A MACHINE LEARNING APPROACH FOR PREDICTION OF COVID DEATH RATE USING DESIGN THINKING

Nandhakumar.K[1], Dr.P.Sumathi[2] , Gunasekar*, Jaganathan.P.V*, Jithendran.R*,Steven.J*,

Assistant Professor[1], Student*,
Department of Information Technology,
SNS College of Engineering Coimbatore, India.
sekarguna534@gmail.com, jagajaga97852@gmail.com, jithendranjithu0710@gmail.com,
stevenjagadish@gmail.com.

ABSTRACT :
At the end of 2019, coronavirus (COVID-19) has immediately shown a high rate of transmission, forcing the World Health Organization (WHO) to declare in March 2020 that this unknown coronavirus, named severe acute respiratory syndrome coronavirus 2 (SARS CoV-2), can be characterized as a pandemic. The COVID-19 pandemic has led to a dramatic loss of human life worldwide and presents an unprecedented challenge to public health, food systems and the world of work. The metabolic activity of immune cells is enhanced after a viral infection, such as the one driven by COVID-19. Dietary approaches that support a healthy gut microbiome can benefit the immune system and ensure a good nutritional status that would help the host deal with pathogens. For most, no income means no food, or, at best, less food and less nutritious food. Many researches are being done based on several habits like food, age etc., To predict the COVID - 19 death rate based on dietary habits of 170 countries using machine learning techniques that group the countries together according to the distribution of fat, energy and protein across 23 different types of food as well as amount of kilograms ingested. Results show how obesity and the high consumption of fats appear in countries with the highest death rates, whereas countries with a lower rate have a higher level of cereal consumption accompanied by a lower total average intake of kilocalories. A transformation of the data has been carried out using Principal Component Analysis (PCA). Once the data is reduced, K-Means has been applied with the intention of grouping 170 countries into clusters based on the food consumption. The project is to apply the machine learning concept to analyse the possibility of survival based on the country they live in and if the person is affected by any other diseases. To predict the possibility of the person's survival based on the country he survives using the higher death rate and normal death rate of country clusters and other diseases affected clusters. Based on the cluster using KNN the possibility of survival of the particular person is predicted.

I.INTRODUCTION

In December 2019, the novel coronavirus was discovered in the Wuhan city of China and it was reported to the WHO (World Health Organization) at the end of December 2019. The virus created a threat all over the world and it was named as COVID - 19 by the World Health Organization on 11th February 2020. The COVID - 19 is the family of viruses including Severe Acute Respiratory Syndrome (SARS), Acute Respiratory Distress Syndrome (ARDS). The World Health Organization declared these outbreaks as a public health emergency and mentioned the virus is being transmitted through the respiratory system when a normal healthy person comes in contact with the COVID infected person. According to the World Health Organization, the signs and symptoms of mild to average cases are dry cough, fever and fatigue while as in severe cases dyspnea (shortness of breath), Fever and tiredness may occur. The person having any other disease like diabetes, asthma, liver disease and heart disease are more in danger of the virus and many become severely ill. But in some cases no signs or symptoms are found in the person they are asymptomatic. The asymptomatic nature makes the disease even more dangerous and spreads the diseases to the common people easily. To develop a modified predictive model for mortality rates among people due to coronavirus (covid-19).covid-19 has the highest mortality rate compared to previous pandemic like influenza, Spanish flu and so on furthermore it has the highest death rate compared to that of its predecessor. Various machine learning and deep learning techniques have been implemented to determine the mortality rate among people due to covid-19. Through machine learning we can identify patterns, analyze data and make decisions without any human 2 intervention. Machine learning has broadly classified into three categories : Supervised learning,

Unsupervised learning, Reinforcement learning The first step in a machine learning model is to collect data.various data sets are related to covid-19 is available online and the most widely used data set is from john hopkins university data set on covid-19 posted on Github. Then we have to classify the data based on various factors and many methods are available to classify these data sets some of them are logistic regression, Naive bayes, K-nearest neighbours, Decision tree, Random forest, Support vector machine and so on. Logistic regression is the largely applied machine learning model in the industry.

