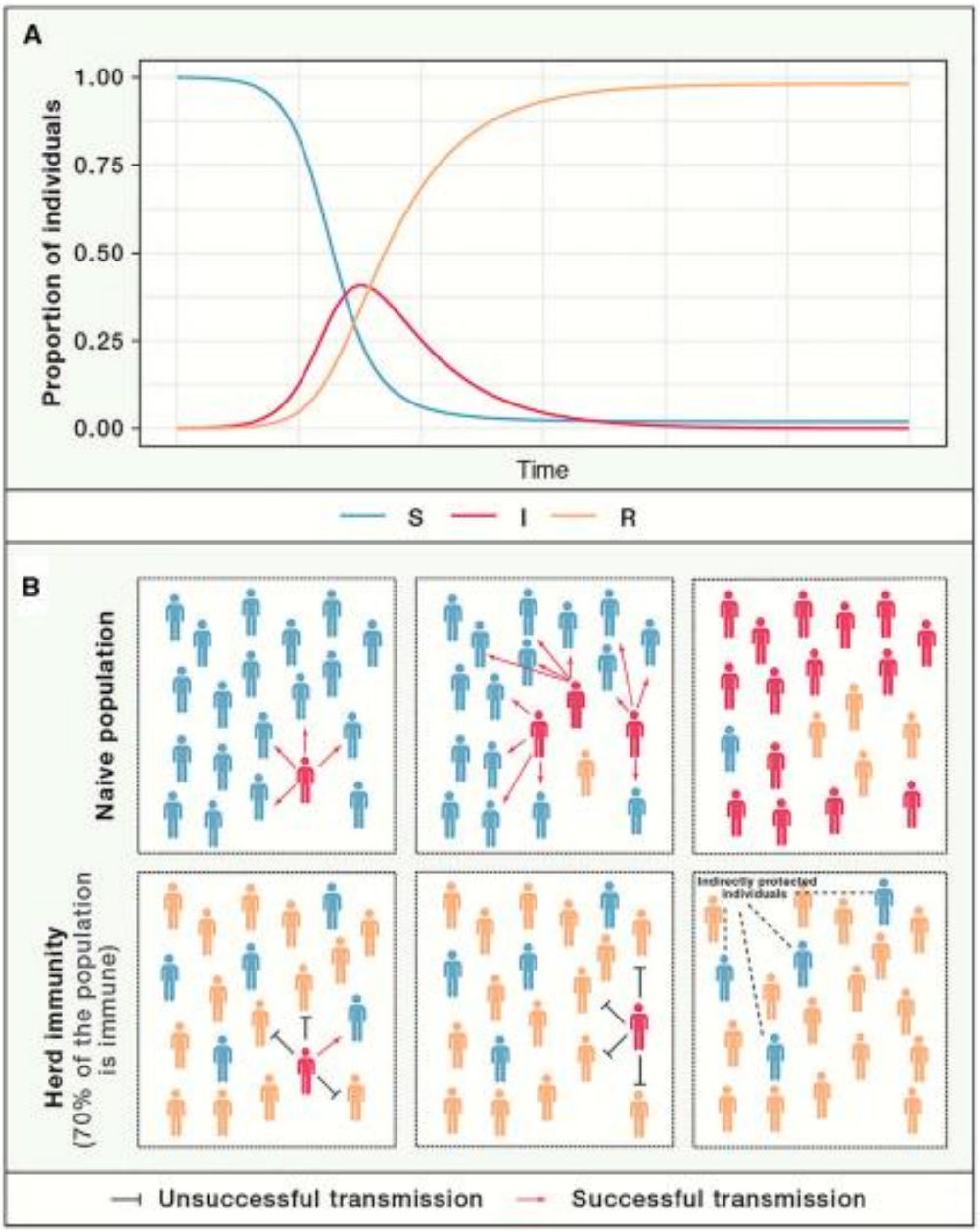
1. INTRODUCTION

In December 2019, an outbreak of pneumonia of unknown origin as reported in Wuhan, Hubei Province, China. (Marco Ciotti et al, 2020) Pneumonia cases were epidemiologically linked to the Huanan Seafood wholesale market. Aa a result of a few experiments and investigations of virus genome analysis showed it to be a novel coronavirus related to SARS-Cov-1 and therefore named severe acute respiratory syndrome Coronavirus2 (SARS-Cov-2). The global spread of SARS-Cov-2 and the thousands of deaths caused by Coronavirus disease (Covid-19) led the World Health Organization to declare a “Pandemic” on 12 March 2020. Till date, the world has paid a high price in this pandemic in terms of human lives lost, economic repercussions and increased poverty (Marco Ciotti et al, 2020).

