Teesside University

MSc Computer science

School of computing, engineering, and digital technologies

Module – Machine learning

(CIS4035-N)

STUDENT ID- C2727569

NAME – ILOANUSI RICHARD CHIBUEZE

Academic year: 2023/2024

ICA ON APPLICATION OF MACHINE LEARNING AND ITS ALGORITHMS IN A COVID-19 DATASET.

20th April 2024

INTRODUCTION

The novel coronavirus SARS-CoV-2 was the cause of the COVID-19 pandemic that surfaced in late 2019. The virus was initially discovered in Wuhan, China's Hubei province, in December 2019. The first cases were connected to a Wuhan seafood market, indicating a possible animal-to-human spread. Subsequent cases, however, suggested that the virus could travel from one individual to another, which accelerated its global spread.

On January 30, 2020, the World Health Organisation (WHO) designated COVID-19 as a Public Health Emergency of International Concern. On March 11, 2020, the WHO upgraded the designation to pandemic. Governments across the world implemented several measures to stop the virus's rapid spread, including travel bans, lockdowns, social distancing orders, mask laws, mass testing, and contact tracking.

Wide-ranging effects of COVID-19 include disruptions to almost every facet of daily life, such as social interactions, healthcare systems, economy, and educational institutions. Significant deaths, financial suffering, and interruptions to livelihoods have resulted from it. Along with continued research into cures and preventive measures, efforts to battle the virus have included the creation and distribution of vaccinations.

The most typical symptoms include

fever
coughing
fatigue
loss of smell or taste.

Severe symptoms include

shortness of breath or difficulty breathing.

speech or movement difficulties

disorientation, or chest pain

Take the following precautions to stop spreading infection and reduce Covid-19 transmission:

- When a vaccine is accessible to you, get it done.
- Maintain a minimum of one meter's distance from people around you, even if they do not appear sick.
- In situations where physical distance is not feasible or in poorly ventilated environments, wear a mask that fits appropriately.
- Select open areas with good ventilation over enclosed ones. If it's inside, open a window.
- Hands should be cleaned with an alcohol-based hand rub or frequently washed with soap and water.
- When you sneeze or cough, keep your mouth and nose covered.
- Stay at home and sequester yourself until you feel better if you're sick.

About the Dataset

This dataset was originated from Kaggle website and the link of the dataset is https://www.kaggle.com/datasets/meirnizri/covid19-dataset.it comprises of a large amount of anonymized patient data, including preconditions, all are contained in this dataset. the total of 1,048,576 distinct patients and 21 unique features makes up the raw dataset.

The dataset includes information on the number of people who have received the vaccine at least once in general, the number of new doses given on a given date, the total number of doses distributed nationwide, the percentage of the

population who has finished the entire vaccination series, the total number of Pfizer and Moderna vaccine doses given in each state, and seven-day rolling averages of newly administered and distributed dosages, among other things.

