IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Machine learning models for analysis, evaluation and prediction of the covid-19 dataset.

Umar Mohammed Abatcha, OUEDRAOGO Pengdwende Leonel Camille, ALEMNGE Nadine

Abstract

The aim of our study is to analyze the current spread of the COVID-19 in the world and build a predictive system for the future evolution of this disease based on specific parameters. Especially since its existence has greatly affected current living conditions worldwide. The methods used in this research include data collection, data cleaning, and the transformation of data using supervised learning methods, model optimization, and visualization of prediction. Jupiter Notebook is used for cleaning and pre-processing and other supervised learning techniques are used for training and forecasting. This work makes use of four different models and the most preferment is used to derive the predictive system. The results display a strong correlation between parameters like diabetes, cardiovascular diseases, and age groups with the spread of the coronavirus. Other parameters which might also have some inverse effect with the spread of the virus are identified after exploratory data analysis. Thus the conclusions drawn reveal a high likelihood of individuals between the ages of 60-75 to get the coronavirus as well as individuals with cardiovascular diseases and diabetes. The results of this research also revealed the random forest model as the most per formant amongst the four used. Recommendations made include increased awareness on the importance of vaccination as it will help curb the spread, further in-depth research on the possible relation between parameters like extreme poverty and life expectancy on the spread of the coronavirus.

Key words: COVID-19, SARS-COV-2, Random Forest, AdaBoost, KNeighbors, Support Vector Machine(SVM).

1. Introduction

The outbreak of the SARS-COV-2 in December 2019 was what triggered the spread ofcoronavirus. Prior to this outbreak coronavirus had long existed with over six different types. The world health organization identified the SARS-COV-2 in 2020 as a new type of coronavirus and this outbreak was quickly spread in different parts of the world. SARS-COV-2 is designated as a severe acute respiratory syndrome and due to its widespread use, there have been several mutations hence new Variants emerging.

Since the outbreak's spread, the globe has been driven into despair and terror, with corporate and educational sectors shutting down on a regular basis. Both of these factors have greatly impeded people's quality of life by affecting their source of income. Human contact spreads the virus, which causes symptoms such as cough, fever, body pains, runny nose, loss of smell, and, in the worst-case scenario, diarrhea or breathing difficulties, but healthy carriers are occasionally kept symptom-free. Researchers have proposed social separation, wearing masks, and hand washing as ways to slow the spread of the virus thus far.

This work focuses on deriving a predictive system for the spread of coronavirus based on the different continents. Data is scraped from the Kaggle website to derive the datasets and methods of data cleaning are implored to restructure the dataset appropriately. The findings made are used for establishing this predictive system. This project also makes use of four models namely; Random Forest, AdaBoost, KNeighbors, and Support Vector Machine and chooses the model with the best result. Of all the models, Random Forest proved to be the most preferment.

The main objective or goal is that of establishing a predictive system for the spread of coronavirus in the world and in function to different pertinent parameters. The parameters chosen are after conducting an exploratory data analysis and observing the relationship between the different continents and features. Based on the findings, age, cardiovascular disease and diabetes proved to be some of the most intriguing features.

In terms of limitations, we faced difficulties at the level of timeframe. The time frame during which this project was done was relatively short and not much time available for exploratory data analysis of all features. Also, the dataset used contained statistics on coronavirus valid till the month of October. However, recently there's been a spread in the Omicron variant which hasn't been included in this work.

2. Literature review

Supervised learning is aimed at predicting something from a dataset and it usually involves training a set of output and input variables fed to an algorithm, to produce target results. There exist different types of models for supervised learning accordingto the regression and classification problems. Examples include linear regression, regression trees, random forest, and support vector machines.

Some of the existing literature focused on the prediction of coronavirus is reviewed here in the paragraphs that follow. Kolla Bhanu Prakash and Mohammed Ismail proposed the Random Forest Regressor and Random Forest Classifier to analyze, predict and evaluate the COVID-19 disease. Their objective was based on the analysis of the COVID-19 data to understand which age group is most affected by COVID-19. For their research, they imported the dataset from the Kaggle website. Compare to SVM models, KNN+NCA, Decision Tree Classifier, Gaussian Naïve Bayes Classifier, Multilinear Regression, Logistic Regression, and XGBoost Classifier, Random Forest Classifier and Random Forest Regressor obtained the best results.