Health care sector is a very important aspect of a country. In the middle of the crisis, it is important for a country to form a herd immunity. Herd immunity occurs when a large portion of a community or the herd becomes immune to a disease. The spread of disease from person to person becomes unlikely hen herd immunity is achieved. Total vaccination plays key role here. The countries have to form a herd immunity as soon as possible for the wellbeing of their citizens.

Herd immunity is the indirect protection from a contagious infectious disease that happens when a population is immune either through vaccination or immunity developed through previous infection. This means that even people who doesn’t trigger immunity, are protected because people around them who are immune can act as buffer between them and an infected person. Once herd immunity has been established for a while, and the ability of the disease to spread is hindered, the disease can eventually be eliminated (Haley E. Randolph and Luis B. Barreiro, 2020).

Covid-19 has a global pandemic for almost 2 years. It made us stay in our four walls for almost 24 months. From the country's point of view, the herd community formation is the best way to resist the increase of the covid-19 cases in a country.



In the figure, SIR (Susceptible, Infectious, Recovered or Vaccinated) (Haley E. Randolph and Luis B. Barreiro, 2020).

The objective of this project is to build a classification model using the data to classify whether the country has formed a herd immunity or not. The total vaccination and population of the country are the parameters are used for this classification. Using python programming language, build models using Logistic Regression and looked at the performance metrics such as model score, confusion matrix and classification report to evaluate the models.

Logistic regression is one of the most important statistical and data mining techniques employed in the field of data science for the analysis and classification of binary and proportional response database. Some of the main advantages of the Logistic Regression are that it can naturally provide probabilities and extend to multi class classification problems (Maalouf, 2011). A supervised learning approach as implemented in the study to classify herd immunity.

The Naïve Bayes classification algorithm has been widely used because of its simplicity and easy to use in both the training and classification. It allows each attribute to contribute towards the final decision equally and independently from each other attributes (S.L. Ting et al, 2011). This assumption is called class conditional independence. It is made to simplify the computation involved and, in this sense, is considered "naïve" (Leung, 2007).

Decision Tree are one of the most popular models foe classification in many application domains (F. Li et al, 2018). Decision Tree is a "tree-shaped diagram representing a sequential decision process in which attribute values are successively tested to infer an unknown state" (S. Ohta et al, 2008). A Decision Tree is composed of a root node, internal/test nodes, and leaf nodes having a class or a label.

1.1 RESEARCH QUESTIONS/ HPOTHESIS

In this project I focused to

- Whether a country has become a herd immunity or not?
- What are the factors that influence the herd immunity classification?
- Do the total deaths in a country affect its herd immunity classification?
- Which classification model performs best?

2. BACKGROUND

Nowadays, Machine Learning (ML) algorithms are used in wide computational areas due to its effective performance. The health care system is one of the most important parts of a country. Covid-19 has infected the whole world that it is associated with the rapid growth of essential data and the need to analyze the relationship and hierarchy between data leads to the need for machine learning and data mining is in the health system that is very effective in prevention, diagnosis and treatment (Naseeba M. Abdulkareem et al, 2021). Covid-19 vaccination progress using machine learning algorithms is also an important subject of researches.

(Muhammad et al, 2020) performed research on data mining models for forecasting the recovery of patients infected by Covid-19. The model will estimate the total time it will take for Covid-19 patients to recover and discharge from isolation centers. The model assist health care providers in determining the recovery and stability of freshly contaminated individuals.

The models were created using Korea Centers for Diseases and Prevention (KCDC) dataset. Python programming and data mining techniques such as Support Vector Machine (SVM), Naïve Bayes, Decision Tree, Logistic Regression, an K-Nearest Neighbor (KNN) were used to create the models. The results indicate that the model Develop using Decision Tree algorithm is more effective in predicting the probability of recovery of patients, with a 99.85% overall accuracy.

(Ritonga et al, 2021) developed the Naive Bayes classification algorithm on twitter data with the keyword 'Covid-19' indexed by the keyword 'Vaccine', in Indonesia tweets in the second and third week of January 2021. The study reported 39 percent optimistic sentiment, 56 percent pessimistic sentiment, and 1 percent favorable sentiment. A lot of negative opinion was created because most of them think that the vaccination was not effective at that time.