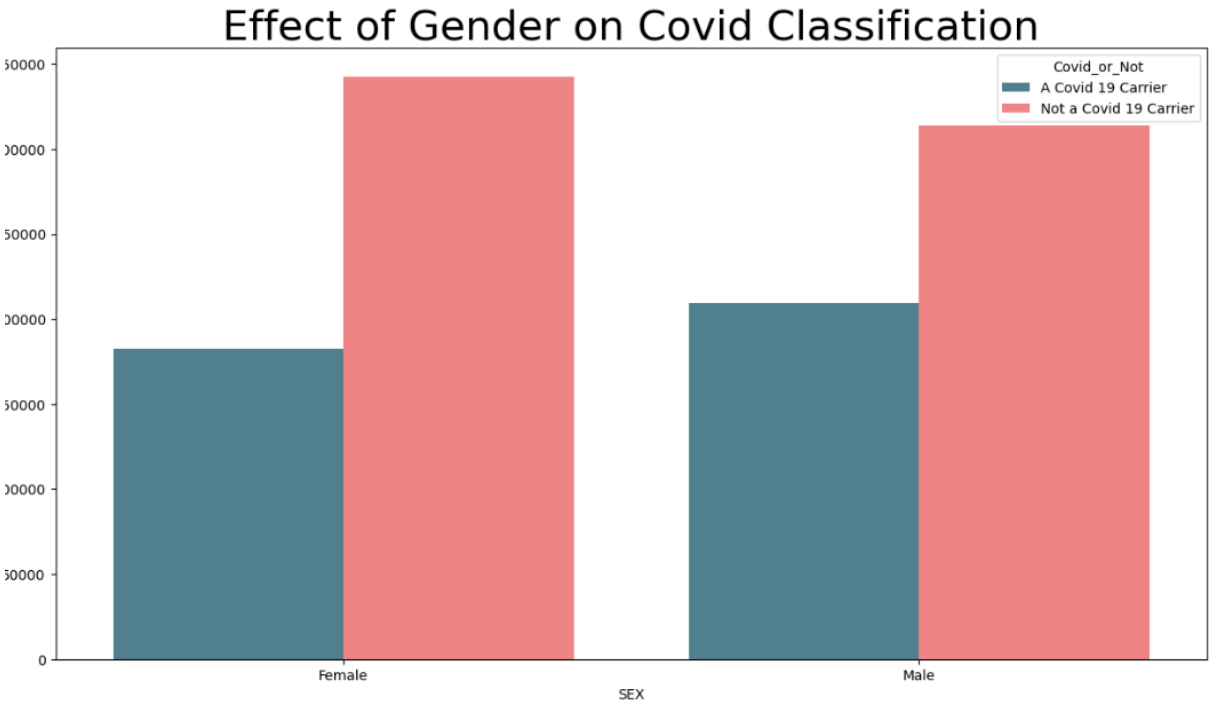
Data Pre-processing

The primary goal of using Deep Transfer-learning models in the majority of covid classification tasks is to reduce the computational complexity. Data augmentation is used to expand the number of training sets to many COVID-19 samples. In addition, the testing data of the non-COVID-19 including various photos fall into 21 subcategories, including pneumonia, asthma, icu, pregnancy, and so on. Thus, some data preparation approaches on date of death and others of infected and noninfected individuals were analysed in this direction-focused study, and the impact of these procedures on COVID-19 diagnosis was investigated.

	mean	std	min	25%	50%	75%	max
USMER	1.632	0.482	1.0	1.0	2.0	2.0	2.0
MEDICAL_UNIT	8.981	3.723	1.0	4.0	12.0	12.0	13.0
SEX	1.499	0.500	1.0	1.0	1.0	2.0	2.0
PATIENT_TYPE	1.191	0.393	1.0	1.0	1.0	1.0	2.0
INTUBED	79.523	36.869	1.0	97.0	97.0	97.0	99.0
PNEUMONIA	3.347	11.913	1.0	2.0	2.0	2.0	99.0
AGE	41.794	16.907	0.0	30.0	40.0	53.0	121.0
PREGNANT	49.766	47.511	1.0	2.0	97.0	97.0	98.0
DIABETES	2.186	5.424	1.0	2.0	2.0	2.0	98.0
COPD	2.261	5.132	1.0	2.0	2.0	2.0	98.0
ASTHMA	2.243	5.114	1.0	2.0	2.0	2.0	98.0
INMSUPR	2.298	5.463	1.0	2.0	2.0	2.0	98.0
HIPERTENSION	2.129	5.236	1.0	2.0	2.0	2.0	98.0
OTHER_DISEASE	2.435	6.647	1.0	2.0	2.0	2.0	98.0
CARDIOVASCULAR	2.262	5.195	1.0	2.0	2.0	2.0	98.0

Encoding categorical values: Using categorical data directly is not how machine learning algorithms operate. This makes it possible for machine learning models

to accurately convert categorical input into their numerical counterparts. In this study, the one-hot encoding approach was used to convert the categorical gender feature in the dataset into binary number representation. Given that it is intended to differentiate between the male and female genders.



