School of Computer Science and Engineering

VIT Chennai

Vandalur - Kelambakkam Road, Chennai - 600 127

Review Report

Programme: SCOPE

Course: CSE4020 Machine Learning

Slot: B2

Faculty: Dr Premaltha

Component: J Component (Project)

Title: Predicting country of origin for given data, and how many vaccinations are projected to be completed at a certain date

Team Members:

Advait Deochakke (20BCE1143)

Harish T R (20BCE1403)

Shah Siddh Tejaskumar (20BCE1937)

Predicting country of origin for given data, and how many vaccinations are projected to be completed at a certain date

*Note: Sub-titles are not captured in Xplore and should not be used

Advait Deochakke
SCOPE
VIT Chennai
20BCE1143
advait.deochaakke2020@vitstudent.ac.in

Shah Siddh Tejaskumar
SCOPE
VIT Chennai
Chennai, India
shahsiddh.tejaskumar2020@vitstudent.ac.in

Harish T R
SCOPE
VIT Chennai
Chennai, India
harish.tr2020@vitstudent.ac.in

***Abstract*—This project deals with utilizing a cleaned dataset about Covid-19, and using it to train Decision Tree classifiers and regressors, and Random Forest regressors to predict country of origin if data is missing, and predict the number of complete vaccinations at a given date. These models can help us in checking the validity of provided data, and can providing benchmarks for Vaccination rates, along with reliable predictions.**

Keywords— Decision Tree, Regression Tree, Accuracy, Precision, MAE, RMSE, Date

I. Introduction

In this project, we start with the raw Covid-19 Vaccinations dataset made available from OWID and Oxford, and clean it appropriately. We then divide it into three subsets, according to Country, Region, and Income. Then, we proceed to train a Machine Learning model with inputs cleaned for population differences, to predict our targets; Location of origin for data, and predicted vaccinations, both in a certain time duration classified by year and month.

As predicting the amount of vaccinations is a numerical result, we need to first scale our data so that evaluation metrics such as MAE, RMSE, can use utilized.

II. Implementation

A. Selecting a Dataset

First, we have to first evaluate our dataset. We see that it contains many columns, such as Location; Date; Cases, Deaths, Vaccinations both total and per million; and population. Where the dataset is marked by countries, we have further indicators such as HDI, GDPPP, etc.

B. Choosing Indicators

Then we decide what combination of indicators do we want our model to consider. As we are trying to classify which location the data belongs to, we ignore all metrics which give a very precise mark of the origin location, like population; raw number of cases, deaths, vaccinations; HDI; GDP, etc and only focus on normalized stats like per million cases, deaths, vaccinations and the month in which the data was taken.

Similarly, when vaccinations are to be predicted, we remove vaccinations from the list of indicators.

Since the date column in our dataset is organized as a string, we convert it to a recognizable datetime object, separating the year and month into new columns in the process.

C. Fitting the Data

Next, we split the full dataset into training and testing data, with a 60:40 split.

Then, we create a Decision Tree Classification model, and fit it to the training dataset. For vaccination prediction, we fit a Decision Tree Regressor and a Random Forest Regressor to a scaled version of the dataset.

D. Metrics

Finally, we apply the model to the testing data and evaluate using various appropriate metrics.

For Classification, we use the accuracy score, balanced accuracy score, precision score, and Jaccard score.

For Regression, we use mean squared error, mean absolute error, and root mean squared error.

III. Libraries and Methods Used

For the project, we use the extensive libraries available under the sklearn library, or scikit-learn. Furthermore, the dataset is stored in a pandas dataframe, The reason for that is use of datetime functions to separate year and month into integer columns is simple and efficient with the datetime library on a pandas dataframe.

A. *Packages imported*

- Overall packages →

Pandas, datetime, sqrt (math)

- Packages from fitting and modeling→

(all under sklearn package)

model_selection.train_test_split

preprocessing.StandardScaler

tree.DecisionTreeRegressor

tree.DecisionTreeClassifier

ensemble.RandomForestRegressor

- Packages for metrics (from metrics.xx) →

accuracy_score

balanced_accuracy_score

precision_score

jaccard_score

mean_squared_error

mean_absolute_error

B. *About Decision Tree (Classification)*

A supervised learning technique, it creates a tree with decision nodes and result leaves. It starts with a root node, and expands upon it with each classification.

A decision node asks effectively, a simple yes/no question, the answer to which decides which branch will be walked upon next, until a leaf node is hit. When we hit a leaf node, we stop and present the answer available on the leaf.

We can use two popular methods for attribute selection at a decision node, Information Gain and GINI Index.