Ramesh Kumar Mojjada, Arvind Yadav, A.V. Prabhu et al. proposed the linear regression model to predict the future of covid-19. The main objective of their study was to provide the World Health Organization with a tool for early prediction of the spread of the new coronavirus known as SARS-CoV-2. They imported a dataset from the GitHub registry, provided by the Center for Systems Science and Engineering at Johns Hopkins University. For their research, they used linear regression, LASSO regression, support vector machine, and exponential smoothing. Among these models, linear regression gave a better result.

Vartika Bhadana, Anand Singh Jalal, Pooja Pathak conducted a comparative study on standard machine learning models and proposed six models such as Least Absolute Shrinkage and Selection Operation (LASSO), Random Forest, Decision Tree Regressor, Linear regression, Support vector machine, Polynomial regression. Their objective was to see which model will give the best prediction. Their experience showed that linear regression and LASSO gave the best results. To conduct this study, they imported the dataset from API. covidl9india.org which is the official website of India.

Yue Gao, Guang-Yao Cai, Wei Fang, et al. proposed a mortality risk prediction model for COVID-19 (MRPMC). This algorithm takes patients' clinical data at admission to classify them by mortality risk, and it can predict physiological decline and death up to 20 days in advance. To do this, they used four machine learning methods such as logistic regression, support vector machine, gradient boosting decision tree, and neural network. Their goal was to create a mortality risk prediction model for COVID-19.

Considering the increasing threats posed by the coronavirus, and its widespread evenmore research and predictions are bound to be done. Especially with the rise of Variants like the Omicron, predictions on its spread and pertinent features inversely related to its spread are very necessary for deciding strategies to curb its spread. There is of course a humongous amount of data generated daily with the

existence of the coronavirus and this data needs to be studied extensively for accurate predictions and decision making. In all, the importance of these literary works can not be overemphasized as they help provide answers to the unknown, fill gaps in knowledge, and promote informed decision-making.

3. Methodology

Research work is appraised based on quality, reliability and accuracy of the methodology based on which it was conducted and the analysis of information provided at the end. This section looks at how data was gathered for the research.

3.1 Material

With respect to hardware, I made use of;

- Laptop: To fulfil all programming and research needs.

Whilst software components used were;

- Jupyter notebook: to read dataset, clean and pre-process
- Google colab and Jupyter notebook: for coding purposes
- Google chrome: for research, resolving errors and findings.

3.2 Methods

The methodology implored hinges on the following: data collection, data cleaning, transforming data into supervised learning, scaling data, splitting dataset, model optimisation and training, deep learning model, evaluating the model and visualising predictions.

3.2.1 Data collection

The dataset used is obtained from Kaggle.com and this dataset includes records of total cases, new cases and deaths recorded in the different continents. The records in this dataset are updated daily. For each continent, data includes cases recorded up until 19th October 2021.

3.2.2 Data cleaning

This is done in the Jupyter Notebook and it involves dropping irrelevant data such as iso-code, tests_units to narrow down information to targets only. The Seaborn library, particularly the heatmap, was used to display features with null values. Features withless than 70% null values were retained and more than that was dropped.

