Early Diagnosis of COVID-19 From Routine Lab Tests: A Machine Learning Model

Youssef Shanan (*Faculty of Business Informatics*) / **Supervisors: Dr. Ayman Al Serafi** (Business Informatics & Operations Management Department) **& Dr. Sally Ibrahim** (Biochemistry Department) **/ The German University in Cairo**

youssef.shanan@student.guc.edu.eg / ayman.alserafi@guc.edu.eg / sally.ibrahim@guc.edu.eg

Code: BINF03

Introduction

The main objective of this thesis is to investigate the effect of building machine learning models on the early diagnosis of COVID-19. The reason why a new approach was sought is that the reverse transcription polymerase chain reaction (RT-PCR) has several acknowledged drawbacks including 3-4 hours turnaround time, the necessity of accredited laboratories, costly equipment, and skilled staff (Brinati et al., 2020: 1). This thesis fulfills one of the sustainable development goals which is health and well being as it aims to early detect COVID-19 and reduce the spread of the virus.

Application: Proposed Approach

As shown in figure 1, we started with searching for an appropriate data set that contains routine lab tests. Once the lab tests were found, we started with preprocessing the dataset to change it from a raw form to a preprocessed dataset. Data preprocessing included handling missing values using the kNN imputer. After preprocessing the dataset, we split it into training and testing data. Furthermore, we constructed and trained four models: decision trees, random forests, logistic regression, and k-nearest neighbors. The next step was to cross validate on the training set to test the model's ability of classifying unseen data. After that, we evaluated the models based on testing data and selected the best performing model. The random forest resulted in the best performance. Future work will include deploying the model and constructing a user interface. This interface will inquire routine lab tests as inputs to predict if a patient is positive or negative on COVID-19.

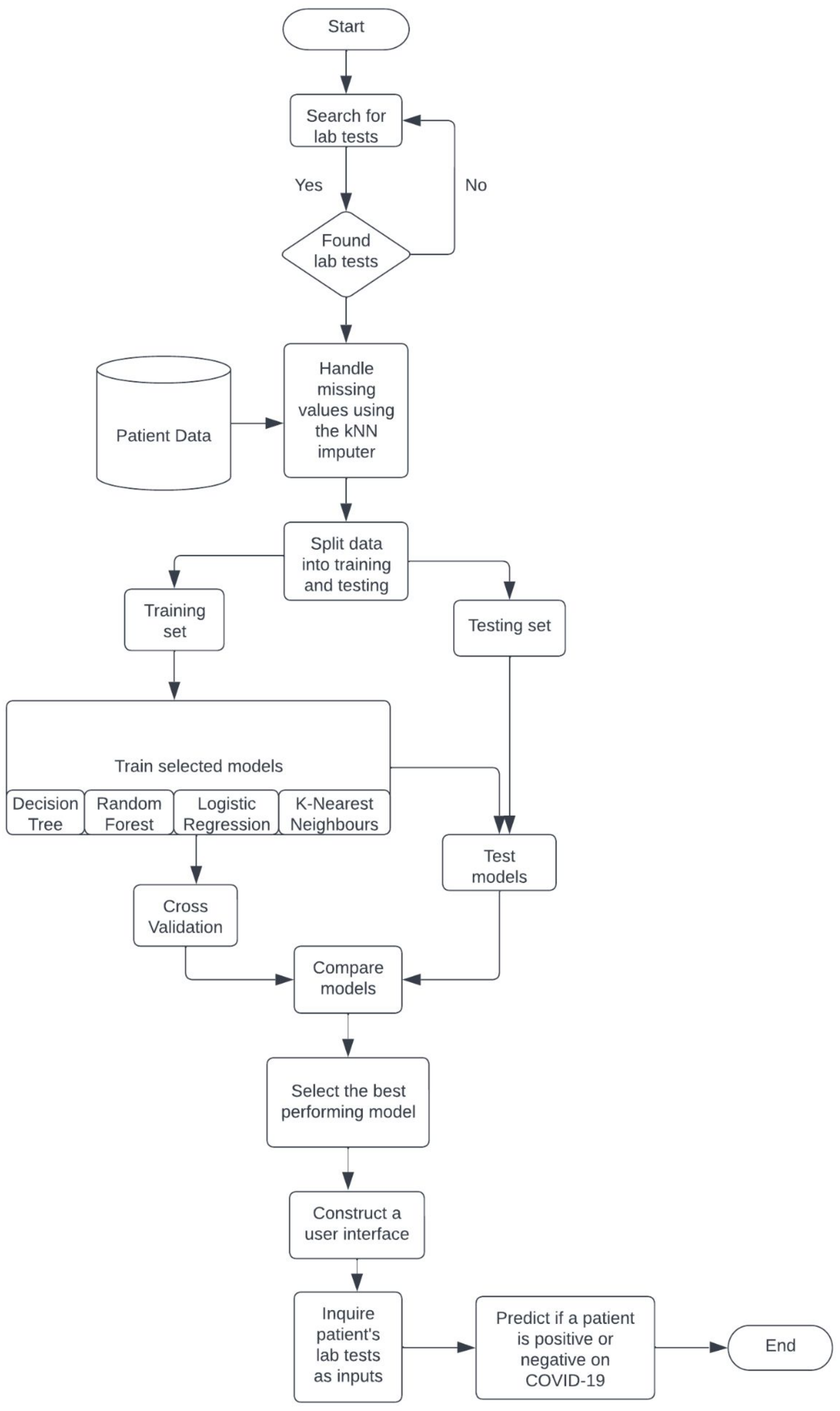
Start
Search for lab tests
Yes
No
Found lab tests
Patient Data
Handle missing values using the kNN imputer
Split data into training and testing
Training set
Testing set
Train selected models
Decision Tree
Random Forest
Logistic Regression
K-Nearest Neighbours
Test models
Cross Validation
Compare models
Select the best performing model
Construct a user interface
Inquire patient's lab tests as inputs
Predict if a patient is positive or negative on COVID-19
End

Figure 1. Proposed approach for the early detection of COVID-19

Methodology

The tool used in this thesis is the cross industry standard process for data mining which is known as the CRISP-DM. As demonstrated in figure 2, CRISP-DM includes six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment (Wirth & Hipp, 2000: 33).

