УНИВЕРЗИТЕТ
ГОЦЕ ДЕЛЧЕВ

*	C	R	Y	P	T	O
C	Y	P	C	O	R	T
R	T	O	Y	R	P	C
Y	P	T	R	C	Y	O
P	O	C	P	Y	T	R
T	C	R	T	P	O	Y
O	R	Y	O	T	C	P

MACHINE LEARNING MODELS FOR PREDICTION OF COVID-19 INFECTION IN NORTH MACEDONIA

Maja Kukusheva Paneva, Natasha Stojkovikj, Cveta Martinovska Bande, Dushan Bikov

Introduction

- The latest pathogenic outbreak of novel Severe Acute Respiratory Syndrome-Coronavirus two (SARS-CoV-2) is responsible for the global pandemic 2019.
- Significant symptoms of COVID-19 include fever, cough and diarrhea.
- Machine Learning (ML) is one of the most advanced concepts of artificial intelligence (AI) and provides a strategic approach to developing automated, complex and objective algorithmic techniques for multimodal and dimensional biomedical or mathematical data analysis.
- ML has already shown potential for diagnosing, detecting, containing, and therapeutic motoring of many diseases.

Machine Learning (ML)

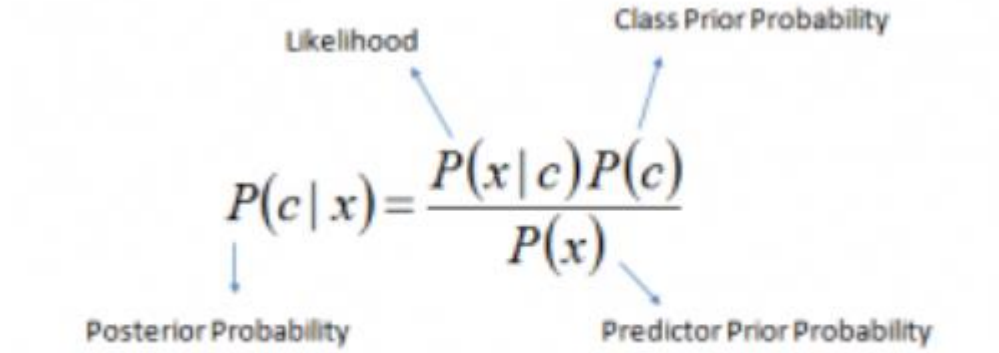
- ML techniques can be classified in four ways:
  - **Supervised learning techniques** are ML learning techniques or algorithms that bind previous and current dataset with the help of labeled data to predict future events
  - **Unsupervised learning** are ML techniques that are used when the training dataset is non-classified or nonlabelled.
  - **Semi-supervised learning techniques** are learning techniques that lie between supervised learning techniques and unsupervised learning techniques, where labeled and non-labeled datasets are used in the training process.
  - **Reinforcement learning techniques** interact with the learning environment by actions to locate errors

Supervised Learning Techniques

- Supervised learning techniques need humans to provide input and required output respectively, in addition to providing feedback about the accuracy of the prediction in the training process.
- In the development of the prediction model naive Bayes, logical regression and decision tree supervised learning algorithms were used.

Naïve Bayes Algorithm

- The ML Naïve Bayes algorithm is used for classification learning tasks in which instances of the dataset are discriminated based on the specified feature.
- The algorithm is probabilistic in nature and at the same time is based on Bayes Theorem



$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \cdots \times P(x_n|c) \times P(c)$$

Decision Tree Algorithm

- Decision tree ML algorithm is used to divide learning activities where the tree is constructed by dividing the dataset into smaller sets units each partition is clean and pure and the data classification depends on the type of the data.

- The decision tree algorithm has been used as one of the most effective learning algorithms due to its ability to handle all types of data, comprehension and simplicity.