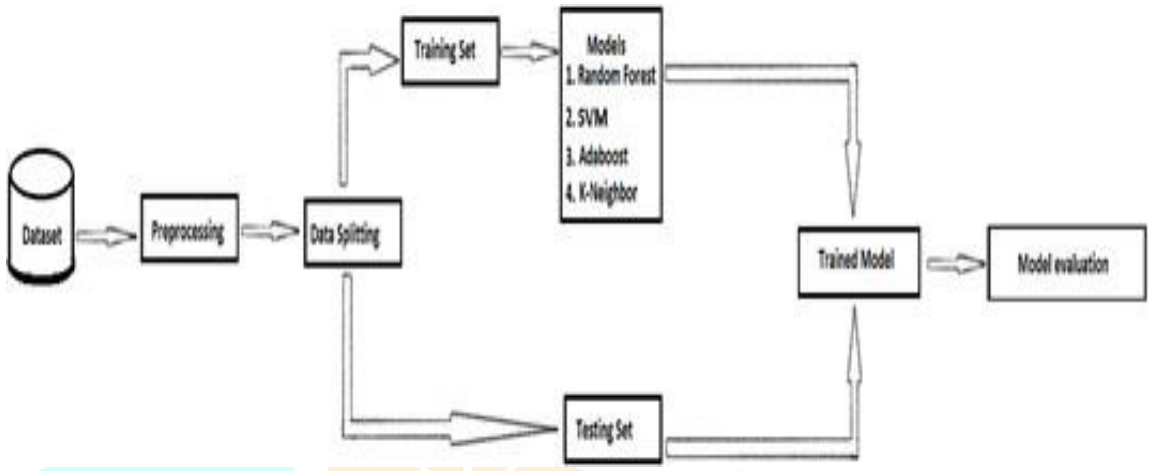

figure 0: Descriptive diagram

3.2.3 Transforming data into supervised learning

This is the first step in the data preprocessing process. Because the chosen models are still trained by supervised learning, we divide the dataset into input samples (x) and targets (y). Before we use our data to train machine learning models, we'll use Label Encoder to convert the target variable to a numerical form. The train test split method is then used to divide our data into two groups: the train set and the test set. The training set was set to 60%, while the testing set was set at 40%.

3.2.4 Supervised learning model

Random forest

This is a test that evaluates various decision trees for a variety of datasets, sub-samples, and averages the results to improve prediction accuracy. The RF regression algorithm is a set learning algorithm that incorporates a wide range of regression trees. A regression tree is a set of hierarchical criteria and constraints that stretch from the root to the leaf of the tree.

A support vector machine (SVM)

It is a supervised machine learning technique used for regression and classification. As a non-parametric approach, SVM regression relies on a collection of mathematical functions. Kernel is a group of functions that converts data inputs into the desired format. Because SVM handles regression issues using a linear function, it translates the input vector(x) to an n-dimensional space termed a feature space when dealing with non-linear regression problems (z). After applying linear regression to space, non-linear mapping techniques are used to complete the mapping. Using a multivariate training dataset (xn) with N number of observations and yn as a collection of observed responses, we can put the notion into ML context. The representation of the linear function is:

$f(x)=x'a+b;$

The idea is to make it as flat as possible, and therefore obtain the value of f(x) using (a′a) as the minimum norm value. As a result, the issue falls into the minimization function J () =12a′a, with the particular requirement that the values of all residuals be less than ε, as shown in the equation:

$\forall n:|yn-(x'na+b)|\leq\varepsilon;$

Smoothing on the Exponential

Forecasting is done using data from prior periods in the exponential smoothing family of approaches. As time passes, the effect of previous data observations diminishes exponentially. As a result, the weight allocated to various lag values decreases exponentially. ES is a sophisticated time series forecasting approach for univariate data that is relatively simple to use. In ES, the forecast for the current time (Ft) is:

$Ft=\alpha At-1+(1-\alpha)\ Ft-1;$

Smoothing cost, where $0\leq\alpha\leq1$ is the actual value of the preceding period in the time series, At-1 is the forecast value of the previous forecast, and Ft-1 is the forecast value of the previous forecast.

KNeighborsClassifier

A k-nearest neighbors' algorithm, abbreviated as KNN, is a data categorization method that calculates how probable a data point is to belong to one of two groups based on which group the data points closest to it belong to.

A k-nearest-neighbor algorithm is a data categorization technique that looks at the data points surrounding it to try to figure out what group a data point belongs to.

When an algorithm examines one point on a grid to see if it belongs to group A or B, it looks at the states of the points nearby. The range is chosen at random, although the goal is to get a sample of the data. If the bulk of the points is in group A, the data point in question is likely to be in group A rather than B, and vice versa.

AdaBoost

It is a statistical classification meta-algorithm that refers to adaptive boosting.
To improve performance, it can be used with a variety of different learning methods. AdaBoost is a method for learning a boosted classifier that is unique. A boosted classifier is a classifier that looks like this:

$$G_T(x) = \sum_{t=1}^{T} g_t(x)$$

And each g is a weak learner that receives an object x as input and returns a valuethat indicates the class of the object..

For each sample in train set, each weak learner generates an output hypothesis, $h(x_i)$. The resulting cumulative learning error E_t of the resultant t-stage boost classifier is reduced by selecting a weak learner and assigning a coefficient α_t at each iteration t.

$$E_t = \sum [F_{t-1}(x_i) + \alpha_t h(x_i)]$$

t

t

$F_{t-1}(x)$ represents the boosted classifier that was created up to the previous training step, E(F) represents some error function, and $g_t(x)=\alpha_t h(x_i)$ represents the weak

learner that is considered for inclusion in the final classifier.

3.2.5 Evaluating the model

In evaluating the model, first, we made use of the learning curve function from the scikit-learn package to see whether the prediction is under-fitting or overfitting. Then a function was written which took into consideration all the four supervised learning models. The metrics used in this function include; f1_score, confusion matrix, and classification reports. All of which is necessary to display precision, recall, and f1_score.