Exploratory Data Analysis

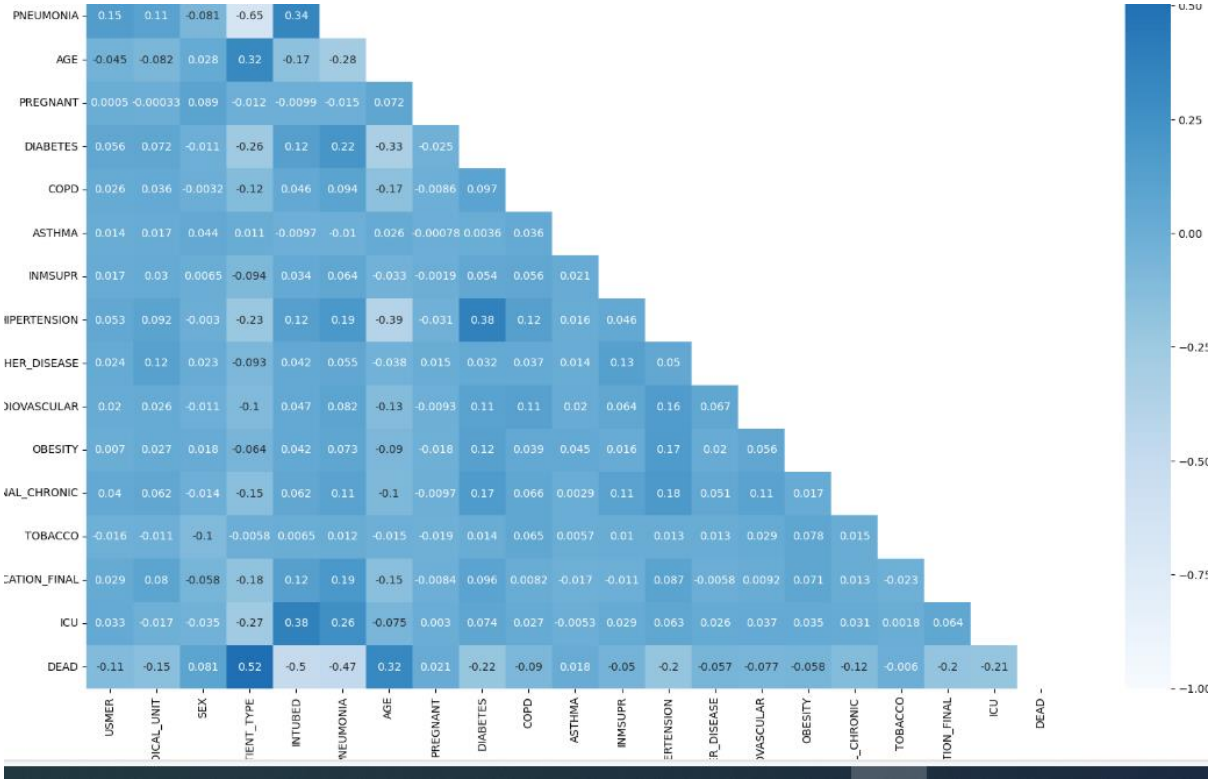
Exploratory Data Analysis (EDA) helps us comprehend trends and patterns by giving us insight into the properties of the data. We loaded the required libraries, including Seaborn, Pandas, and NumPy, which provide the chance to run exploratory studies like value counts to find the frequency of the categorical variables. We can understand the dataset's imbalanced nature thanks to our EDA, as the figure below illustrates.

9] :

	PNEUMONIA	PREGNANT	DIABETES	COPD	ASTHMA	INMSUPR	HIPERTENSION	OTHER_DISEASE
0	1.0	2.0	2.0	2.0	2.0	2.0	1.0	2.0
1	1.0	2.0	2.0	2.0	2.0	2.0	1.0	2.0
2	2.0	2.0	1.0	2.0	2.0	2.0	2.0	2.0
3	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0
4	2.0	2.0	1.0	2.0	2.0	2.0	1.0	2.0
...	...	...	...	...	...	...	...	...
1048570	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0
1048571	2.0	2.0	2.0	2.0	2.0	2.0	1.0	2.0
1048572	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0
1048573	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0
1048574	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0

1048575 rows × 13 columns

We were able to ascertain that the majority of COVID-19 virus-infected individuals will suffer from mild to severe respiratory illness and recover without the need for special care. Serious sickness is more likely to strike the elderly and those with underlying medical conditions such diabetes, pregnancy, cardiovascular disease, and chronic respiratory diseases.



Handling Missing Values: We had to deal with the missing values in the dataset for our machine learning models to be able to manage our dataset and to avoid errors. In our dataset, there is only one missing value the date_died out of 21 columns. Other than repeated values, the column has the largest percentage of missing values. Missing values must be handled carefully since they have the potential to skew our machine learning models' results. Since the numerical data were not uniformly distributed, imputation was used to manage the missing values. Text data were replaced with mode, which enhances the dataset's overall quality, and numerical data were replaced with the median.

