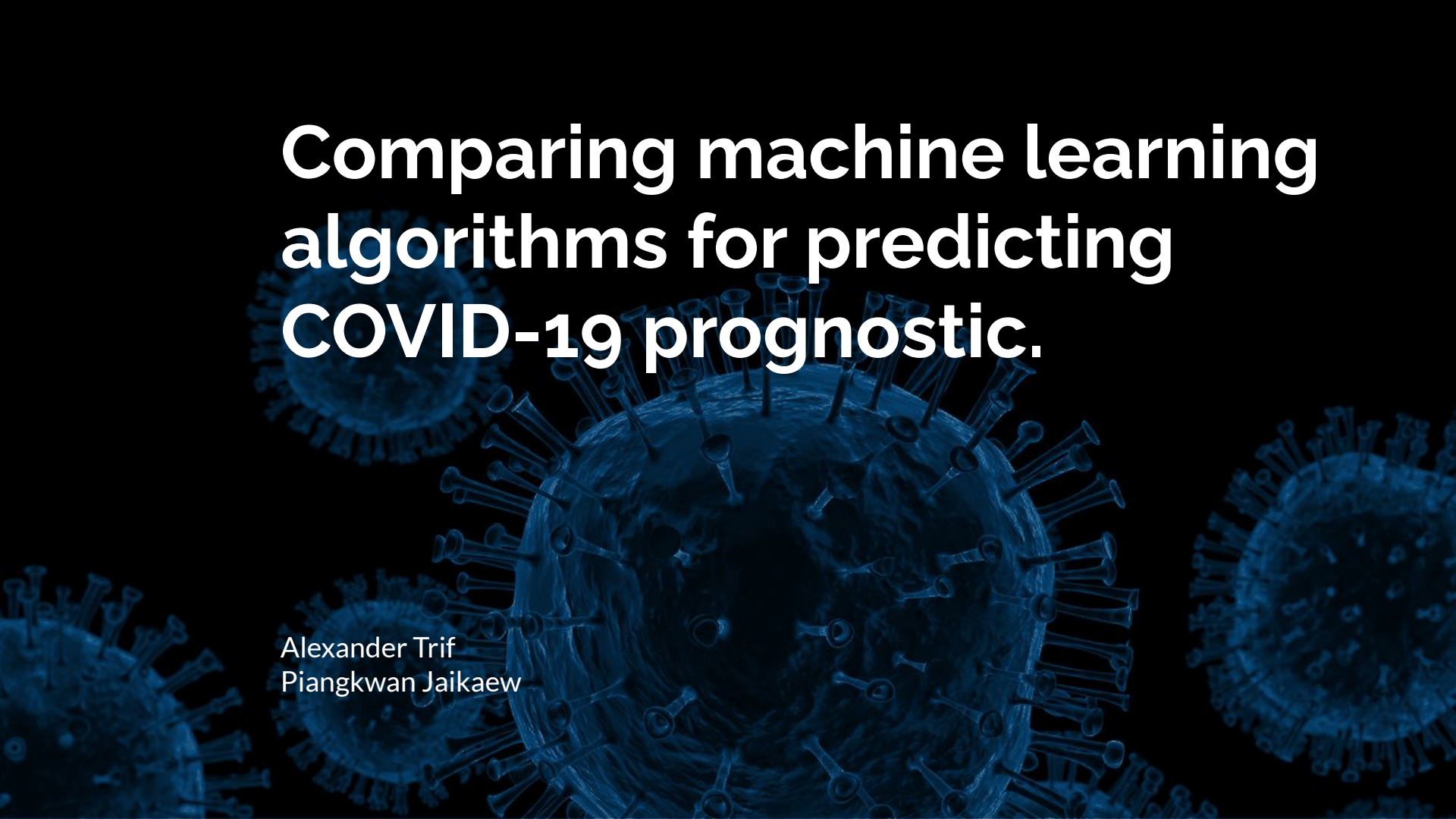
Comparing machine learning algorithms for predicting COVID-19 prognostic.