3.2.6 Visualising predictions

The matplotlib library is used for visualisation that is for plotting graphs to show the predictions made and enable proper interpretation.

4. Results

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split
from sklearn.compose import make_column_transformer, make_column_selector
from sklearn.linear_model import SGDClassifier
from sklearn.pipeline import make_pipeline
from sklearn.impute import SimpleImputer
from sklearn.metrics import f1_score, confusion_matrix, classification_report
from sklearn.model_selection import learning_curve
from sklearn.tree import DecisionTreeClassifier
from sklearn.preprocessing import PolynomialFeatures, StandardScaler
from sklearn.ensemble import RandomForestClassifier, AdaBoostClassifier
from sklearn.feature_selection import SelectKBest, f_classif
from sklearn.svm import SVC
from sklearn.neighbors import KNeighborsClassifier
```

Figure 1: Importing libraries

```
In [9]: df=pd.read_csv('covid.csv')

In [10]: df=df.drop(['iso_code','location','tests_units','last_updated_date'], axis=1)

In [11]: df.head()
```

Out[11]:

	continent	total_cases	new_cases	new_cases_smoothed	total_deaths	new_deaths	new_deaths_smoothed	total_cases_per_r
0	Asia	155801.0	25.0	28.857	7247.0	1.0	2.429	39
1	NaN	8444411.0	6054.0	6225.429	215909.0	240.0	215.286	61
2	Europe	178804.0	616.0	448.571	2841.0	12.0	8.286	622
3	Africa	205453.0	89.0	94.714	5875.0	2.0	2.286	46
4	Europe	15369.0	2.0	8.857	130.0	0.0	0.000	1986

5 rows × 61 columns

```
In [12]: df.shape
Out[12]: (224, 61)
```

Figure 2: Overview of dataset and shape in python

```
In [22]: df.describe()
```

Out[22]:

	total_cases	new_cases	new_cases_smoothed	total_deaths	new_deaths	new_deaths_smoothed	total_cases_per_million	new_cases_per_million	r
count	2.020000e+02	202.000000	202.000000	1.950000e+02	195.000000	202.000000	201.000000	201.000000	
mean	3.781025e+06	6952.306931	6424.722772	7.961521e+04	129.487179	105.035361	49419.259965	154.927572	
std	1.940099e+07	36352.302937	33721.318428	3.967416e+05	677.985816	555.644911	49387.023123	317.880580	
min	1.000000e+00	0.000000	0.000000	1.000000e+00	0.000000	0.000000	8.602000	0.000000	
25%	2.200525e+04	10.750000	23.286000	4.625000e+02	0.000000	0.429000	4468.004000	1.126000	
50%	1.826525e+05	188.000000	286.214500	2.920000e+03	4.000000	4.071500	36996.156000	32.249000	
75%	7.999550e+05	1703.500000	1481.321750	1.649700e+04	22.500000	20.821500	82931.492000	162.252000	
max	2.415685e+08	436256.000000	409047.000000	4.913106e+06	8012.000000	6784.429000	222333.434000	2951.283000	

8 rows × 60 columns

Figure 3: statistical measures of the data

Exploratory Data Analysis

This part consists of understanding our data as well as possible and developing a modeling strategy. Our dataset contains both quantitative and qualitative variables.

After an exploratory analysis of the data, we find that in all continents, people aged 65-70 years are more likely to be infected with COVID-19 (see Figures 7 and 8). Figures 6, 7, and 8 illustrate the percentage of cases obtained by age group.

Figures 8 and 9 show that people with heart disease are at greater risk of infection than people with diabetes.

Figure 11 shows the correlation matrix of the dataset.