Handling Outliers: The EDA provides us with useful insight into the COVID data set, enabling us to comprehend that the dataset's numerical columns contain no outliers.

Feature scaling: as an input variable, each feature in a dataset might have a variety of scales (units). These scale variations in the input variables may result in several issues when the machine learning model is developed. Making ensuring that all features are around the same size helps to reduce potential issues. This is the goal of feature scaling. As a result of feature scaling, every feature gain equal weight, which facilitates processing by machine learning algorithms.

Model Implementation:

Four distinct machine learning methods were used in this research project: Decision Tree (dt), Random Forest Classifier (rf), Gaussian Naïve Bayes (gnb), and Logistic Regression (lr). All the selected algorithms are tried-and-true methods that have been applied extensively to binary classification for many years.

Logistic Regression (LR): is most appropriate in situations where the dependent variable's probabilities fall into one of the two groups. We choose logistic regression for our binary classification challenge because it yields a probability output that can be thresholder to produce binary predictions. It is also an algorithm that is easy to comprehend and straightforward.

The Gaussian Nave Bayes (GNB): algorithm is a widely used technique for binary classification tasks. All that this algorithm does is presume that every characteristic is independent. The approach works effectively on our datasets despite occasionally overgeneralizing because the characteristics are rather independent.

Random Forest Classifier: This technique generates predictions based on several decision trees by utilising an ensemble learning approach. This was taken into consideration for our research because it supports numerical and categorical data, regression analysis, and can be less likely to experience overfitting of algorithms

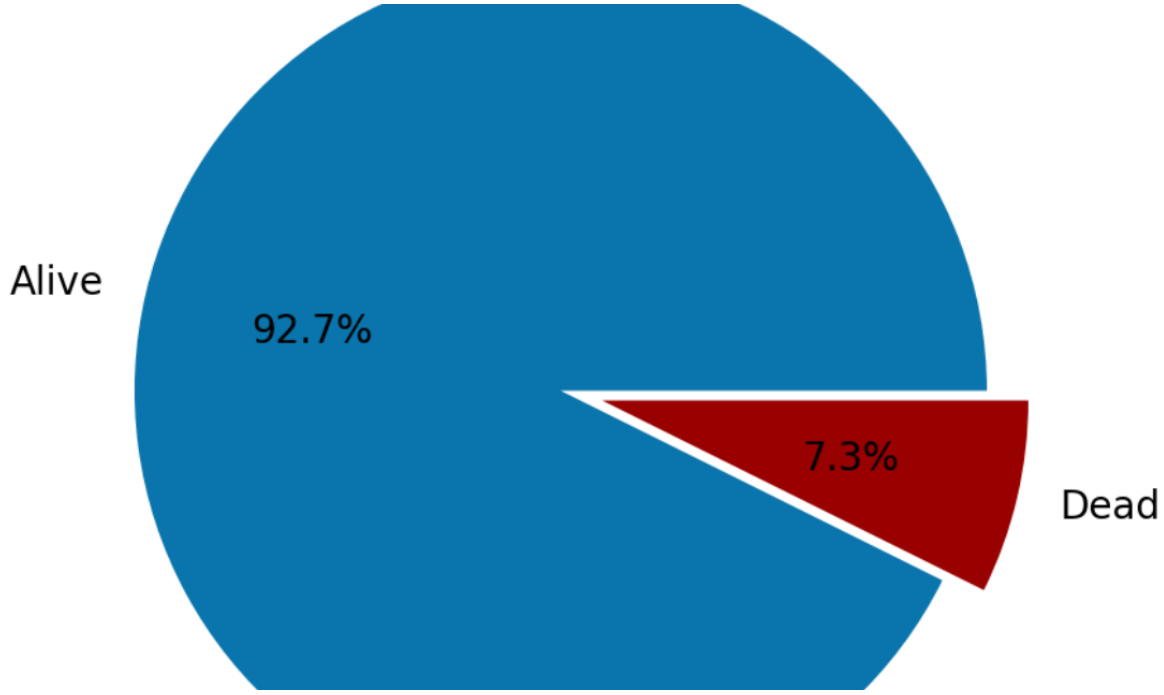
Decision Tree: This algorithm divides the data according to the feature values, producing a structure with the shape of a tree in the process. It is a binary classification algorithm. This algorithm was taken into consideration for our research because to its simplicity in interpretation and visualisation.