II.PROPOSED SYSTEM

Our model classifies the country based on being affected by any other disease other than COVID- 19. It classifies the country into a high death rate country and normal   death rate country. Input from the clustering is used to predict the patient's possibility of survival. Particular person's details are given as input, to predict the particular patient's possibility of survival rate based on the factors like person's country, if the person is affected by any other disease etc.,

KNN

K-Nearest Neighbour is the simplest Machine Learning algorithm on Supervised Learning technique. K-NN algorithm finds the similarity between the new data and already available data or case and adds the new data into the category which is more similar to the data. K-NN stores already available categories and classifies the new data or case based on the similarity between data. So whenever new data comes it classifies easily based on a well

suited category. The K-NN algorithm is used both for Regression and Classification. It is a non-parametric algorithm and doesn't make any assumptions on fundamental data. At the training phase K-NN just stores the data and when new data comes it classifies based on category.

III.METHODOLOGY

KNN (K-Nearest Neighbour) is used to classify the data. K-Nearest Neighbour is the simplest Machine Learning algorithm on Supervised Learning technique. By using K-NN, the algorithm finds the similarity between the new data and already available data and adds the new data into the category which is more similar to the data. K-NN stores already available categories and classifies the new data or case based on the similarity between data. So whenever new data comes it classifies easily based on a well suited category. It classifies the data based on the country the person belongs to. At the training phase K-NN just stores the data and when new data comes it classifies based on category.

Data Preprocessing

The data obtained from the dataset is pretty preprocessed. These data are analyzed, and the dataset is generated from this raw data. To use the dataset with the python functions, the data need to be normalized. Normalization is a technique usually applied as part of data preparation for machine learning. The objective of normalization is to change the values of numeric columns in the dataset to use a standard scale without having differences in the ranges of values or losing information. Thus the data present is normalized in the range of 0 to 1. The feature and the label are splitted and made ready for further processing with the model. The data present in the dataset is normalized using MinMaxScalar.

Training and Testing the model

The model is developed and trained in the python programming environment using various libraries like Pandas, Numpy and matplot. The dataset is split into training and validation dataset. Training dataset is used to fit the model, and Validation Dataset is used to provide an unbiased evaluation of a model fit on the training dataset while tuning model Hyperparameters.

Validating the model

Model validation is referred to as the process where a trained model is evaluated with a testing data set. The testing data set is a separate portion of the same dataset from which the training set is derived. The purpose of model validation is to check the accuracy and performance of the model based on the past data for which we already have actuals. The performance of the proposed system has been evaluated using Accuracy.
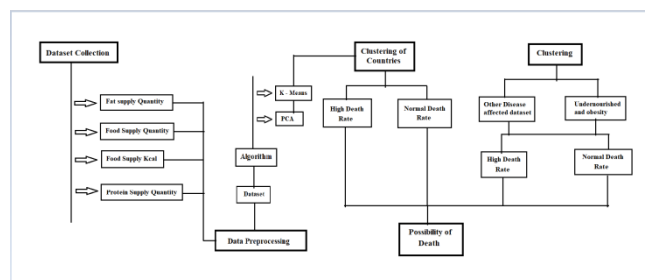
IV.LIBRARIES
Matplotlib

Matplotlib is one of the libraries in Python. It is also a numerical – mathematical extension for the NumPy library. Pyplot is a state-based interface to a Matplotlib module which provides an interface like MATLAB. The imshow() function in the pyplot module of matplotlib library is used to display data as an image; that is on a 2D regular raster.

Scikit-learn

Sklearn is a free open source python library. It is built on NumPy, Matplotlib and SciPy. Scikit-learn is the most used library in machine learning. It contains many efficient tools including Regression like linear and logistic regression, Clustering including K-Means and K-Means++, Classification including K-Nearest Neighbours, dimensionality reduction and processing including Min-Max Normalization.

V. BLOCK DIAGRAM

VI. DATASET

The dataset contains the details of the person affected with any other disease other than COVID - 19. It includes diseases like Diabetes, Chronic Kidney Disease (CKD), Chronic obstructive pulmonary disease (COPD), Obesity and Undernourishment. And also it contains the active cases, confirmed cases, death, and recovered cases are also included.

Cluster	Number of countries
High Death rate	28
Normal Death Rate	142

VII. SYSTEM REQUIREMENTS

HARDWARE REQUIREMENTS

Processor : i3 core
Speed : 2.30 GHz
Hard Disk : 500 GB
RAM : 16 GB

SOFTWARE REQUIREMENTS

Environment : Google Chrome, Google Colaboratory, Operating System (Windows 10).

PYTHON

Python is a general-purpose interpreted, interactive, object-oriented, and high level programming language. Python source code is also available under the General Public License (GPL). It provides constructs that enable clear programming on both small and large scales. Python features a dynamic type system and automatic memory management. It supports multiple programming paradigms, including object-oriented, imperative, functional and
procedural, and has a large and comprehensive standard library. Python is open source software and has a community-based development model. Python is managed by the non-profit Python Software Foundation. The features of python include the below mentioned.

