Machine Learning Predictive Models Applied to COVID-19 Datasets to Estimate the Infection Risk in Specific Geographic Regions

PhD Student: Diogen Babuc, Computer Science Department, West University of Timișoara

Background

This study investigates statistical indicators and performance evaluation metrics for machine learning models applied to the COVID-19 datasets.

In the present work, the models used for the final projection of the results are Decision Tree (DT), Support Vector Machine (SVM), Extreme Gradient Boost (XGBoost), Classification and Regression Trees (CART), and K-Nearest Neighbors (KNN).

No improvement in results is observed when using dense neural networks on the dataset of Oceania countries in the present experiment because the dataset has few observations. The idea is to build a model that can classify well the unseen data.

In specialized articles [1,2,3,4,5] are used both machine learning models and neural networks.

Aim

The aim of this paper is to discover the best-performing models for two COVID-19 datasets (countries from Oceania; the last 100 countries on Worldometer) and to detect geographic areas with an above-average rate of infection, after a binary classification.

Materials and methods

The dataset that collected data on the Oceania countries had 110 observations, while the second dataset reached up to 1100 observations. Both datasets were taken from the Kaggle platform. The Oceania dataset was downloaded in March 2022, and the dataset of the last 100 countries from Worldometer (WM) by total infected number was downloaded in September 2020.

The statistical indicators that were selected for data processing are mean squared error, mean absolute error, coefficient of determination, and correlation coefficient. The values are calculated for 40 different cases of a variable parameter.

The predictive values taken from the confusion matrix are used to calculate the evaluation metrics: accuracy, sensitivity, specificity, precision, and F1-score. Classes and methods from the *Scikit-learn* library are used. Graphs and charts, obtained using the *Matplotlib* library, are uploaded to the *Streamlit Cloud* platform.

The classes from the binary classification are *A: areas at risk*, and *B: areas without above-average risk of infection* (Oceania: >10% of the population; Last 100 WM: >0.3% of the population).

Results and Discussions

Oceania dataset (DS1) – The best-performing models are CART (DT, criterion='gini') and Decision Tree (DT, criterion='entropy'), with all of the evaluation metrics (accuracy, sensitivity, specificity, precision, and F1-score) above 93.2%. The values for mean squared and absolute errors are less than 3.5 in most portions. The coefficients of determination and correlation are greater than 0.81 for most of the portions.

Dataset of the last 100 countries on Worldometer (DS2) (total cases until September 2020) – The best model-performing models are also DT and CART, both models having the accuracy and specificity bigger than 88%, precision and F1-score bigger than 78%, and the sensitivity of DT is circa 84%. The constructed graphs and charts show that DT have the mean absolute and squared errors less than 2, and the correlation and determination coefficient bigger than 0.85, on average.