Logistic Regression Algorithm

- Logistic Regression ML algorithm is used for classification learning task in which the association versus categorial dependent features against independent features are determined.
- The association between dependent features and independent features of the dataset is:

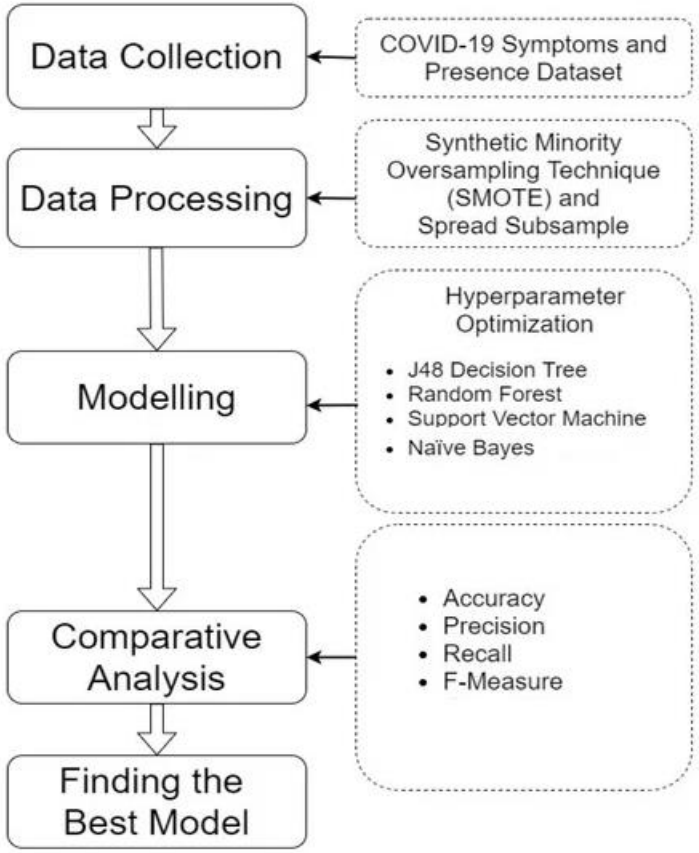
$$i = \text{Logistic regression}(p) = 1n\left(\frac{p}{1-p}\right).$$

Support Vector Machine

- Support vector machine (SVM) is a learning algorithm used for regression and classification learning tasks. The dataset features are represented in space in SVM and divided into points and groups with similar structures that fall into the same group.
- The data is considered p- dimensional for linear SVM that can be partitioned by p-1 planes known as hyper planes.
- The planes are divide the set of boundaries and data space among the data group for regression or classification learning task.