```
In [25]: sns.countplot(x='total_cases', hue='continent', data=df)
Out[25]: <AxesSubplot:xlabel='total_cases', ylabel='count'>
```

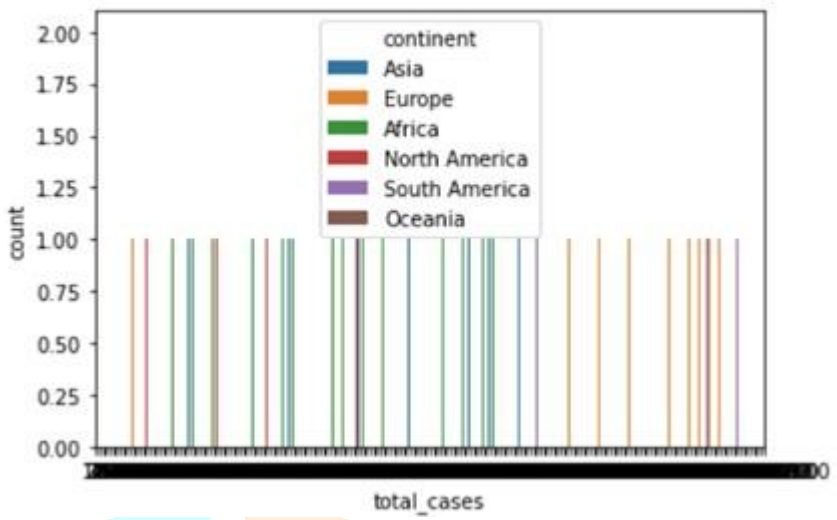

Figure 4: Total cases per continent

```
In [26]: sns.countplot(x='new_cases', hue='continent', data=df)
Out[26]: <AxesSubplot:xlabel='new_cases', ylabel='count'>
```

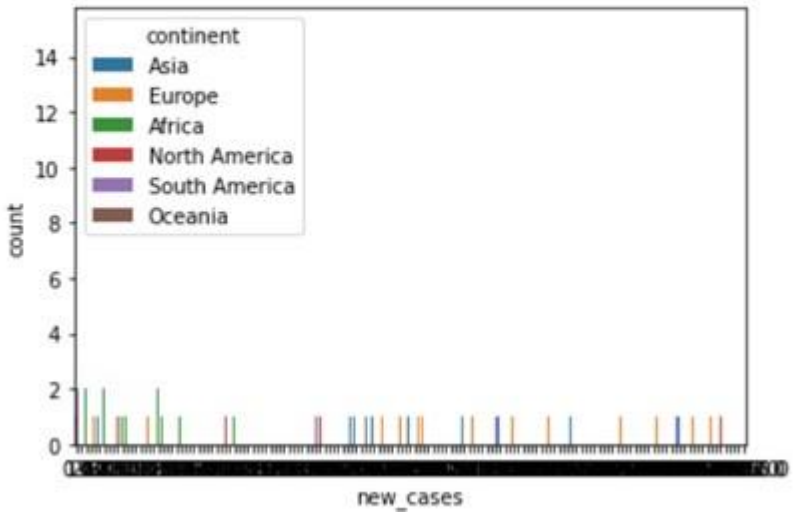

Figure 5: New cases per continent

```
In [27]: sns.countplot(x='median_age', hue='continent', data=df)

Out[27]: <AxesSubplot:xlabel='median_age', ylabel='count'>
```

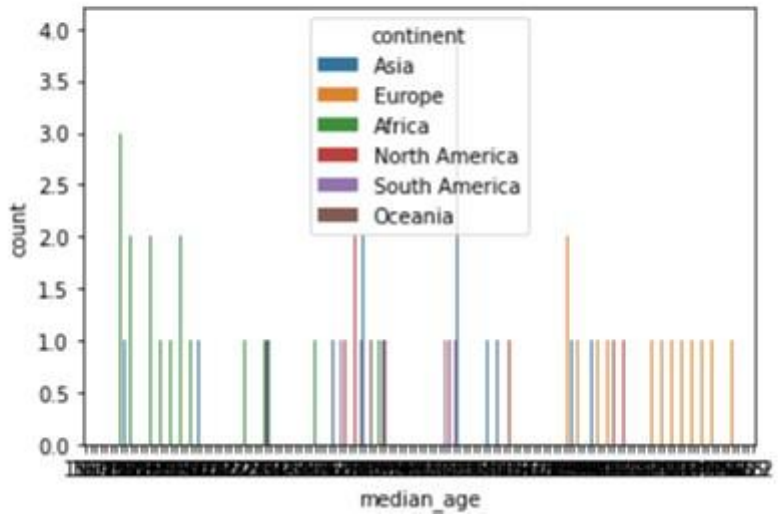

Figure 6: median age per continent

```
In [28]: sns.countplot(x='aged_65_older', hue='continent', data=df)