Best Model?

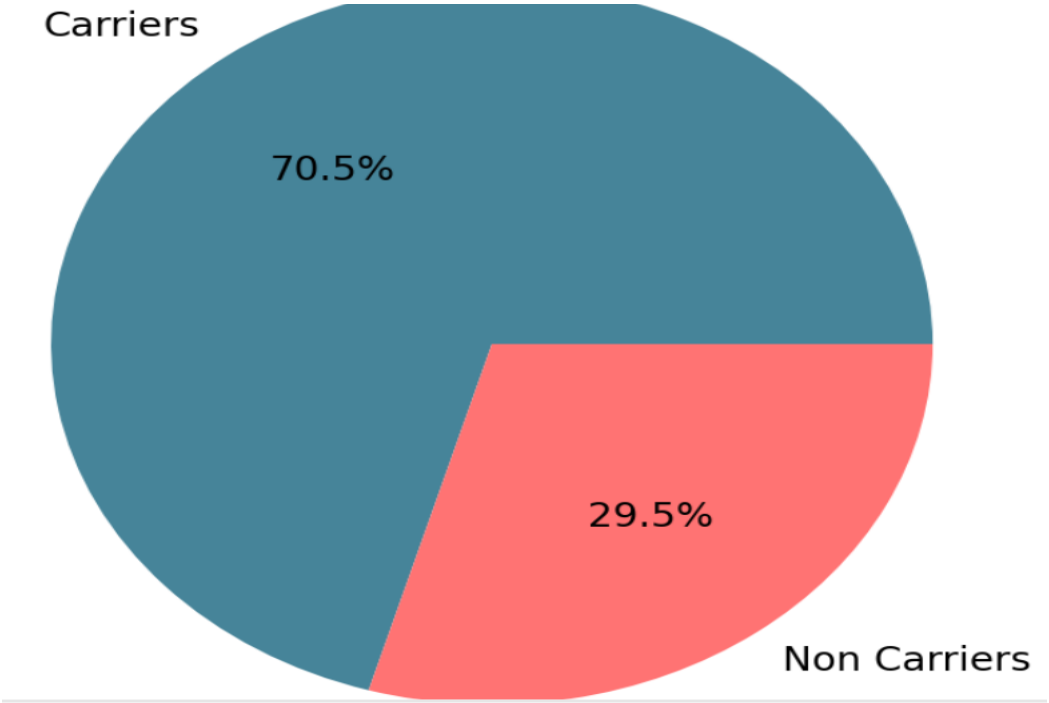
Although both the Decision Tree and Random Forest algorithms produced the highest accuracy (about 92%) as can be seen from the results, the recall in the "Dead" class was extremely poor. Because we are most concerned with our "false negatives" (predicting a patient to be at no risk when in fact they are at high risk), which leads to type 2 error, our judgement will be based on the recall, which is where those algorithms fall short due to their deceptive accuracy.

Thus, when considering accuracy and recall simultaneously, the "Logistic

Regression" model proved to be the most effective, exhibiting almost 90% accuracy, the lowest recall in the "Not dead" class, and somewhat greater recall in the "Dead" class(about 91%).



The "Naive Bayes" algorithm, which has the lowest recall of 88% across both classes and an accuracy of 88%, comes next.



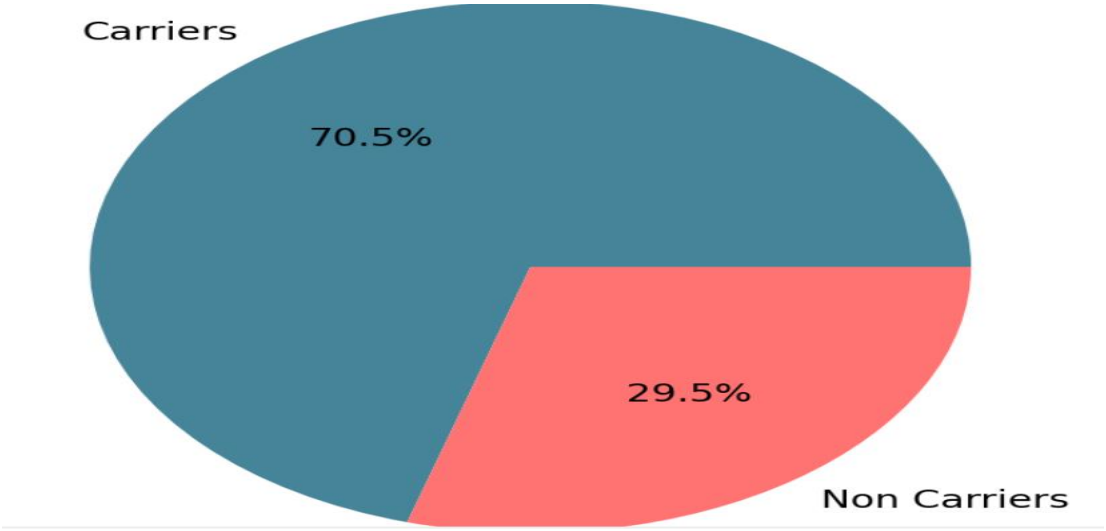
Additionally, we can see that we gave up on precision and f1 score, which performed poorly in all the reports. Since under sampling will lower the amount