Alexander Trif
Piangkwan Jaikaew

Over $305 billion was allocated to health expenditure in 2020

THAT REPRESENT

13.8 % of CANADA'S GDP

Source: [Canadian Institute for Health Information]

Key Challenge

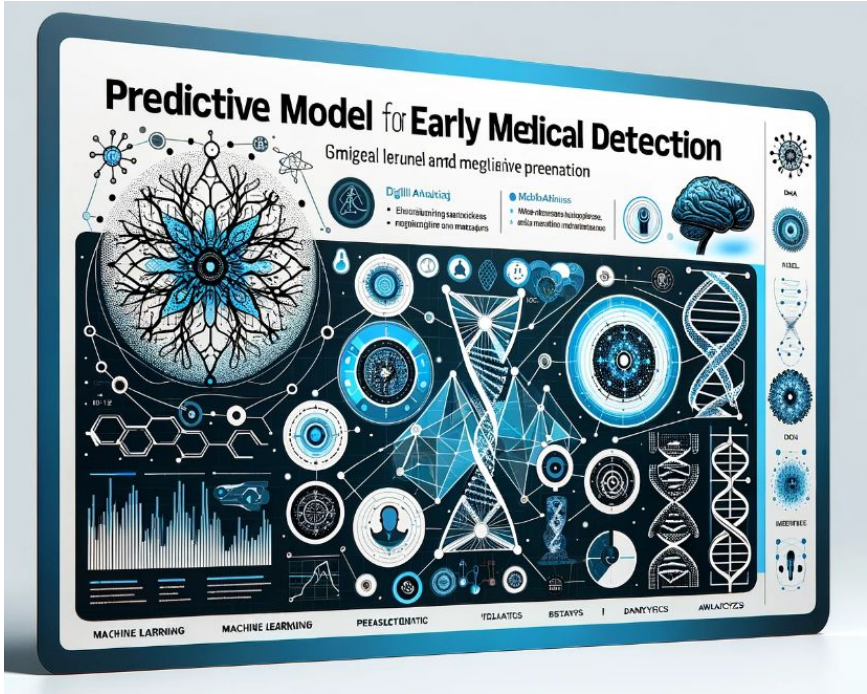
How Machine learning (ML), in augmenting healthcare responses to the pandemic for rapid and accurate COVID-19 detection.

Project Overview

Goal: Develop a predictive model for early COVID-19 detection.

Data: COVID-19 dataset (demographics, symptoms, medical history).

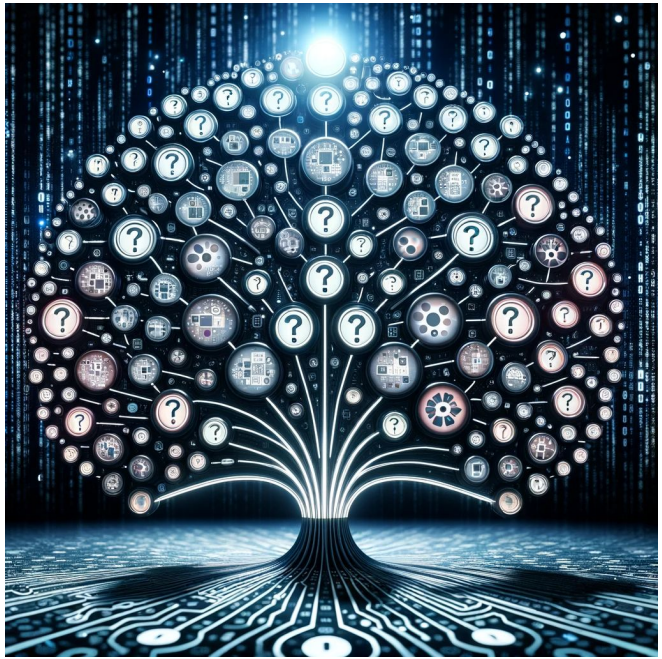
Methodology: Overview of exploratory data analysis and ML model development process.



ML Model in Consideration

Decision Trees:

A model that represents decisions and their possible consequences as a tree-like graph. It's like playing a strategic game of "twenty questions," where each question leads you closer to the answer.



ML Model in Consideration

Random Forests:

This ensemble learning method constructs numerous decision trees at training time and outputs the mode of the classes (classification) or mean prediction (regression) of the individual trees. It is akin to a council of experts where each provides input, and the final decision is a majority vote.

ML Model in Consideration

XGBoost (Extreme Gradient Boosting):

An efficient and scalable implementation of gradient boosting. It builds models sequentially, with each new model correcting the errors made by previous models. Imagine a team where each new member learns from the mistakes of those before them, leading to improved overall performance.