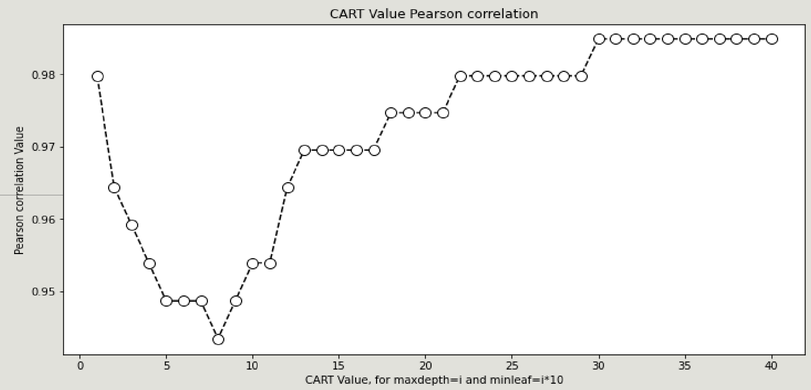

Figure 1. Correlation coefficient for the CART model, for the DS2 dataset

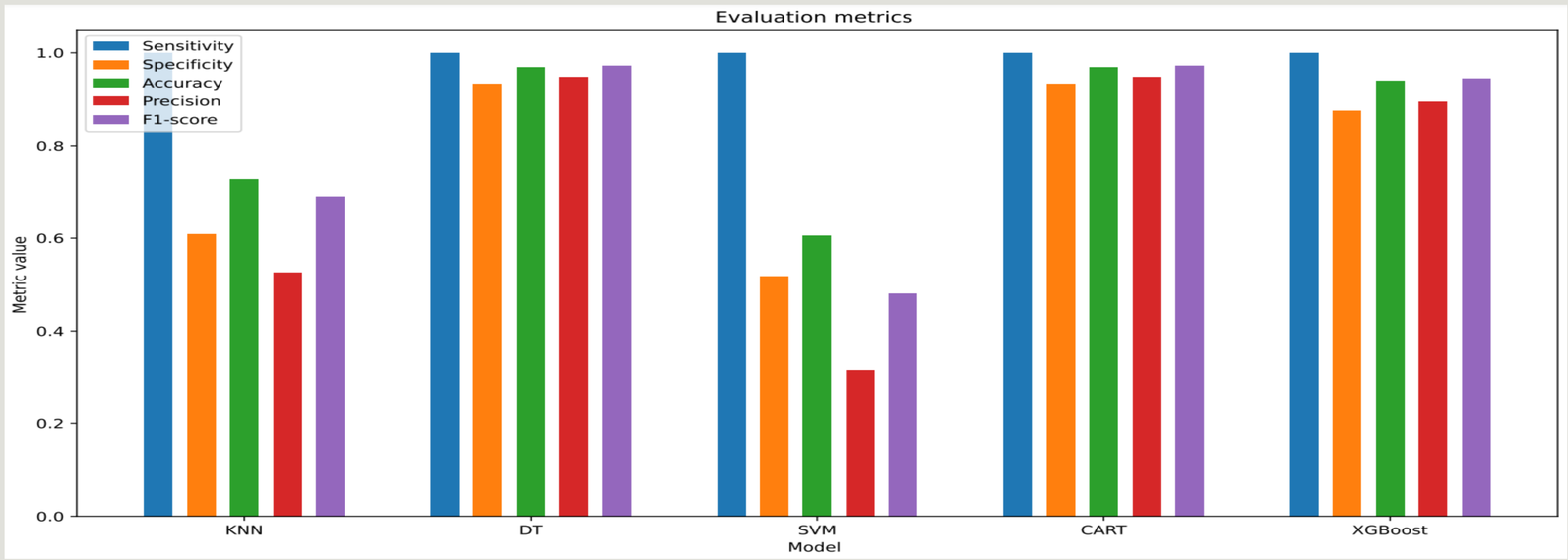

Chart 1. Evaluation metrics values for the DS1 dataset

Conclusions

Some models help discover geographic regions at risk of infection. The best-performing models for both datasets are CART and Decision Tree, while XGBoost is in the third position, considering the statistical indicators and performance evaluation metrics.

A future direction is to augment the datasets to reach a volume that is useful for neural networks as well, without reducing the performance of the other models.

Contact

Diogen Babuc
Computer Science Department, West University of Timișoara
Email: diogen.babuc@e-uvt.ro

References

1. Rustam F. et al. (2020). COVID-19 Future Forecasting Using Supervised Machine Learning Models. IEEE Access vol 8 pp 101489–101499.
2. Kurkina, E. S.; Koltsova, E. M. (2021) Mathematical modeling and forecasting of the spread of the COVID-19 coronavirus epidemic. Designing the future. In Proceedings of the Problems of Digital Reality: Proceedings of the 4th International Conference, Moscow, Russia, 4–5 February 2021.
3. Jiang X., Coffee M., Bari A., Wang J. (2020). Towards An Artificial Intelligence Framework for Data-Driven Prediction of Coronavirus Clinical Severity. Compu Mater Continua. vol 63 no 1 pp 537–551.
4. Caballé, N. C.; Castillo-Sequera, J. L.; Gómez-Pulido, J. A.; Polo-Luque, M. (2020). Machine Learning Applied to Diagnosis of Human Diseases: A Systematic Review. Appl. Sci. 10.
5. Saleem, F.; AL-Ghamdi, A. S. A.-M.; Alassafi, M. O.; AlGhamdi, S. A. (2022). Machine Learning, Deep Learning, and Mathematical Models to Analyze Forecasting and Epidemiology of COVID-19: A Systematic Literature Review. Int. J. Environ. Res. Public Health, 19, 5099.