(Akib Mohi Ud Din Khanday et al, 2020) develop a control system that will detect the coronavirus. In this paper they classified textual clinical report into four classes b using classical ensemble machine learning algorithms. Logistic Regression and Multinomial Naïve Bayes showed better results than other ML algorithms b having 96.2 percent test accuracy.

3. DATA

The data set is about the Covid-19 details of various countries on different dates. The dataset contains details like Location, Date, Total cases, New cases, Total Deaths, Ne Deaths, Total vaccination etc. The data as taken from a Github project which has the link to US public health data set. The total vaccinations and total population of the country has the effect on whether the country has formed herd immunity or not. The other factors that affect the herd classification can be found by defining the correlation between them.

- This is the link where I downloaded the dataset

  https://github.com/owid/covid-19-data/tree/master/public/data

This dataset has 238703 records and 33 features. Out of 33 columns there are 29 features are of float data type and are of string data type. There are almost 3025876 null values in the given dataset.

This dataset is a real and live dataset, as it is updating on daily basis by Covid-19 data repository by the Center for Systems Science and Engineering (CSSE) at john Hopkins University (JHU). The dataset is first updated on 24th February 2020.

The data as in csv format which I uploaded to Jupyter Notebook before pre-processing. Then I pre-processed the data by various data cleaning methods.

These are the attributes in the dataset

```
iso_code
continent
location
date
total_cases
new_cases
new_cases_smoothed
total_deaths
new_deaths
total_cases_per_million
new_cases_per_million
total_deaths_per_million
new_deaths_per_million
icu_patients
icu_patients_per_million
weekly_icu_admissions
weekly_icu_admissions_per_million
weekly_hosp_admissions
weekly_hosp_admissions_per_million
total_tests
new_tests
total_tests_per_thousand
new_tests_per_thousand
positive_rate
total_vaccinations
people_fully_vaccinated
poplation density
median_age
handwashing_facilities
hospital_beds_per_thousand
life_expectancy
human_development_index
population
```

3.1 DATA PREPARATION

Data preparation is the process of cleaning and transforming raw data prior to processing and analysis. It is an important step to prior to processing and that involve reformatting data, making corrections to data and the combining the data sets to enrich the data.

- Collect the data

The data was in csv format which we uploaded to Jupyter File before pre-processing. Shape of the data and null value present in the data found out.

```
df=pd.read_csv("covid-data.csv")
```

```
df.shape
```

```
(238703, 33)
```

```
df.isnull().sum().sum()
```

```
3025876
```

Checked the first and last 5 rows of the dataset.

```
df.head()
```

	iso_code	continent	location	date	total_cases	new_cases	new_cases_smoothed	total_deaths	new_deaths
0	AFG	Asia	Afghanistan	24-02-2020	5.0	5.0	NaN	NaN	NaN
1	AFG	Asia	Afghanistan	25-02-2020	5.0	0.0	NaN	NaN	NaN
2	AFG	Asia	Afghanistan	26-02-2020	5.0	0.0	NaN	NaN	NaN
3	AFG	Asia	Afghanistan	27-02-2020	5.0	0.0	NaN	NaN	NaN
4	AFG	Asia	Afghanistan	28-02-2020	5.0	0.0	NaN	NaN	NaN

```
df.tail()
```

	iso_code	continent	location	date	total_cases	new_cases	new_cases_smoothed	total_deaths	new_deaths
238698	ZWE	Africa	Zimbabwe	25-11-2022	257893.0	0.0	0.0	5606.0	0.0
238699	ZWE	Africa	Zimbabwe	26-11-2022	257893.0	0.0	0.0	5606.0	0.0
238700	ZWE	Africa	Zimbabwe	27-11-2022	257893.0	0.0	0.0	5606.0	0.0
238701	ZWE	Africa	Zimbabwe	28-11-2022	257893.0	0.0	0.0	5606.0	0.0
238702	ZWE	Africa	Zimbabwe	29-11-2022	257893.0	0.0	0.0	5606.0	0.0