of noise in our data, we probably could have gotten better results if we had under sampled instead of oversampled our train dataset. However, doing so might have led to the loss of a significant amount of data, compromising its integrity.

Results

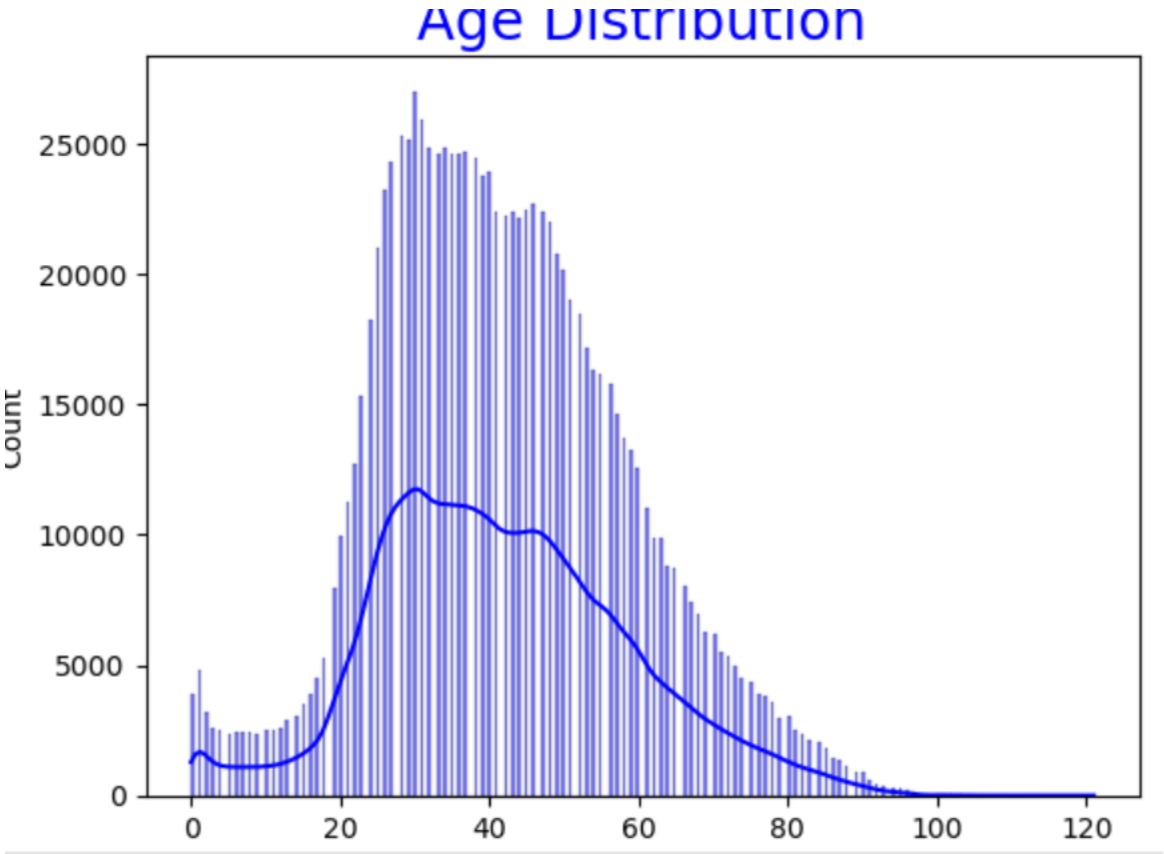
To validate the outcomes of covid 19 incidence, the dataset was properly trained for proper analysis:

Firstly, it is evident that 1,046,683 individuals with comparable diseases are divided into several cohorts of patients with comparable conditions. Additionally, 76,942 patients or 7.3% of the total number of patients in our dataset have passed away. It is evident that all the deceased patients' "Classification" values, ranging from 1 to 7, indicate that some of the deceased patients did not have a Covid-19 diagnosis. Consequently, we can conclude that the deceased patients' outcomes differed from those of most of them, who received a Covid-19 diagnosis. 54,236 patients, or 70.5% of the total, were found to be Covid 19 carriers among the deceased. That illustrates the horrific result of death rate cause by the pandemic.

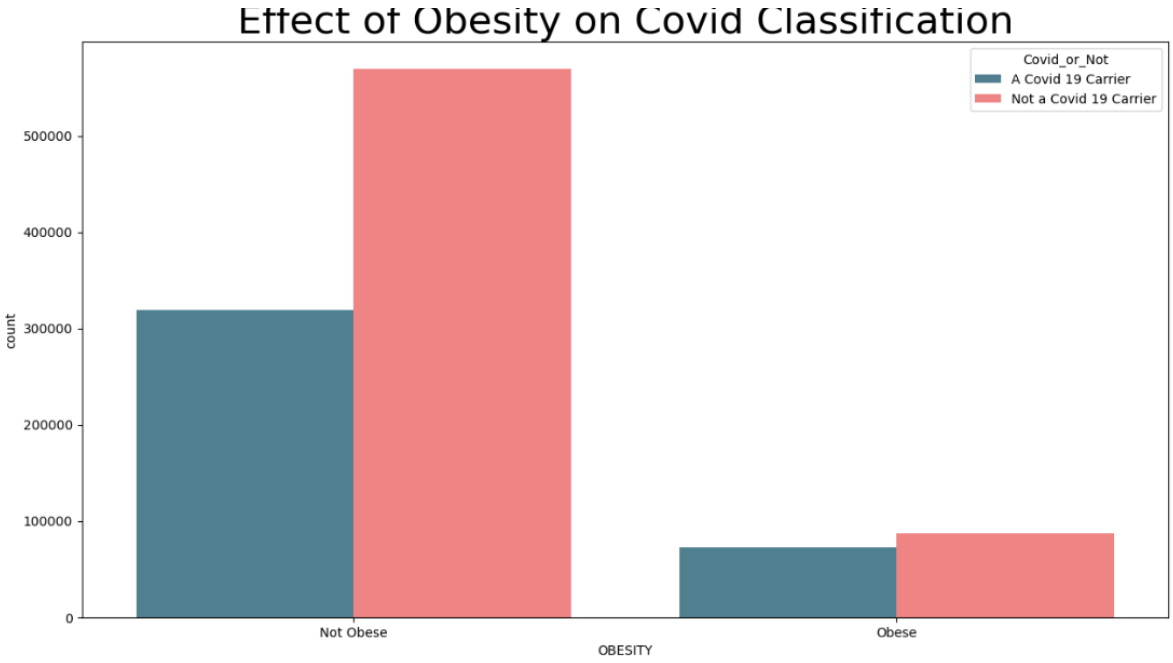


Moreover, with a percentage of 37.4%, we can observe that 391,979 patients overall were Covid 19 carriers. This suggests that most of the patients did not

carry the Covid 19 virus. As we can see, just 13.8% of all Covid 19 carriers were among the 54,236 patients who passed away. This indicates that 86.2% of patients received successful treatment, and to some extent, treatment was proceeding well back then in the world. Additionally, according to records, the bulk consists of individuals in their 20s to late 50s.



Another factor that was considered is obese people and the chances of been exposed to covid-19. Although they are still regarded as a minority, patients with obesity show that the ratio is very close, indicating that a relatively high percentage of patients have the disease. In contrast, the ratio for patients without obesity is approximately 1:2, meaning that only half of every 100 patients without obesity have the potential to have Covid. Therefore, our study indicates that carrying COVID-19 increases the risk for obesity.



Age was another component examined in this investigation, and the results show that gender has little bearing on the situation, although men are marginally more likely than women to possess the virus, but this doesn't matter.