ML Model in Consideration

Logistic Regression:

A statistical model that estimates the probability of a binary outcome. It's similar to determining odds in a race, predicting the likelihood of a particular outcome.

ML Model in Consideration

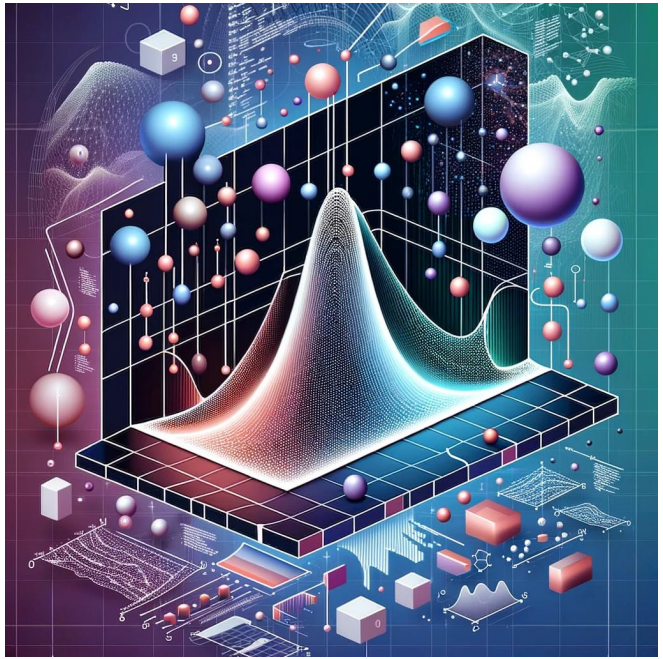
K-Nearest Neighbors (KNN):

This method classifies data based on the classification of its nearest neighbors. It's like asking a group of nearest neighbors for advice and following the most common suggestion.

ML Model in Consideration

Support Vector Machines (SVMs):

These models find an optimal boundary (hyperplane) that separates classes. Imagine drawing the straightest line or plane that divides groups of points.



ML Model in Consideration

SVMs Tuned with Weight Classes:

A variant of SVMs that assigns different weights to classes during training, making it particularly useful for imbalanced datasets. It's an adjusted approach to the standard SVM to address specific data biases.