- Cleaning the data

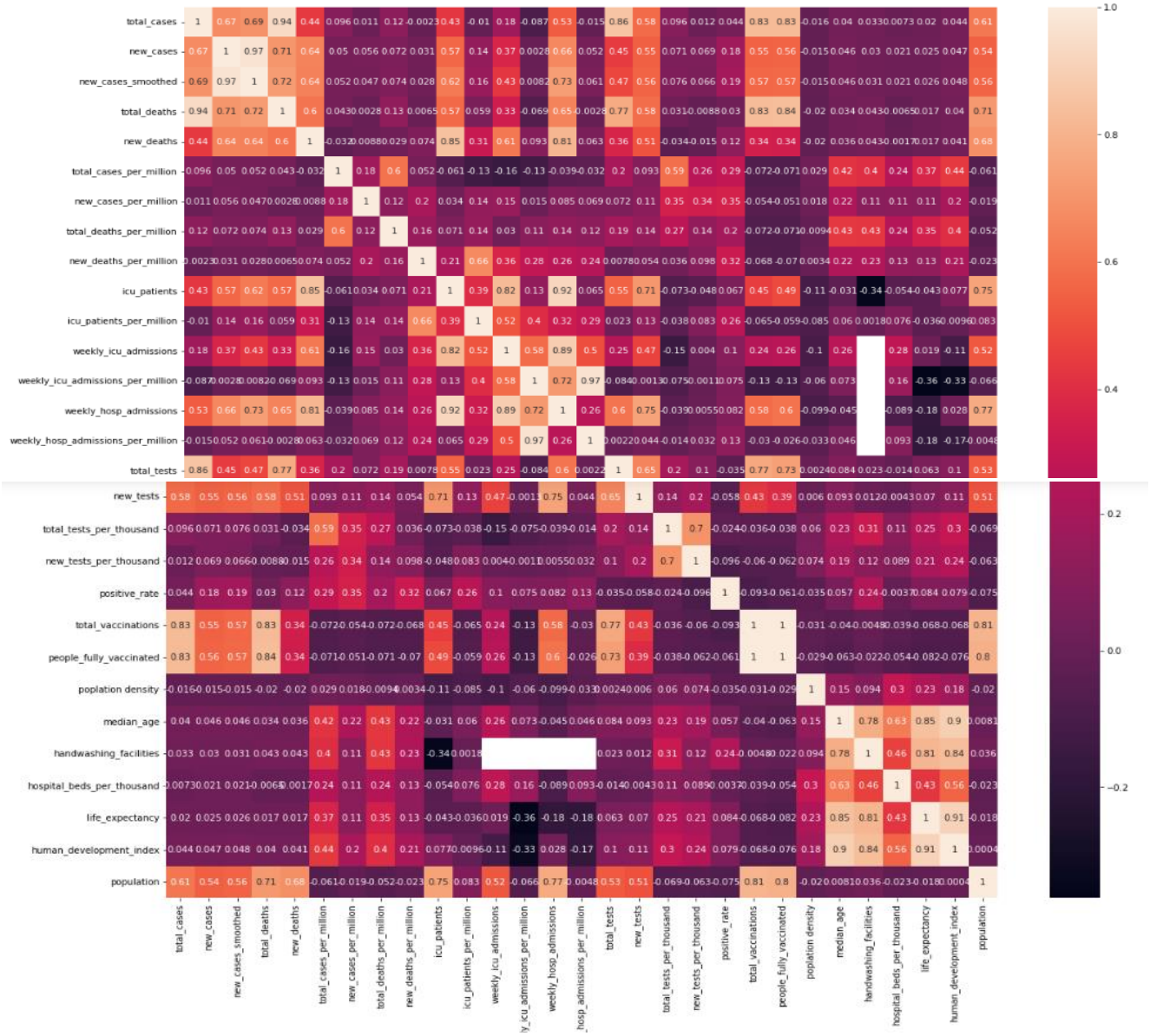
To treat the missing values in numeric columns in the dataset I used SimpleImputer method using the strategy mean. Then checked the missing value again.

```
temp_df.isnull().sum().sum()
```

```
0
```

- Analyzing the data

For analyzing the data, Heat map of correlation of attributes made and find out the correlation between them. From that I deleted one of the highly correlated attributes.



Delated columns

- new_cases_smoothed
- icu_patients
- weekly_hosp_admissios_per_millions
- people_full_vaccinated
- median_age
- Iso_code
- continent

4. APPROACH

This project is aimed to the classification of countries whether they formed herd immunity at a particular date. So for that I created two new columns herd and herdimmunity(yes/no).

herd attribute contains the value that is equal to the 60 percentage of population and herdimmunity(yes/no) contains if total vaccination is greater than herd then the value is True for herdimmunity(yes/no) otherwise False. Then I converted the type of the herdimmunity(yes/no) column Boolean to Int.

```
temp_df['herdimmunity(yes/no)'].value_counts()
```

```
1    210060
0     28643
Name: herdimmunity(yes/no), dtype: int64
```

- Duplicate Rows

Find out an duplicate rows in the dataset using duplicated()

```
dups = df.duplicated()
print('Number of duplicate rows = %d' % (dups.sum()))
print(df.shape)
```

```
Number of duplicate rows = 0
(238703, 26)
```

- Outlier Treatment

Then I tried to find out the outliers present in the dataset using the equation - The outliers are when value < Q1 - 1.5 x IQR or Q3 + 1.5 x IQR < value, where IQR = Q3 - Q1 (Interquartile range).