GOOGLE COLABORATORY

Python can be coded in Integrated Development Environment (IDE). There are manyIDE"s like Jupyter, Spyder, PyCharm and Rodeo. All these IDEs run locally in the personal computers. Since the IDE"s are local, they are not portable so that the files which are needed should be installed in every computer. Moreover, the speed of the IDE depends on the speed of the computer whereas Colaboratory (Colab) has many advantages over the IDEs. Colab is a product of Google Research. Colab is a free cloud service based on Jupyter Notebooks which allow you to write python code. It is configured for Deep Learning free of charge to access GPU resources. It does not require any pre-setup configuration.

VIII. RESULT AND DISCUSSION

A time-varying description of certain behavior of the natural system as computed by the mathematical model. Simulation is a set of techniques that uses computers to imitate the operations of various real-world tasks or processes through simulation.

Sample Dataset

In this process, the dataset consists of the details of the person affected by any other disease other than

COVID - 19. With the help of this dataset we can cluster the country into high death rate countries and normal death rate countries. Figure 6.1 describes the sample dataset containing the factors like name of the person, country, confirmed cases, death cases, active cases, and other disease like diabetes, chronic kidney disease, COPD etc.,

Name	Country	Animal fats	Oilcrops	Confirmed	Deaths	Recovered	Active	Diabetes	ckd	copd
Robert	Afghanistan	0.0277	0.2034	0.142134196	0.006185779	0.123373921	0.012574497	Yes	No	Yes
John	Albania	0.0711	0.3722	2.967300916	0.050951374	1.792635659	1.123713883	No	Yes	No
Michael	Algeria	0.0054	0.183	0.244897085	0.006558153	0.167572198	0.070766734	Yes	Yes	No
William	Angola	0.0277	2.1534	0.061687473	0.00146055	0.056907699	0.003419224	No	No	No
David	Antigua and Barbuda	0.1289	0.7674	0.293877551	0.007142857	0.190816327	0.095918367	Yes	Yes	No
Richard	Argentina	0.0097	0.0097	4.356147387	0.108226635	3.905192058	0.342728695	Yes	Yes	Yes
Joseph	Armenia	0.1438	0.2619	5.681224628	0.109345061	5.386410014	0.177469953	No	No	Yes
Thomas	Australia	0.074	0.6994	0.112025316	0.003529549	0.10128912	0.007206648	Yes	No	Yes
Charles	Austria	0.3376	1.5028	4.739982051	0.089679156	4.498870092	0.153432802	No	No	No
Christopher	Azerbaijan	0.0596	0.0541	2.285536209	0.031222794	2.225573803	0.028739612	No	No	No
Daniel	Bahamas	0.0695	0.8052	2.100763359	0.044783715	1.735114504	0.32086514	No	Yes	No
Matthew	Bangladesh	0.0083	0.5972	0.316691106	0.004923066	0.284343586	0.027524454	No	Yes	Yes
Anthony	Barbados	0.112	1.2324	0.583972125	0.006271777	0.470034843	0.107665505	No	Yes	Yes
Mark	Belarus	0.1328	0.5207	2.740696	0.018912	2.612704	0.10928	Yes	No	Yes
Donald	Belgium	0.4678	0.2216	6.286322188	0.185427703	0	6.100864485	Yes	No	Yes
Steven	Belize	0.0893	1.7864	2.872792363	0.073031026	2.739619136	0.060143198	Yes	No	No
Paul	Benin	0.0078	3.2508	0.034343517	0.000450487	0.029183389	0.00470964	No	Yes	No
Andrew	Bolivia	0.0019	0.5114	1.952445629	0.092435313	1.438614287	0.421396029	No	Yes	No
Joshua	Bosnia and Herzegov	0.0633	0.6863	3.76208473	0.145534899	2.960316977	0.656232856	No	Yes	Yes
Kenneth	Botswana	0.1534	1.5572	1.014372033	0.007034959	0.844799309	0.162537764	Yes	Yes	No
Kevin	Brazil	0.1266	0.4788	4.460165146	0.108602912	3.979468585	0.372093649	Yes	No	No
Brian	Bulgaria	0.24	0.7619	3.227255888	0.134416053	2.788956258	0.303883355	Yes	No	No
George	Burkina Faso	0.0086	5.7595	0.053709994	0.000641056	0.048031383	0.005037554	Yes	No	Yes
Edward	Cabo Verde	0.0649	0.3534	2.586330935	0.024280576	2.461151079	0.100899281	No	Yes	No
Ronald	Cambodia	0.0153	4.5722	0.003058657	0	0.002929599	0.000129057	No	No	No
Timothy	Cameroon	0.0144	3.6301	0.118031431	0.001782089	0.110914354	0.005334988	No	No	No
Jason	Canada	0.8868	2.416	2.109960723	0.054202871	1.909848128	NA	Yes	No	No