Materials and Model

- Methodology to build machine learning classification models for COVID-19

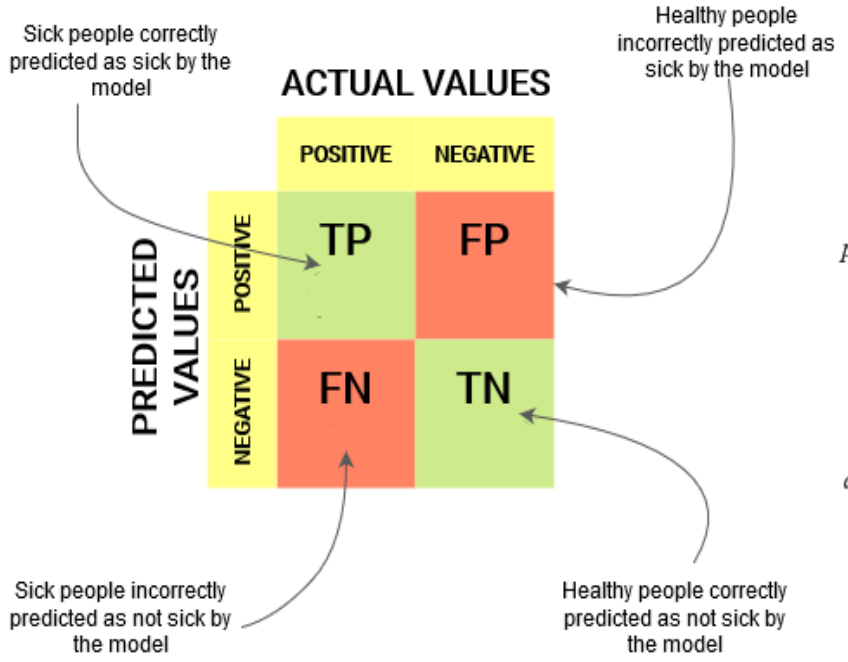


Dataset

- An epidemiology dataset of positive COVID-19 cases from North Macedonia obtained from Public Health Institute of North Macedonia
- The data set encompasses 14 features:
  - demographic variables: age and gender
  - 12 clinical indicators: pregnancy, pneumonia, cardiovascular diseases (CVDs), diabetes, pneumonia, hepatitis, neuromuscular, hypothyroid/ Hashimoto's, immunodeficiency/ HIV, cancer, chronic kidney disease (CKDs) and the outcome (deceased or recovered).

Results and discussion

- The performance of the classification models was evaluated with commonly used metrics precision, recall and F1 score.
- The F1 score is the harmonic mean of precision and recall. It combines both precision and recall into a single metric.



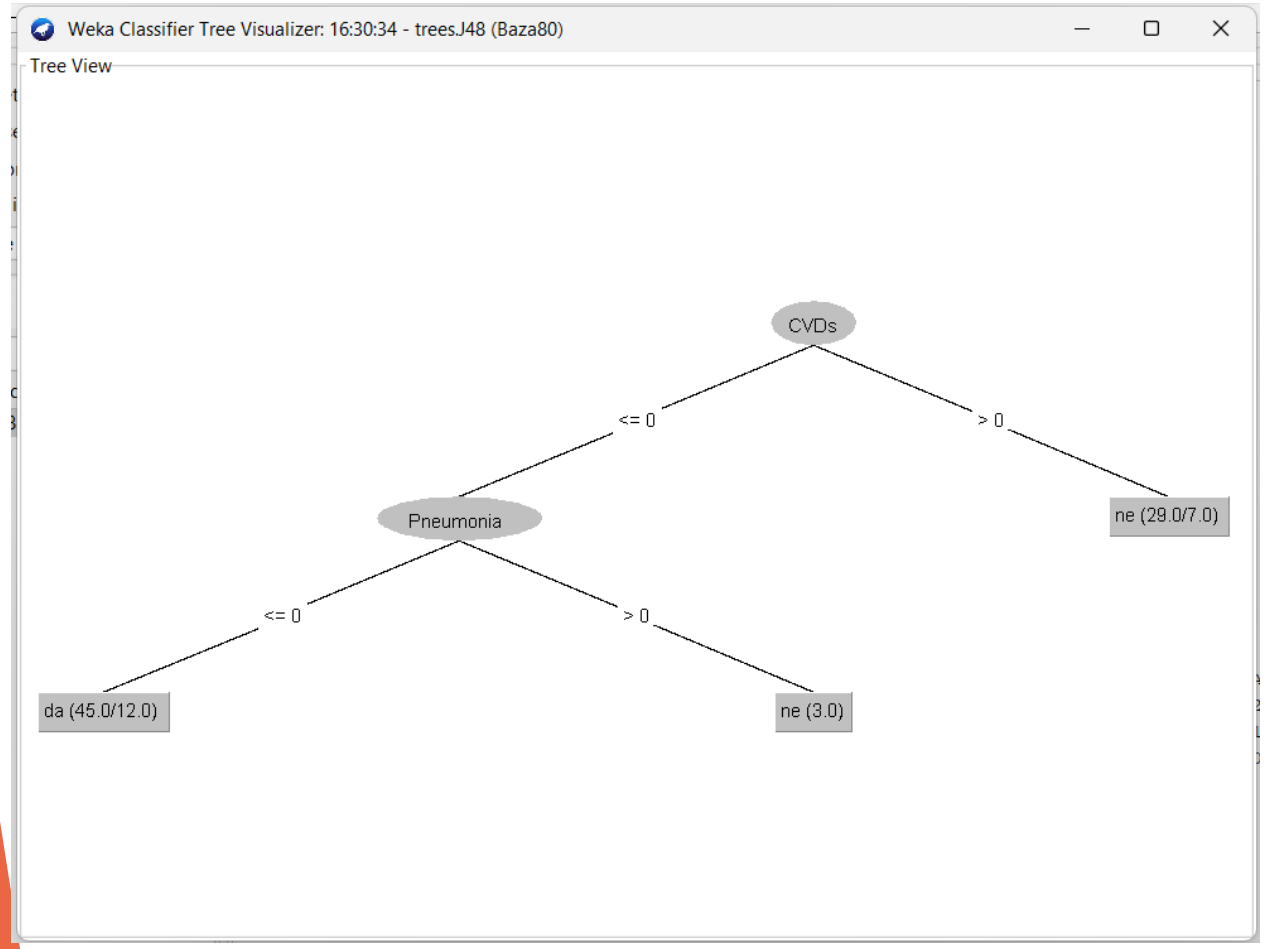
$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall}$$