Decision trees are easy to understand, as they operate similar to a human in exercising their ability.

A key problem in decision trees is overfitting, where the depth of the tree is increased such that it gives too accurate results on train data, but may not always have ideal accuracy in arbitrary testing data.

C. *About Decision Tree (Regression)*

We know decision trees predict a category, so if we think of continuos data in the terms of small sections of numbers, with the value of the category being the average of the section, we can utilize decision trees for regression.

The model sifts through the training data, making a section at with each major point of difference (decided by minimizing MSE). It takes the average value of all the points in the section as the value of the section.

A key point to note in this implementation is that overfitting can be a serious concern. If the points are spread out such that the model puts each point in a single section, it becomes terrible for testing data.

As such, we put limits on how many points are needed to make a section, and this prevents overfitting to a high degree.

D. *About Random Forest (Regression)*

A forest is nothing but a collection of similar, but different trees. That is exactly was random forest algorithm does, it makes multiple decision trees for the given data with many subsets, and they have low correlation.

When a testing data comes in, all the trees process it, and present their outcome. The outcome with the majority wins and is presented as the result.

E. *About Metrics*

- Accuracy Score

  The set of labels predicted for a sample must exactly match the corresponding set of labels in y_true.

  1 is best, 0 is worst.

- Balanced Accuracy Score

  Balanced accuracy is used in binary and multiclass classification problems to deal with imbalanced datasets. It is defined as the average of recall obtained on each class.

1 is best, 0 is worst.

- Precision Score

  The precision is the ratio tp / (tp + fp) where tp is the number of true positives and fp the number of false positives. The precision is intuitively the ability of the classifier not to label as positive a sample that is negative.

  1 is best, 0 is worst.

- Jaccard Score

  A similarity index, it compares two sets (y true and y predicted), and checks for similarities and dissimilarities. Very sensitive to small sample sizes, but we have large dataset.

  Due to it being intersection divided by union, it is lower than the other scores we use by design, highlighting the dissimilarity.

  1 is best, 0 is worst.

- MSE

  Mean of the Squares of the Errors. Take the estimated value, the real value; square of differences, then mean.

  Also has RMSE (root of the MSE)

- MAE

  Instead of squaring, we just use the absolute value.

IV. Results and Discussion

From the Country Dataset :

Predicting origin location

Jaccard ~ 95%

Rest 3 ~ 97.5%

Predicting Vaccinations

DT MSE ~ 3-6;

MAE ~ 0.3-0.5;

RMSE ~ 1.5-2.2

RF MSE ~ 2-4;

MAE ~ 0.2-0.4;

RMSE ~ 1.5-2

From the Region (Continent) Dataset :

Predicting origin location

Jaccard ~ 85%

Rest 3 ~ 92%

Predicting Vaccinations

DT MSE ~1-2;

MAE ~0.23-0.26;

RMSE ~1.2-1.4

RF MSE ~0.8-1;

MAE ~ 0.22-025.;

RMSE ~ 0.9-1

From the Income Dataset :

Predicting origin location

Jaccard ~ 97%

Rest 3 ~ 98%

Predicting Vaccinations

DT MSE ~1-3;

MAE ~0.2-0.5;

RMSE ~1-1.7

RF MSE ~0.7-1.5;

MAE ~ 0.07-0.1.;

RMSE ~ 0.6-1.3

As we can see, the origin location prediction gets extremely high scores when we are predicting the country, lower when we predict the continent, and highest when we run it on the income dataset.

This puts forth a suggestion that there is extreme similarity between how any group of countries based on economic factors (income) performs with relation to Covid-19, but when it comes to geographic grouping, we see less consistent results, but still at good correlation.

When predicting number of vaccinations, our output Y is in units of “Fully Vaccinated per Hundred”. So we can consider mean errors close to 1 as very accurate, and those nearer to 5 as having considerably more variance and hence less accurate.

The results suggest that there is once again great similarity between how vaccinations are proceeding in any group of countries when divided according to their wealth gaps. However, it produces very similar results this time for countries grouped by continent as well. However, we see a bit higher variance when countries are not grouped.

CONCLUSION

After evaluating our model, we come to the conclusion that countries with similar economic characteristics tend to have very similar infection and fatality profiles, but vaccination efficiency tends to be similar regardless of country.