Where The first quartile, Q1, is the middle number that falls between the smallest value of the dataset and the median.

The third quartile, Q3, is the middle number that falls between the median and the largest value of the dataset.

Then replaced the values lower than the Q1 - 1.5 x IQR value with the lower range (Q1 - 1.5 x IQR) and replaced the values higher than Q1 + 1.5 x IQR value with the upper range (Q1 + 1.5 x IQR.

Outliers in columns in the dataset

```
total_cases:42656
new_cases:45405
total_deaths:19591
new_deaths:17574
total_cases_per_million:25270
new_cases_per_million:23402
total_deaths_per_million:19793
new_deaths_per_million:21768
icu_patients_per_million:32233
weekly_icu_admissions:8139
weekly_icu_admissions_per_million:8139
weekly_hosp_admissions:19210
total_tests:8426
new_tests:57830
total_tests_per_thousand:77921
new_tests_per_thousand:75403
positive_rate:84030
total_vaccinations:65125
poplation density:10670
handwashing_facilities:94664
hospital_beds_per_thousand:23647
life_expectancy:4887
human_development_index:2961
population:30790
herd:30790
```

- Imbalance data

By looking at the targeted attribute ie herdimmunity(yes/no), So from this we can summarise that the dataset has the imbalanced data.

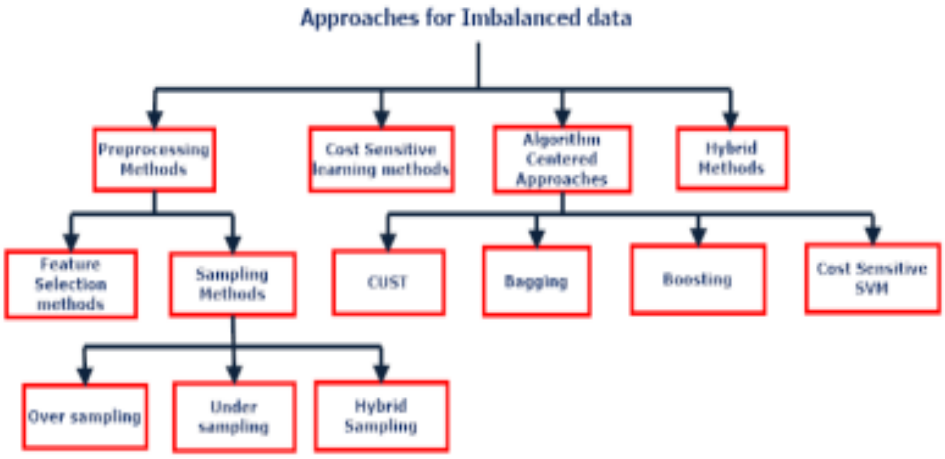
```
temp_df["herdimmunity(yes/no)"].value_counts()
```

```
1    210060
0     28643
Name: herdimmunity(yes/no), dtype: int64
```

The class showing 1 or true value has 210060 records and 0 or false value has 28643 records. This Skewed class proportions is called imbalanced. The ratio of 1 and 0 classes is 88% : 12%. 1 value class is known as majority class an 0 value class is known as minority class (Harsurinder Kaur et al, 2019).

The major imbalance problems, which makes predictions of trained models to be biased towards the majority class. There are various approaches are proposed to overcome the imbalanced data.

Classification of approaches for imbalanced data (Kaur, 2019)



- Sampling Methods.