$$accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

Decision Tree- J48



=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.568	0.175	0.750	0.568	0.646	0.408	0.666	0.672	ne
	0.825	0.432	0.673	0.825	0.742	0.408	0.666	0.611	da
Weighted Avg.	0.701	0.309	0.710	0.701	0.696	0.408	0.666	0.640	

Correctly Classified Instances	54	70.1299 %
Incorrectly Classified Instances	23	29.8701 %
Kappa statistic	0.3962	
Mean absolute error	0.3865	
Root mean squared error	0.4603	
Relative absolute error	77.3761 %	
Root relative squared error	92.0873 %	
Total Number of Instances	77	

Random Forest

Correctly Classified Instances	53	68.8312 %
Incorrectly Classified Instances	24	31.1688 %
Kappa statistic	0.3731	
Mean absolute error	0.3395	
Root mean squared error	0.4575	
Relative absolute error	67.9669 %	
Root relative squared error	91.5322 %	
Total Number of Instances	77	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.622	0.250	0.697	0.622	0.657	0.375	0.756	0.706	ne
	0.750	0.378	0.682	0.750	0.714	0.375	0.756	0.757	da
Weighted Avg.	0.688	0.317	0.689	0.688	0.687	0.375	0.756	0.733	

Naïve Bayes Algorithm

Correctly Classified Instances	57	74.026 %
Incorrectly Classified Instances	20	25.974 %
Kappa statistic	0.4776	
Mean absolute error	0.3208	
Root mean squared error	0.4416	
Relative absolute error	64.2309 %	
Root relative squared error	88.3574 %	
Total Number of Instances	77	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.676	0.200	0.758	0.676	0.714	0.480	0.781	0.740	ne
	0.800	0.324	0.727	0.800	0.762	0.480	0.781	0.766	da
Weighted Avg.	0.740	0.265	0.742	0.740	0.739	0.480	0.781	0.754	

Support Vector Machine with

Correctly Classified Instances	58	75.3247 %
Incorrectly Classified Instances	19	24.6753 %
Kappa statistic	0.5032	
Mean absolute error	0.2468	
Root mean squared error	0.4967	
Relative absolute error	49.4041 %	
Root relative squared error	99.3811 %	
Total Number of Instances	77	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.676	0.175	0.781	0.676	0.725	0.508	0.750	0.684	ne
	0.825	0.324	0.733	0.825	0.776	0.508	0.750	0.696	da
Weighted Avg.	0.753	0.253	0.756	0.753	0.752	0.508	0.750	0.690	

Decision

- The analyses show that Support Vector Machines and Multilayer Perceptron with precision 76%, recall 75% and F1 score 75% have better evaluation values than the other classifiers.
- Similar results are obtained with Naïve Bayes (precision 74%, recall 74% and F1 score 74%), Logistic Regression (precision 71%, recall 71% and F1 score 71%), and Decision Trees (precision 71%, recall 70% and F1 score 70%).
- The research demonstrated that Machine Learning (ML) can achieve a notable degree of accuracy in predicting COVID-19 outcome.

	F1 Measure	Precision	Recall
J48	0.696	0.71	0.70
Random Forest	0.687	0.71	0.70
Naïve Bayes	0.739	0.74	0.74
Support Machine Vector (SMV)	0.752	0.76	0.75
Logistic Regresion	0.710	0.710	0.70