Code Available at : https://github.com/Advait177013/CSE4020_ML_JComp

Dataset Available at : https://github.com/Advait177013/CSE4020_ML_JComp/tree/main/cleaned%20data

Related Work and references

1) Satu, M.S.; Howlader, K.C.; Mahmud, M.; Kaiser, M.S.; Shariful Islam, S.M.; Quinn, J.M.W.; Alyami, S.A.; Moni, M.A. Short-Term Prediction of COVID-19 Cases Using Machine Learning Models. Appl. Sci. 2021, 11, 4266. https://doi.org/ 10.3390/app11094266

Short-Term Prediction of COVID-19 Cases Using Machine Learning Models

The paper discusses the case in its infancy for Bangladesh, where limited awareness and various socio-economic factors, combined with very high population density can lead to explosive growth. Such growth is very different from global trends, and should be recognized and evaluated as such, for eg. By taking factors like population density, GDP, HDI, portion of population in urban areas, etc into account.

In the paper, we see models such as Naive-Bayes, Support Vector Machines, Linear Regression, Decision Trees, LSTM, and Random Forest applied to data from various countries. Many advanced models such as LASSO, RNN, GRU, and VAE also see use, to compare the effectiveness of the modelling

In the initial phase of the tuning, Linear Regression showed the best prediction versus the actual statistics recorded during the same period of 9 March 2020 through 9 April 2020. Due to the nature of Linear Regression, there were obvious unconformities owing to the straight line nature of the curve.

In the second round of testing data, SVR showed the best results, but the predictions continued to be slightly off from actual. In this case, they were on average higher than the actual rate owing to higher tuned growth rate.

The 3rd algorithm onwards, the prediction is almost spot on, with the PR algorithm and Poly-MLR giving the best results in round 3 and 4 of the iterations. Since the 5th iteration upto the 34th, Prophet algorithm gives the best results and confirms correctly to the real world data.

2) Usherwood, T., LaJoie, Z. & Srivastava, V. A model and predictions for COVID-19 considering population behavior and vaccination. Sci Rep 11, 12051 (2021). https://doi.org/10.1038/s41598-021-91514-7

A model and predictions for COVID-19 considering population behaviour and vaccination

Many Covid-related papers have been published in the past year or two, with a majority of them dealing with modeling the spread, including models involving segmentation and compartmentalizing of infected populace into sections of Susceptible, Infected, and Recovered (SIR Models).

Furthermore, models have also been made by taking into account SIR models across age ranges, to optimally distribute vaccinations to facilitate lowest infected and death counts, including vaccine effectiveness across age ranges and such .

Function :

β = β0 fI fV

fI = e^(−dI I)

fV = 1/fI + (1 − 1/fI)^(e−dV V)

The Level of Caution function is related to the various countermeasures and precautions taken by various people in the populace and how common they are, as a measure of modeling the degree of prevention. It includes things such as social distancing, personal protective hygiene, and government schemes (eg. Covid-0 policy in China).

The input factor depends on various states of the population which change throughout a period of time such as awareness, fatigue caused by the pandemic, changes in seasons which may affect daily routines, and changes in government policies such as mandatory lockdowns or not.

The level of caution is inversely related to e to the power of the above factor, so we can see that as the population is affected with more and more of the above changes in state (factor approaches 1), the level of caution correspondingly approaches 0, showing that there is closer to 0 chance of transmission between any two persons.

As the factor responsible for the Sense of Safety function approaches 0 (no drop in safety precautions despite high sense of safety; i.e,

increasing vaccinations), the Sense of Safety function approaches a value of 1, signalling that safety measures continue to be adopted appropriately.
If the populace feels that higher vaccination can allow for worse safety measures, then the Sense of Safety value neutralizes the Level of Caution, leading to no change in transmission rate from the base transmission rate.

Further classifying the population's response to the rollout of vaccinations, the paper classifies that the population shows no changes in its state as a result of the rollout of the vaccination, preserving the high alertness level, as opposed to having falling precautions and alertness with the rollout of the vaccines.

3) Muhammad, L.J., Algehyne, E.A., Usman, S.S. et al. Supervised Machine Learning Models for Prediction of COVID-19 Infection using Epidemiology Dataset. SN COMPUT. SCI. 2, 11 (2021). https://doi.org/10.1007/s42979-020-00394-7

Supervised Machine Learning Models for Prediction of COVID-19 Infection using Epidemiology Dataset