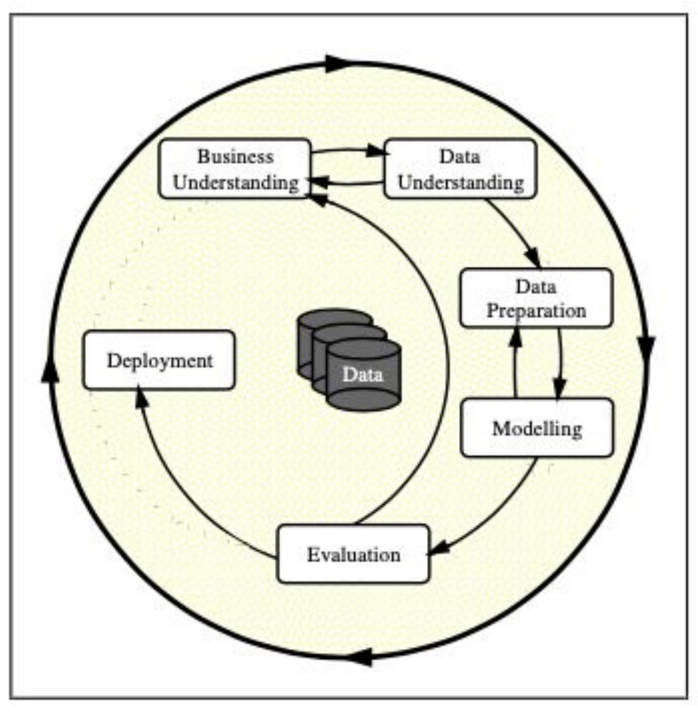

Figure 2. CRISP-DM phases [Wirth & Hipp, 2000: 33]

Dataset Description

Attribute
Gender
Age
White Blood Cells (WBC)
Platelets
Neutrophils
Lymphocytes
Monocytes
Basophils
C-Reactive Protein (CRP)
Aspartate Aminotransferase (AST)
Alanine Aminotransferase (ALT)
Alkaline Phosphatase (ALP)
Gamma-Glutamyl Transferase (GGT)
Lactate Dehydrogenase (LDH)
Eosinophils
PCR Result

Table 1. Dataset features

As shown in table 1, the dataset contains 16 features from which 13 are routine lab tests. Furthermore the other 3 are categorical features: age, gender, and PCR result which is the target variable.

Results

As demonstrated in table 2, four evaluation metrics were used to assess every model: accuracy, f1-score, precision, and recall. The random forest is the best performing model showing a higher accuracy, f1-score, precision, and recall compared to other models.

Model	Accuracy	F1-Score	Precision	Recall
Random Forest	0.798	0.862	0.828	0.898
Logistic Regression	0.750	0.824	0.817	0.895
K-Nearest Neighbor	0.702	0.786	0.793	0.780
Decision Tree	0.726	0.810	0.790	0.831

Table 2. Performance metrics

Conclusion

The study has showed how it is important in the current time to acquire machine learning in the healthcare sector. Furthermore, research gap has emerged by Brinati et al., (2020:10) proposing that further research is required to explore other approaches for detecting COVID-19. In this study, research gap was filled through proposing and constructing an alternative to the RT-PCR which is a machine learning model that is trained and tested with routine lab tests to be able to early detect COVID-19 in a cheaper and faster way. The limitation in this study is that the dataset found, had a low number of cases. On the other hand, larger datasets are needed to produce better results.

References

1. Ahamad, M. M., Aktar, S., Rashed-Al-Mahfuz, M., Uddin, S., Liò, P., Xu, H., ... & Moni, M. A. (2020). A machine learning model to identify early stage symptoms of SARS-Cov-2 infected patients. Expert systems with applications, 160, 113661.

2. Brinati, D., Campagner, A., Ferrari, D., Locatelli, M., Banfi, G., & Cabitza, F. (2020). Detection of COVID-19 infection from routine blood exams with machine learning: a feasibility study. Journal of medical systems, 44(8), 1-12.

3. Chen, M., & Decary, M. (2020, January). Artificial intelligence in healthcare: An essential guide for health leaders. In Healthcare management forum (Vol. 33, No. 1, pp. 10-18). Sage CA: Los Angeles, CA: SAGE Publications.

4. Davenport, T., & Kalakota, R. (2019). The potential for artificial intelligence in healthcare. Future healthcare journal, 6(2), 94.

5. Hashmi, H. A. S., & Asif, H. M. (2020). Early detection and assessment of covid-19. Frontiers in medicine, 7, 311.

6. Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. International journal of data mining & knowledge management process, 5(2), 1.

7. Toh, C., & Brody, J. P. (2021). Applications of Machine Learning in Healthcare. Smart Manufacturing: When Artificial Intelligence Meets the Internet of Things, 1-25.

8. Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. In Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining (Vol. 1, pp. 29-40).

Prepared for Thesis Poster Display Conference
Conference of the Parties (COP 27)

Faculty of Management Technology
THESIS Poster Display Conference 2022

11th -12th June 2022