It is an easy and popular approach to balance the class distributions of the training data. (Harsurinder Kaur et al, 2019) The sampling methods are effective alternative for supervised learning (Lu Cao and Yiku Zhai, 2015). (Bee Wah Yap et al) have proved that sampling methods outperforms bagging and boosting. There are various options to perform the sampling:

Over-sampling: The basic idea of over-sampling is to increase the size of the minority class to obtain balanced classes. Duplication of samples is done in random over-sampling in which samples are randomly selected. over-fitting is the main issue arises in over-sampling (Ganganwar, 2012). In (Nitesh V. Chawla et al, 2009) Chawla proposed Synthetic Minority Over-Sampling technique (SMOTE). In SMOTE, synthetic samples are produced by the help of minority class samples. It also suggests the idea that using combination of sampling methods can be a good option for improving the classifier performance while dealing with the imbalanced class distributions

Under-sampling: It is a pre-processing method that draws the random set of samples from the majority class to balance the classes and rest of the samples are ignored (Nguyen Ha Vo and Yonggwan Won, 2007). The size of the data space is measured to draw desirable class distribution ratio.

Hybrid sampling: Hybrid sampling methods are those that apply both re-sampling techniques to attain balance in the data. (Qiang Wang, 2014)proposed a technique of combining sampling methods, under-sampling and over-sampling to handle the problem of imbalance data (Harsurinder Kaur et al, 2019).

The most popular approach, over-sampling adds artificial samples to the data space, known as SMOTE is use in this project to deal the imbalance.

After applying SMOTE on the train data set, the ratio of 1(value) : 0(value) has become 1:1.

```
from imblearn.over_sampling import SMOTE
sm=SMOTE(random_state=2)
X_train_res,y_train_res=sm.fit_resample(X_train,y_train.ravel())
```

```
from scipy.stats import itemfreq
itemfreq(y_train_res)
```

```
array([[     0, 147042],
       [     1, 147042]], dtype=int64)
```

5. RESULT

Logistic Regression

Using this balanced train data set I build a model using Logistic Regression and get the model accuracy as

Accuracy

```
# Accuracy - Training Data
model.score(X_train_res, y_train_res)
```

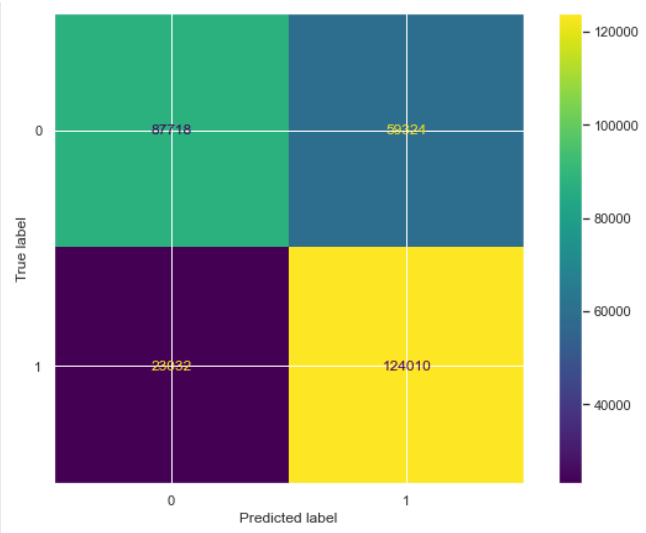
```
0.7199575631452239
```

```
# Accuracy - Test Data
model.score(X_test, y_test)
```

```
0.8123751937551493
```

Its confusion matrix is

Train data



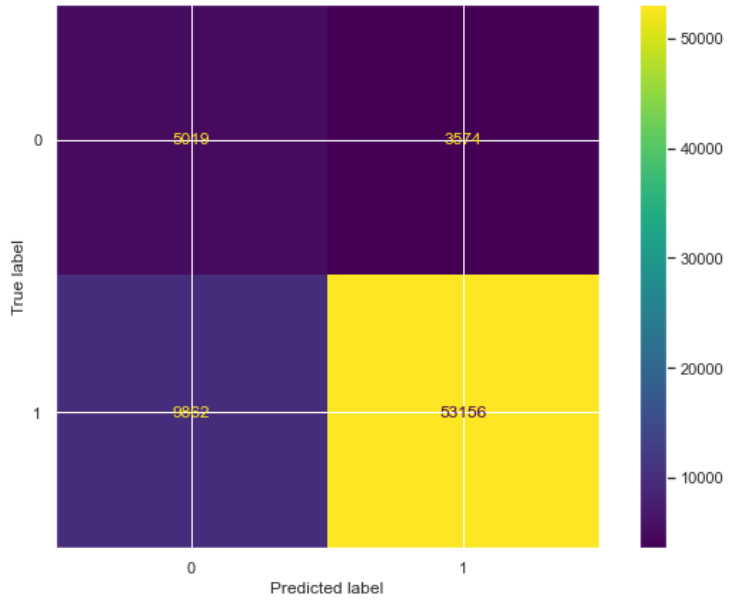
Given above is the confusion matrix for Training data using the Logistic regression model.