The current paper accounts for predicting whether an
individual is infected or not based on symptoms experienced and various different categorization techniques, whereas the previous papers focussed on predicting the overall trend across a subset of the populations with machine learning concepts.
The dataset which was being worked on was rehashed to only contain the Age and Sex, along with
eight other indicators, the likes of which included Diabetes, Pneumonia, Tobacco consumers, etc.
The authors continue on to apply specific supervised learning algorithms such as naive bayes,
Logistic Regression, and Decision Tree. The authors also use Artificial Neural Networks, and analyze the Coefficient of Correlation between the various independent variables (Age, Sex, Pneumonia, Diabetes, Tobacco consumption, etc) and the dependent variable (Positive RT-PCR test result for Covid-19 in the tested patients).
A brief look on the complex topic of ANNs is also given, touching on how bias and transfers between layers simulate the connection of neurons inside the brain. Lastly, the analysis of regression coefficient can show the degree of relatedness amongst the various independent decider variables and also their relation to the actual output variable, the dependent variable that is the frequency of occurrence of Covid-19 cases amongst the population. Post analysis, the paper shows that their results have indicated a weak positive correlation between any independent variable and a dependent variable. They further show the Specificity (True Negative Rate), Sensitivity (True Positive Rate), and the overall Accuracy of the various methods applied.

4) Malki, Zohair, et al. "The COVID-19 pandemic: prediction study based on machine learning models." Environmental science and pollution research 28.30 (2021): 40496-40506.

The COVID-19 pandemic: prediction study based on machine learning models

This paper used a machine learning approach to predict the spread of the virus in several selected countries. However, the nature of the virus is more or less the same everywhere, so the same approach can be applied to predict the spread of COVID-19 infections in other countries. A machine learning model is presented using statistical visualization graphs to better predict the spread of the COVID-19 pandemic.
They used the quality and density of the WHO data collected to determine the predictive value of the technique. Data used in this study were collected from official data repositories such as the official website of Johns Hopkins University, WHO and Worldometer. The data show the daily total number of confirmed COVID-19 positive cases, daily and total deaths, and total and daily recoveries.
This study is primarily based on a decision tree algorithm for global realtime data of COVID-19. The core idea is to use supervised machine learning algorithms for time series forecasting. The algorithms proposed for this work, namely decision tree algorithms and linear regression, are powerful models for predicting problems related to sequence and time series data. Using machine learning, they developed a predictive model using available data from COVID-19 found on a wellknown data storage website.

The proposed model was compared with various state-of-the-art models (random forest, ARIMA, deep learning) and the accuracy of the machine learning model on the training data set was measured by root mean square error (RMSE) and mean absolute error (MAE). . The machine learning models were developed to predict an estimate of the spread of his COVID-19 infection in many countries and how long the virus could be expected to be stopped thereafter. Their findings predicted a sharp decline in his COVID-19 infections worldwide in the first week of September 2021

5) Rahimi, Iman, Fang Chen, and Amir H. Gandomi. "A review on COVID-19 forecasting models." Neural Computing and Applications (2021): 1-11.

A review on COVID-19 forecasting models

This paper presents an overview and brief analysis of the leading machine learning predictive models for COVID-19. The work presented in this study consists of two parts. In the first section, a detailed Scientometric analysis presents an influential tool for literature analysis performed on his COVID-19 data from the Scopus and Web of Science databases. Keywords and topics are covered in the above analysis, and later in the work, classification of machine learning prediction models, criteria evaluation, and comparison of solution approaches are described.
Putra and Khozin Mu'tamar used a particle swarm optimization (PSO) algorithm to estimate the parameters of a susceptibility, infection and recovery (SIR) model. The results show that the proposed method is accurate and has sufficiently small error compared with other analytical methods. Mbuvha and Marwala [2] adjusted her SIR model for reported cases in South Africa after considering different reproduction number (R0) scenarios for reporting transmission and medical resource estimates. . Qi and Xiao suggest that both daily temperature and relative humidity can affect COVID-19 outbreaks in Hubei and other provinces. Salgotra and Gandomi developed his two his COVID-19 prediction models based on genetic programming and applied these models to India. The results of a study by show that the genetic evolutionary programming model of COVID-19 cases in India has proven to be highly reliable.
The paper presented a keyword-based network visualization, and the top keywords researchers focused on, including coronavirus, forecast, epidemic, humans, statistical analysis, quarantine, hospitalization, mortality, and weather.
A detailed analysis of the total number of works cited and the number of records vs. affiliation is also presented.
This paper aims to review the main predictive models for COVID-19 and provides a brief analysis
of the published literature. This article used keyword analysis to highlight the most important subject areas. In addition, several criteria were identified that will help researchers in their future work. The paper also acknowledges the most useful models researchers have used to predict this pandemic.

6) Xiang, Yue, et al. "COVID-19 epidemic prediction and the impact of public health interventions: A review of COVID-19 epidemic models." Infectious Disease Modelling 6 (2021): 324-342.