IX..CONCLUSION

The goal of this project is to predict the person's possibility of survival based on the country they belong to and if the person is affected by any other disease. Covid-19 is being spread rapidly and many technologies have been identified for prediction of survival rate. KNN is the most commonly used classification algorithm. Though many inventions have been made, using KNN algorithms is the best and fastest way to predict the survival rate of people. It stores all the data that is fed through clustering of the data based on country dietary habits and other diseases and classifies the new data, based on similarity in the test data and training data. The purpose of this project is to predict the person's possibility of survival.

X.FUTURE WORK

The future work of this project is to develop a web based application which will be useful for the hospital management system and make it free of cost. And have planned to develop two additional modules, assessment using RNN and IoT. The analysis using RNN module will act as a voice enabled chat bot and ask some questions to the users. It is intelligent to change the questions from the user's mental state and their previous answers. The final module is analysis using IoT which makes use of pulse rate monitors and EEG sensors to detect depression at a very accurate level. This will be very effective and useful for the doctors and also the public.

XI.REFERENCE

1. Severity Detection for the Coronavirus Disease 2019 (COVID-19) Patients Using a Machine Learning Model Based on the Blood and Urine Tests (2020) Haochen Yao1, Nan Zhang2, Ruochi Zhang3, Meiyu Duan3, Tianqi Xie4, Jiahui Pan1, Ejun Peng5, Juanjuan Huang1, Yingli Zhang2, Xiaoming Xu2, Hong Xu2*, Fengfeng Zhou3* and Guoqing Wang1*.

2. A novel specific artificial intelligence-based method to identify COVID-19 cases using simple blood exams (2020).

3. Care of patients with liver disease during the COVID-19 pandemic: EASL-ESCMID position paper (2020) TobiasBoettler, Philip N.Newsome, Mario U.Mondelli, MojcaMaticic, ElisaCordero, MarkusCornberg, ThomasBerg.

4. Use of Real-World Electronic Health Records to Estimate Risk, Risk Factors, and Disparities for COVID-19 in Patients With Cancer (2020) Aakash Desai, MBBS, MPH1; Ali Raza Khaki, MD2; Nicole M. Kuderer, MD.

5. Impact of COVID-19 Outbreak on Cancer Patient Care and Treatment: Data from an Outpatient Oncology Clinic in Lombardy (Italy) (2020) Erica Quaquarini, *ORCID, Giuseppe Saltalamacchia, Daniele Presti, Giulia Caldana, Valentina Tibollo, Alberto Malovini, Raffaella Palumbo, Cristina Maria Teragni, Emanuela Balletti, Ludovica Mollica, Elisa Biscaldi, Mara Frascaroli, Antonio Bernardo and Federico Sottotetti.

6. Active and Effective Measures for the Care of Patients With Cancer During the COVID-19 Spread in China (2020) Zhijie Wang, MD, Jie Wang, MD, PhD1, Jie He, MD 51.

7. COVID-19 Coronavirus Vaccine Design Using Reverse Vaccinology and Machine Learning (2020) Edison Ong, Mei U Wong, Anthony Huffman and Yongqun He.

8. COVID-19 World Vaccination Progress Using Machine Learning Classification Algorithms (2021) Nasiba M. Abdulkareem, Adnan Mohsin Abdulazeez, Diyar Qader Zeebaree, Dathar A. Hasan.

9. T.R.Lekhaa, "Efficient Crop Yield and Pesticide Prediction for Improving Agricultural Economy using Data Mining Techniques" International Journal of Modern Trends in Engineering and Science, 2016.

10. RETRACTED: Obesity and mortality of COVID-19. Meta-analysis (2020) Abdulzahra Hussaina, Kamal Mahawar, Zefeng Xiac, WahYang, ShamsiEL-Hasanie
.

11. COVID-19: Transmission, prevention, and potential therapeutic opportunities (2020) Melika Lotfiab, Michael R.Hamblin, NimaRezaei.