Out[28]: <AxesSubplot:xlabel='aged_65_older', ylabel='count'>
```

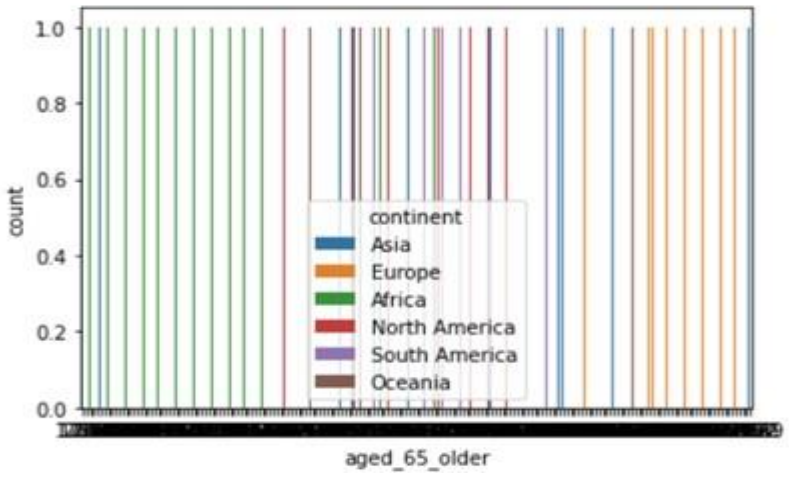

Figure 7: aged 65 older per continent

```
In [29]: sns.countplot(x='aged_70_older', hue='continent', data=df)
Out[29]: <AxesSubplot:xlabel='aged_70_older', ylabel='count'>
```

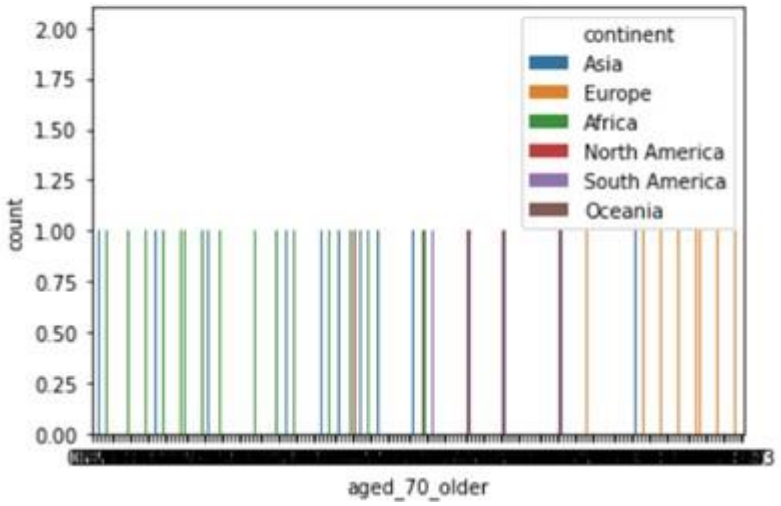

Figure 8: aged 70 older per continent

```
In [30]: sns.countplot(x='cardiovasc_death_rate', hue='continent', data=df)
Out[30]: <AxesSubplot:xlabel='cardiovasc_death_rate', ylabel='count'>
```

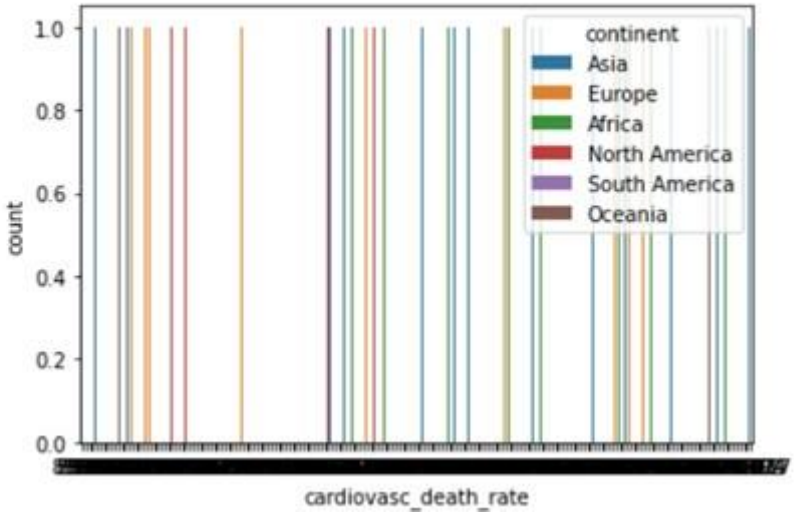

Figure 9: cardiovascular death per continent

```
In [31]: sns.countplot(x='diabetes_prevalence', hue='continent', data=df)