COVID-19 epidemic prediction and the impact of public health interventions: A review of COVID-19 epidemic models

In light of the COVID-19 global pandemic, governments around the world rely on mathematical predictions to support decision-making during the epidemic. Extrapolation of epidemic impacts from early stages of growth is affected by model, data, and behavioral uncertainties. Although the complexity and methodologies vary, simple models can still be used as a reference during the early growth stages of the epidemic and provide the basis for more complex models of infection to understand the course of disease development.
Various countries should take appropriate public health measures to control the progression of the epidemic. In particular, there is evidence that recessive infections are non- or poorly contagious, suggesting that existing COVID-19 epidemiological models may overestimate epidemic risk. The input epidemiological parameters of the prediction models show significant differences in predicting the severity of epidemic spread.
Therefore, preventive control organizations should be cautious when basing public health strategies

on the predictive outcomes of mathematical models.

7) Chu, Xu, et al. "Data cleaning: Overview and emerging challenges." Proceedings of the 2016 international conference on management of data. 2016.

Data Cleaning: Overview and Emerging Challenges

Industry and academics have recently shown a surge in interest in several facets of data cleansing, including innovative methods. These have included techniques such as crowdsourcing, abstractions, and new approaches towards scalability. Xu Chu et. al focus on not only looking into these innovative approaches but also give a breakdown into good practices, and different approaches.

The authors outline and evaluate two statistical approaches to cleaning up qualitative data, where methods either evaluate how cleaning will affect either upcoming numerical queries or employ machine learning techniques to increase accuracy, or efficiency. They place these methods within a single, comprehensive listing of data cleansing, enabling the demonstration of how numerous qualitative methods can lend themselves to such statistical analysis.

In the next section the authors introduce to the readers what data cleaning means from the perspective of statistics. Prominent problems like deduplication, repairing missing values, and presence of incorrect values is discussed. Furthermore, they also touch upon how to remove irrelevant and erroneous data.

The paper showcases its real value here as it provides a trove of detailed papers which check through and investigate each aspect of such errors in detail.

8) Villavicencio, C.N.; Macrohon, J.J.E.; Inbaraj, X.A.; Jeng, J.-H.; Hsieh, J.-G. COVID-19 Prediction Applying Supervised Machine Learning Algorithms with Comparative Analysis Using WEKA. Algorithms 2021, 14, 201. https://doi.org/10.3390/a14070201

COVID-19 Prediction Applying Supervised Machine Learning Algorithms with Comparative Analysis Using WEKA

Early diagnosis is essential to prevent the development of a disease that can cause danger on human lives. COVID-19, which is an infectious disease that has mutated into several variants, has become a global pandemic that requires it to be diagnosed as soon as possible. Using technology, the available information about COVID-19 is increasing every day useful information from massive data can be extracted using data mining. Many models have been proposed which can predict the presence of coronavirus in a person and one such published paper consists of some models predicted by some machine learning algorithms and comparing which one of those algorithms produces the best model, the one having less bias and less variance.

The paper first gives the statistics of covid- 19, the countries it has affected maximally, number of cases of recovery, number of cases, number of deaths, etc. Then it goes on to explain the symptoms of the disease, how the virus spreads, the parts of the body the disease affects, common prevention measures of covid- 19, how humanity is going through a failed attempt to control this disease as the number of cases everyday has been rising despite thorough human intervention through quarantine, vaccination and other measures. Then it says how prediction models help us in preventing the disease to a great extent.

The machine learning algorithms J48 Decision Tree, Random Forest, Support Vector Machine, KNearest Neighbours and Naïve Bayes algorithms were applied through WEKA machine learning software. Each model was evaluated using 10 fold cross validation and compared according to major accuracy measures, correctly or incorrectly classified instances, kappa, mean absolute error, and time taken to build the model. The results show that Support Vector Machine using Pearson VII universal kernel outweighs other algorithms by attaining 98.81% accuracy and a mean absolute error of 0.012.

There are other algorithms, eg deep learning algorithm called encoder which uses original patient features to get reconstructed patient features, which the algorithm itself generates and is different from the original patient features, which predict if a person having those particular features is covid19 susceptible or not. Although these provide accurate results, the accuracy of SVM which is 98.81% for this data which is more than enough, also that the cost of implementation of encoders are also very high, it is more plausible to with SVM's taking cost factors as well into consideration. In [4] a different dataset is used and it turns out that the naive- bayes classification turns out to be the best algorithm for the dataset, it actually depends which algorithm suits which dataset, one algorithm may produce the highest accuracy in one dataset while another may prove to be the best dataset for another algorithm.