Furthermore, we noticed that the following diseases and habits has the highest impact on covid-19 namely: Pneumonia, hypertension, diabetes, tobacco usage. Above all we concluded that pneumonia patents are more to contact covid.

Conclusion

Unfortunately, 7.3% of the patients have passed on, but roughly 70.5% were Covid Carriers. Out of all patients, carriers made up roughly 37.5% of the total and of those carriers, 14% or so have passed away. We also discovered that ageing significantly affects the situation because it raises the risk of contracting the virus. Additionally, we discovered that the virus is more likely to be carried by obese individuals. Regarding our modelling, we observed that the "Decision Tree" and "Random Forest" algorithms yielded the best accuracy of 92%; however, upon verifying the recall, we discovered that these accuracy levels are false. As a result, we decided to use "Logistic Regression," which scored highly

in terms of accuracy and recall (around 90%). Pregnancy-wise, we were unable to classify the impacts on covid.

In conclusion, we observed that individuals with "pneumonia," "hypertension," "diabetes," and tobacco usage have a higher risk of contracting the virus, with the most people having "pneumonia." In addition, we found that there is a good association between the disease's "hypertension" and "diabetes," since most people who have one of those conditions also have the other. We observed that, of all the patients with these conditions, those identified as having the third degree of COVID-19 are by far the most. Furthermore, we were unable to determine whether pregnancy had any bearing on the Covid classification.

Reference

Mahesh, B., 2020. Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, *9*(1), pp.381-386.

Singh, A., Thakur, N. and Sharma, A., 2016, March. A review of supervised machine learning algorithms. In *2016 3rd international conference on computing for sustainable global development (INDIACom)* (pp. 1310-1315). Ieee.

Sarker, I.H., 2021. Machine learning: Algorithms, real-world applications and research directions. *SN computer science*, *2*(3), p.160.

Sharun, K., Dhama, K., Pawde, A.M., Gortázar, C., Tiwari, R., Bonilla-Aldana, D.K., Rodriguez-Morales, A.J., de la Fuente, J., Michalak, I. and Attia, Y.A., 2021. SARS-CoV-2 in animals: potential for unknown reservoir hosts and public health implications. *Veterinary Quarterly*, *41*(1), pp.181-201.

Chavda, V.P., Vuppu, S., Mishra, T., Kamaraj, S., Patel, A.B., Sharma, N., and Chen, Z.S., 2022. Recent review of COVID-19 management: Diagnosis, treatment, and vaccination. *Pharmacological Reports*, *74*(6), pp.1120-1148.

Yuki, K., Fujiogi, M. and Koutsogiannaki, S., 2020. COVID-19 pathophysiology: A review. *Clinical immunology*, *215*, p.108427.

Shi, Y., Wang, G., Cai, X.P., Deng, J.W., Zheng, L., Zhu, H.H., Zheng, M., Yang, B. and Chen, Z., 2020. An overview of COVID-19. *Journal of Zhejiang University. Science. B*, *21*(5), p.343.

Dhama, K., Khan, S., Tiwari, R., Sircar, S., Bhat, S., Malik, Y.S., Singh, K.P., Chaicumpa, W., Bonilla-Aldana, D.K. and Rodriguez-Morales, A.J., 2020. Coronavirus disease 2019–COVID-19. *Clinical microbiology reviews*, *33*(4), pp.10-1128.

World Health Organization (WHO) (2019) Coronavirus disease (COVID-19) Pandemic. https://www.who.int/emergencies/diseases/novel-coronavirus-2019 [Accessed: April 20, 2024].

World Health Organization (WHO) (2020a) Novel Coronavirus (2019-nCoV) Situation Report-10: data as reported by 30 January 2020. WHO. https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200130-sitrep-10-ncov.pdf?sfvrsn=d0b2e480_2 [Accessed: April 20, 2024].

World Health Organization (WHO) (2020b) Coronavirus disease 2019 (COVID-19), Situation Repor-11: data as reported by 31 January 2020. WHO. https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200131-sitrep-11-ncov.pdf?sfvrsn=de7c0f7_4 [Accessed: April 20, 2024].

World Health Organization (WHO) (2020c) Coronavirus disease 2019 (COVID-19), Situation Report-12: data as reported by 01 February 2020. WHO. https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200201-sitrep-12-ncov.pdf?sfvrsn=273c5d35_2 [Accessed: April 20, 2024].