Out[31]: <AxesSubplot:xlabel='diabetes_prevalence', ylabel='count'>
```

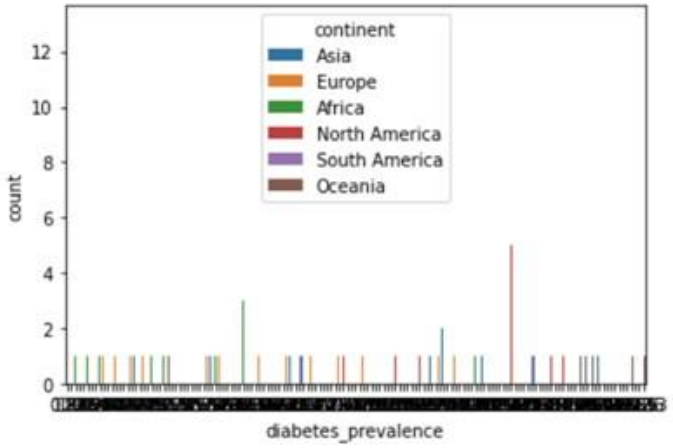

Figure 10: Diabetes prevalence per continent

```
In [42]: plt.figure(figsize=(30,20))
         sns.clustermap(df.drop('continent',axis=1).corr())

Out[42]: <seaborn.matrix.ClusterGrid at 0x1de700a68b0>

         <Figure size 2160x1440 with 0 Axes>
```

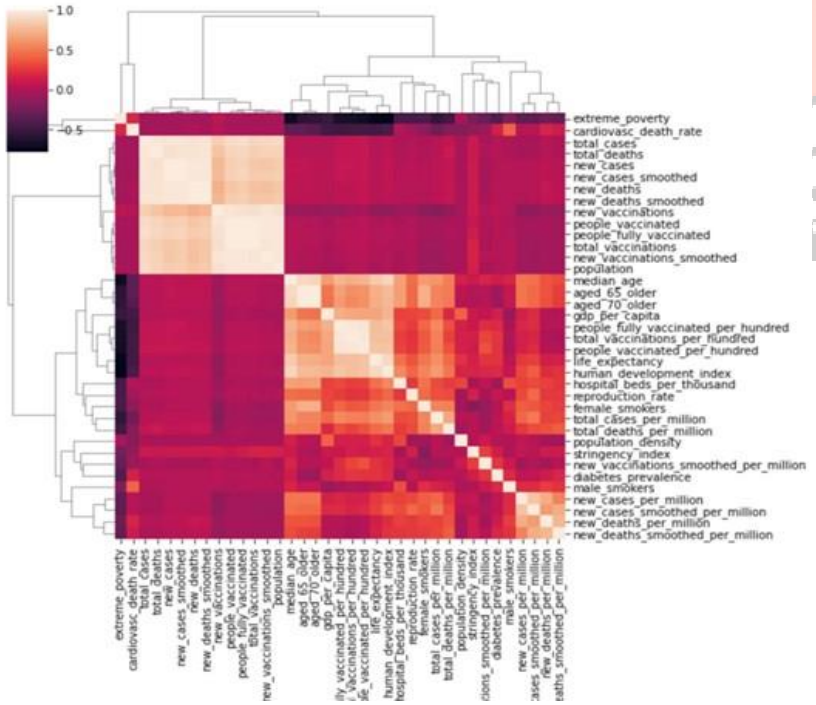

Figure 11: Correlation matrix of the dataset

In our case study, we found that the random forest (Figure 11) gave a better prediction than other models such as Support Vector machine (Figure 13), AdaBoost Classifier (Figure 14) and KNeighborsClassifier (Figure 15).

Figure 12: Random Forest Classifier

Figure 13: Support Vector Machine

Figure 14: AdaBoost Classifier

Figure 15: KNeighborsClassifier

6. Conclusion

Firstly the objective of this project was to analyse, evaluate and predict the evolution and propagation of coronavirus based on specific learning models. Judging from the results obtained and the correlation matrix used to understand relationships, individuals within age groups of 65-70 are more exposed to COVID-19 disease in all continents as well as individuals with cardiovascular diseases and diabetes. This raises questions as to

1) Are these the only parameters which adversely affect the spread ofcoronavirus?
2) Is it only cardiovascular disease and respiratory patients that are highly at riskwith the spread of the coronavirus?
3) The relation between different diseases and their possible effects oncoronavirus.

All of which require further research to be answered.

Giving the findings uncovered in this research, the random forest classifier outperformed other models in terms of accuracy. There is a need to evaluate other machine learning models to derive more accurate predictive systems.

This study revealed an asymmetric relationship between life expectancy, extreme poverty and vaccination rate with the spread of coronavirus in the world. In light of this, the following suggestions are made:

1) Government and health officials should do more to sensitise the population on the importance of vaccines and the benefits procured from immunisation. This can be done through various outreaches and health campaigns.
2) The relation between cases of extreme poverty and their susceptibility to being infected by coronavirus needs to be studied further so as to make informed decisions.
3) Also, the relation between life expectancy and the spread of coronavirus needs to be studied further.

REFERENCE

Menni, C., Valdes, A.M., Freidin, M.B. et al. Real-time tracking of self-reported symptoms to predict potential COVID-19. Nat Med (2020). https://doi.org/10.1038/s41591-020-0916-2.

Kolla Bhanu Prakash et al., "Analysis, Prediction and Evaluation of COVID-19 Datasetsusing Machine Learning Algorithms" International Journal of Emerging Trends in Engineering Research, 8(5), May 2020, 2199 – 2204.