• True Positive =124010, which is actually True (value=positive or 1) and has been predicted True too

• False Negative=23032, which is actually True (value=positive or 1), but predicted False (value=negative or 0)

• False Positive =50324, which is actually False (value=negative or 0), but predicted True (1)

• True Negative =87718, which is actually False and has been predicted False too.



Given above is the confusion matrix for test data using the Logistic regression model.

• True Positive =53156, which is actually True (value=positive or 1) and has been predicted True too

• False Negative=3574, which is actually True (value=positive or 1), but predicted False (value=negative or 0)

• False Positive =9862, which is actually False (value=negative or 0), but predicted True (1)

• True Negative =5019, which is actually False and has been predicted False too.

HYPER PARAMETER TUNING

Grid search method

The grid search method is used for logistic regression to find the optimal parameters for building the model.

```
grid={'penalty':['l2','none'],
      'solver':['sag','lbfgs'],
      'tol':[0.1,0.01]}
```

The above is the list of parameters that has been loaded into the grid search CV.

```
LogisticRegression(max_iter=10, n_jobs=2, solver='sag', tol=0.1)
```

The above set of parameters were considered the best for this modelling.

Its accuracy is

```
# Accuracy - Training Data
best_model.score(X_train_res, y_train_res)
```

```
0.6846683260565009
```

```
# Accuracy - Test Data
best_model.score(X_test, y_test)
```

```
0.7874209269525632
```

Naïve Bayes

I build the model using Naïve Bayes Algorithm and got the Accuracy as

```
model_score=nb_model.score(X_train_res,y_train_res)
print(model_score)
```

```
0.8820575073788441
```

```
model_score=nb_model.score(X_test,y_test)
print(model_score)
```

```
0.8413092960578682
```

Decision Tree

I build the model using Decision Tree Algorithm ad got the Accuracy as

```
# Accuracy - Train Data
dt_model.score(X_train_res, y_train_res)
```

```
0.9947123950979992
```

```
# Accuracy - Test Data
dt_model.score(X_test, y_test)
```

```
0.9878230998031029
```

6. DISCUSSION

To find the solution I have built models using Logistic Regression, Naïve Bayes and Decision Tree from the results we can summarise that Decision Tree model showed the high accuracy, train and test data accuracies as 99% and 98% respectively.

6.1 Limitation and Challenges

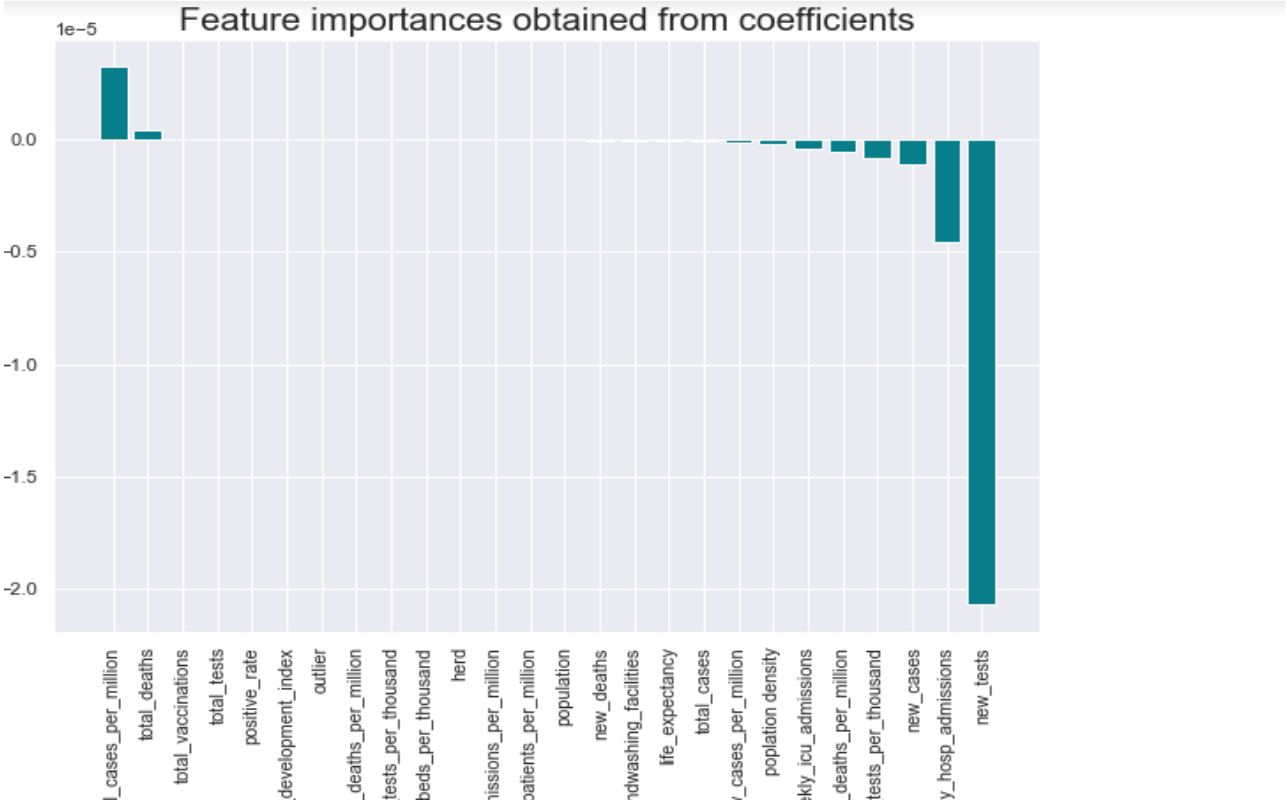
I could have created model using KNN algorithm and SVM algorithm and compare the accuracy of the models. I could have found which date the countries have become a herd community and found their population: herd ratio.

