A NOVEL APPROACH TO APPLY DIFFERENT ALGORITHMS TO PREDICT COVID-19 DISEASE

U. Mahesh (170031326), B. SYAM JASON (170030161), S. Nithya Tanvi Nishitha(170031524)

Name of the Supervisor: J. Surya Kiran, Asst. Professor,
Department of CSE

Introduction

Objective of your work

The Covid-19 outbreak occurred in Wuhan in December 2019 and spread everywhere in the world. The Covid-19 communicable disease have a vaccine and drug for its treatment. The foremost important factors in reducing the spread of the virus are washing hands, employing a mask, and reducing social distance. Today additionally to clinical studies, computer-aided studies also are widely administered for the Covid-19 outbreak. AI methods are successfully applied during this study. In this study, we used different algorithms for the prediction and analysis of Covid-19 daily cases. As a result of the study, the number of daily cases was successfully estimated with these different types of algorithms.

Origin of your proposal

Coronavirus disease (COVID-19) is an inflammation disease from the latest virus. This particular infection causes respiratory ailment (like influenza) with manifestations, for example, cold, cough, and fever, and in progressively serious cases, The problem in breathing. COVID-2019 has been perceived as a worldwide pandemic and a couple of examinations are being led utilizing different numerical models to anticipate the likely advancement of this pestilence. These numerical models hooked into various factors and investigations are dependent upon potential inclination. Here, we presented a model that could be useful to predict the Covid-19 daily cases by using different algorithms.

Methods

Methods and Materials

Here there are five scenarios/Algorithms

1. ***Decision Tree***
2. ***K Nearest Neighbors***
3. ***SVM***
4. ***Naïve Bayes***
5. ***Random Forest***

I. Decision Tree:

Decision Trees (DTs) are probably one of the most useful supervised learning algorithms out there. As opposed to unsupervised learning (where there is no output variable to guide the learning process and data is explored by algorithms to find patterns), in supervised learning your existing data is already labelled, and you know which behaviour you want to predict in the new data you obtain. This is the type of algorithms that autonomous cars use to recognize pedestrians and objects, or organizations exploit to estimate customers lifetime value and their churn rates.

II. K Nearest Neighbors:

K-Nearest Neighbors (KNN) is a standard machine-learning method that has been extended to large-scale data mining efforts. The idea is that one uses a large amount of training data, where each data point is characterized by a set of variables. Conceptually, each point is plotted in a high-dimensional space, where each axis in the space corresponds to an individual variable. When we have a new (test) data point, we want to find out the K nearest neighbors that are closest (i.e., most "similar" to it). The number K is typically chosen as the square root of N, the total number of points in the training data set.

III. SVM:

Support vector machines so called as SVM is a supervised learning algorithm which can be used for classification and regression problems as support vector classification (SVC) and support vector regression (SVR). It is used for smaller dataset as it takes too long to process. In this set, we will be focusing on SVC. SVM is based on the idea of finding a hyperplane that best separates the features into different domains. It uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.

IV. Naïve Bayes:

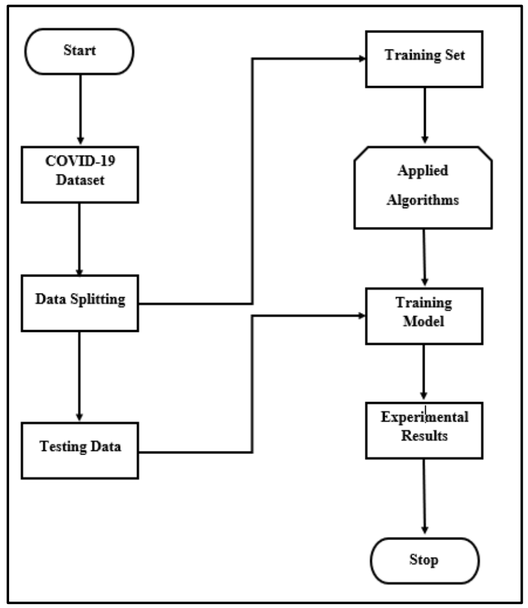
In statistics, naive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naïve) independence assumptions between the features. They are among the simplest Bayesian network models, but coupled with kernel density estimation, they can achieve higher accuracy levels. Naïve Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. Maximum-likelihood training can be done by evaluating a closed-form expression, which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers.

V. Random Forest:

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean/average prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set. Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees. However, data characteristics can affect their performance.

Block Diagram, Flowchart, Models, Results

Block Diagram



Step 1: Start
Step 2: Required dataset collection using Kaggle and GitHub repositories.
Step 3: Pre-processing the dataset to remove the noisy data.
Step 4: Splitting the data into training and testing data.
Step 5: Creation and training the models using the training data.
Step 6: Calculating accuracy of the models using the test data.
Step 7: Comparison of experimentation results.
Step 8: Stop.

Results

Accuracy and Error Rate Comparisons:

ALGORITHMS	ACCURACY	ERROR RATE
Decision Tree	96%	7.46%
K Nearest Neighbors	94%	8.01%
SVM	93%	7.57%
Naïve Bayes	92%	7.45%
Random Forest	92%	8.51%

The above table illustrates the accuracy and error rate of different algorithms used for the Covid-19 prediction during experimentation.

Conclusion

Discussion/Conclusion

In this paper, we predicted the COVID-19 cases by taking the data records of 3617 and by using the Decision Tree, SVM, Naïve Bayes, Random Forest, and KNN algorithms out of all the algorithms Decision Tree algorithm got the highest accuracy rate when compared with other algorithms.

Limitations

For Huge data sets it takes a lot of time for processing, it requires HDFS to implement it in faster way. Usage of Hadoop is not shown in this.

Future Direction

Currently, we predicted whether the person is cured or not cured due to COVID in future we can create models in such a way to predict the possible affected regions Based on the speed of increase of cases and the places that the affected people have visited and their geolocation we can be able to track what may be the affected areas in the future. And we can also use Hadoop and the concept of HDFs in such a way that we can be able to process huge datasets.

References and Affiliations

References

[1] Fuzzy rule-based system for predicting daily case in COVID-19 outbreak. (2020). 2020 4th international symposium on multidisciplinary studies and innovative technologies (ISMSIT).

[2] Nath, M. K., Kanhe, A., & Mishra, M. (2020). A novel deep learning approach for classification of COVID-19 images. 2020 IEEE 5th international conference on computing communication and automation (ICCCA).

[3] Gambhir, E., Jain, R., Gupta, A., & Tomer, U. (2020). Regression analysis of COVID-19 using machine learning algorithms. 2020 international conference on smart electronics and communication (ICOSEC).

[4] Kumar, N., & Susan, S. (2020). COVID-19 pandemic prediction using time series forecasting models. 2020 11th international conference on computing, communication and networking technologies (ICCCNT).

Acknowledgements

We would like to express our indebtedness and deep sense of gratitude to our supervisor "Mr. J. Surya Kiran" for his valuable guidance and support throughout our project.

We would also extend our thanks to the Head of the Department for Computer Science Engineering "Mr. V. Hari Kiran" for providing us with the faculty and support needed.