Coronavirusdisease2019(COVID-19),

https://www.mayoclinic.org/diseases-conditions/coronavirus/diagnosis-treatment/drc-20479976.

Yang, P., Wang, X. COVID-19: a new challenge for human beings. Cell Mol Immunol17, 555–557 (2020). https://doi.org/10.1038/s41423-020-0407-x.

Y. Zhang, B. Jiang, J. Yuan, Y. Tao The impact of social distancing and epicenter lockdown on the COVID-19 epidemic in mainland China: a data-driven SEIQR model study medRxiv (2020).

Ramesh Kumar Mojjada, A. Yadav, A.V. Prabhu et al., Machine learning models for covid-19 future forecasting, Materials Today: Proceedings, https://doi.org/10.1016/j.matpr.2020.10.962.

Y. Bai, L. Yao, T. Wei, F. Tian, D.-Y. Jin, L. Chen, M. Wang Presumed asymptomatic carrier transmission of COVID-19 JAMA (2020) [PMC free article] [PubMed].

Y. Bai, L. Yao, T. Wei, F. Tian, D.-Y. Jin, L. Chen, M. Wang Presumed asymptomatic carrier transmission of COVID-19 JAMA (2020) [PMC free article] [PubMed].

Katz, J. N. et al. Disruptive modifications to cardiac critical care delivery during the Covid-19 pandemic: an international perspective. J Am Coll Cardiol. https://doi.org/10.1016/j.jacc.2020.04.029 (2020).

World Health Organization. Coronavirus 2019 (COVID-19) (World Health Organization, 2020). https://covid19.who.int/.

[11] Wang, D. et al. Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus-infected pneumonia in Wuhan, China. JAMA https://doi.org/10.1001/jama.2020.1585 (2020).

[12] Liang, W. et al. Development and validation of a clinical risk score to predict the occurrence of critical illness in hospitalized patients with COVID-19. JAMA Internal Med. https://doi.org/10.1001/jamainternmed.2020.2033 (2020).

[13] Yue, H. et al. Machine learning-based CT radionics method for predicting hospitalstay in patients with pneumonia associated with SARS-CoV-2 infection: a multicenter study. Ann Transl Med 8, 859 (2020).

[14] Gao, Y., Cai, GY., Fang, W. et al. Machine learning based early warning system enables accurate mortality risk prediction for COVID-19. Nat Commun 11, 5033 (2020). https://doi.org/10.1038/s41467-020-18684-2

[15] N.S Punn, S. K Sonbhadra and S. Agarwal, "COVID-19 Epidemic Analysis using Machine Learning and Deep Learning Algorithms", medRxiv preprint, 2020

[16] Z Car, S Baressi Šegota, N Anđelić, I Lorencin and V Mrzljak, "Modelling the Spread of COVID-19 Infection Using a Multilayer Perceptron", Computational and Mathematical Methods in Medicine, vol. 20, pp. 1-10, 2020.

[17] W. Huang, Y. Nakamori and SY Wang, "Forecasting stock market movement direction with support vector machine", Computers & operations research, vol. 32.10, pp. 2513-2522, 2005.

[18] V. Bhadana, A. S. Jalal and P. Pathak, "A Comparative Study of Machine Learning Models for COVID-19 prediction in India," 2020 IEEE 4th Conference on Information & Communication Technology (CICT), 2020, pp. 1-7, doi: 10.1109/CICT51604.2020.9312112.