7. CONCLUSION

Using the model, we can find out whether the country become a herd immunity or not and the model built using Decision Tree algorithm showed the best accuracy. So we can say that Decision Tree Model is the best Model.

The Features that are affecting the targeted feature ie herdimmunity(yes/no) are

- Total_cases_per_million
- Total_deaths
- Weekly_icu_admissions
- Total_deaths_per_thousands
- New_cases
- Weekly_hosp_admissions
- New_tests



8. REFLECTIONS ON OWN WORK

Firstly, when I got the dataset from the Github for the Covid-19 Classification Project, I studied the features and find out the scope of classifying the countries whether they formed herd immunity or not. For doing this project I refer some ML Classification Projects and read some articles related to classification.

If I would start it all over again, then I will concentrate more on data pre-processing techniques and the most proper ways to handle imbalance data and the classification algorithms that handle imbalanced data.

References

Akib Mohi Ud Din Khanday et al. (2020). Machine learning based approaches for detecting COVID-19 using clinical text data.

Bee Wah Yap et al. (u.d.). An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets. 2014.

F. Li et al. (2018). Cost-sensitive and hybrid-attribute measure multi-decision tree over imbalaced dataset.

Ganganwar, V. (2012). An overview of classification algorithms for imbalanced datasets.

Haley E. Randolph and Luis B. Barreiro. (2020). Herd Immunity: Understanding COVID-19.

Harsurinder Kaur et al. (2019). A Systematic Review on Imbalanced Data Challenges in Machine Learning: Applications ans Solutions.

Kaur, H. (2019). A Syetematic Revie on Imbalanced Data Challenges in Machine Learning.

Leung, K. M. (2007). Naive Bayesian Classifier.

Lu Cao and Yiku Zhai. (2015). Imbalanced data classification based on a hybrid resampling SVM method. In Proceedings of the Ubiquitous Intelligence and Computing.

Maalouf, M. (2011). Logistic regression in data analsis:an overview.

Marco Ciotti et al. (2020). The COVID-19 pandemic.

Muhammad et al. (2020). Preictive data mining models for novel coronavirus(COVID -19) infected patients' recovery.

Naseeba M. Abdulkareem et al. (2021). Covid-1 world vaccination progress using machine learning classification algorithms.

Nguyen Ha Vo and Yonggwan Won. (2007). Classification of unbalanced medical data with weighted regularized least squares.

Nitesh V. Chawla et al. (2009). SVMs modeling for highly imbalanced classification.

Qiang Wang. (2014). . A hybrid sampling SVM approach to imbalanced data classification. In Abstract and Applied Analysis.

Ritonga et al. (2021). Sentimental analysis of COVID-19 vaccine in Indonesia using Naive Bayes Algorithm. IOP Publishing. .

S. Ohta et al. (2008). Minimizing false positive of a decision tree classifier for intrusion detection on the internet.

S.L. Ting et al. (2011). Is Naive Bayes a good classifier for document classification.