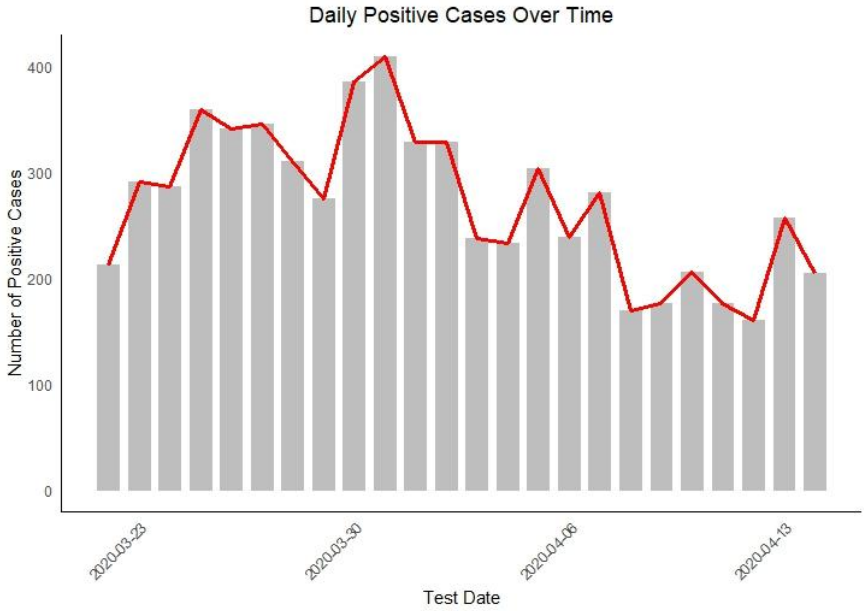
Data Insight

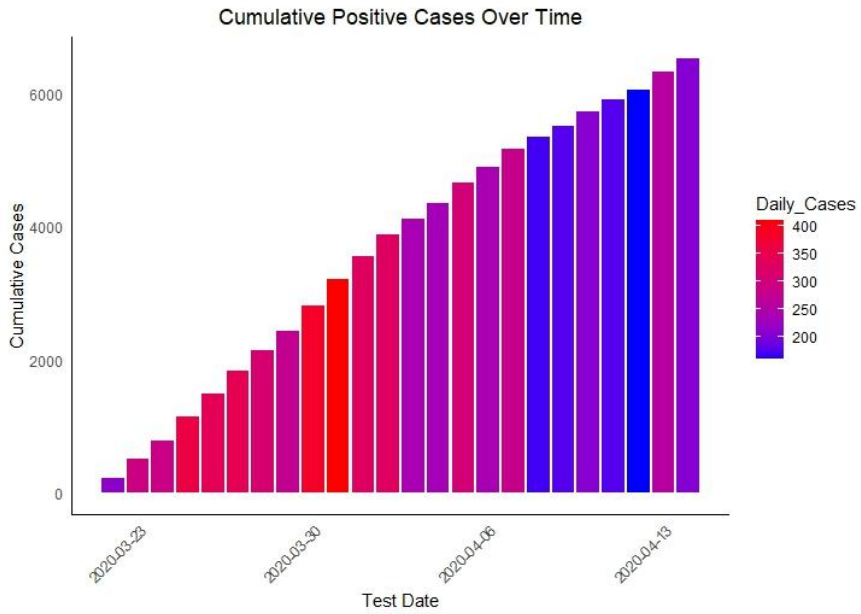
6,516
Confirmed Covid-19 Cases

21,936
No. of Observation

2,959 %
Percentage change
(Growth Rate)

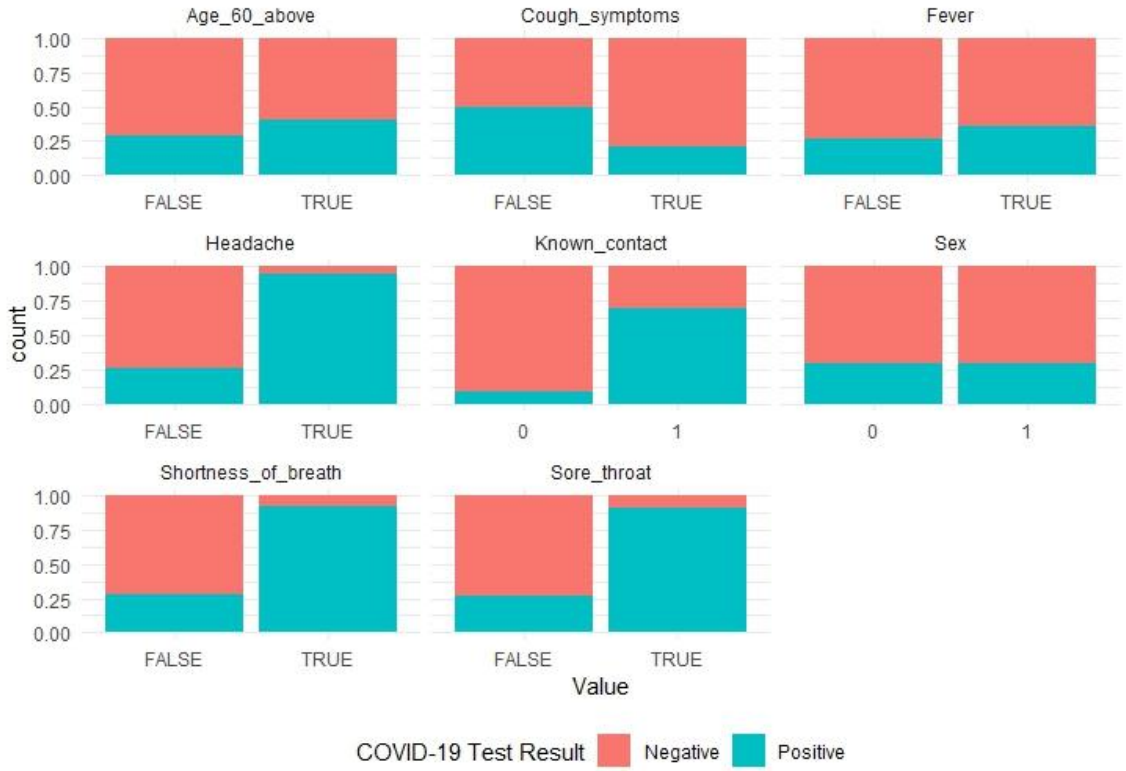
9
Features
(Demographic & Symptoms)





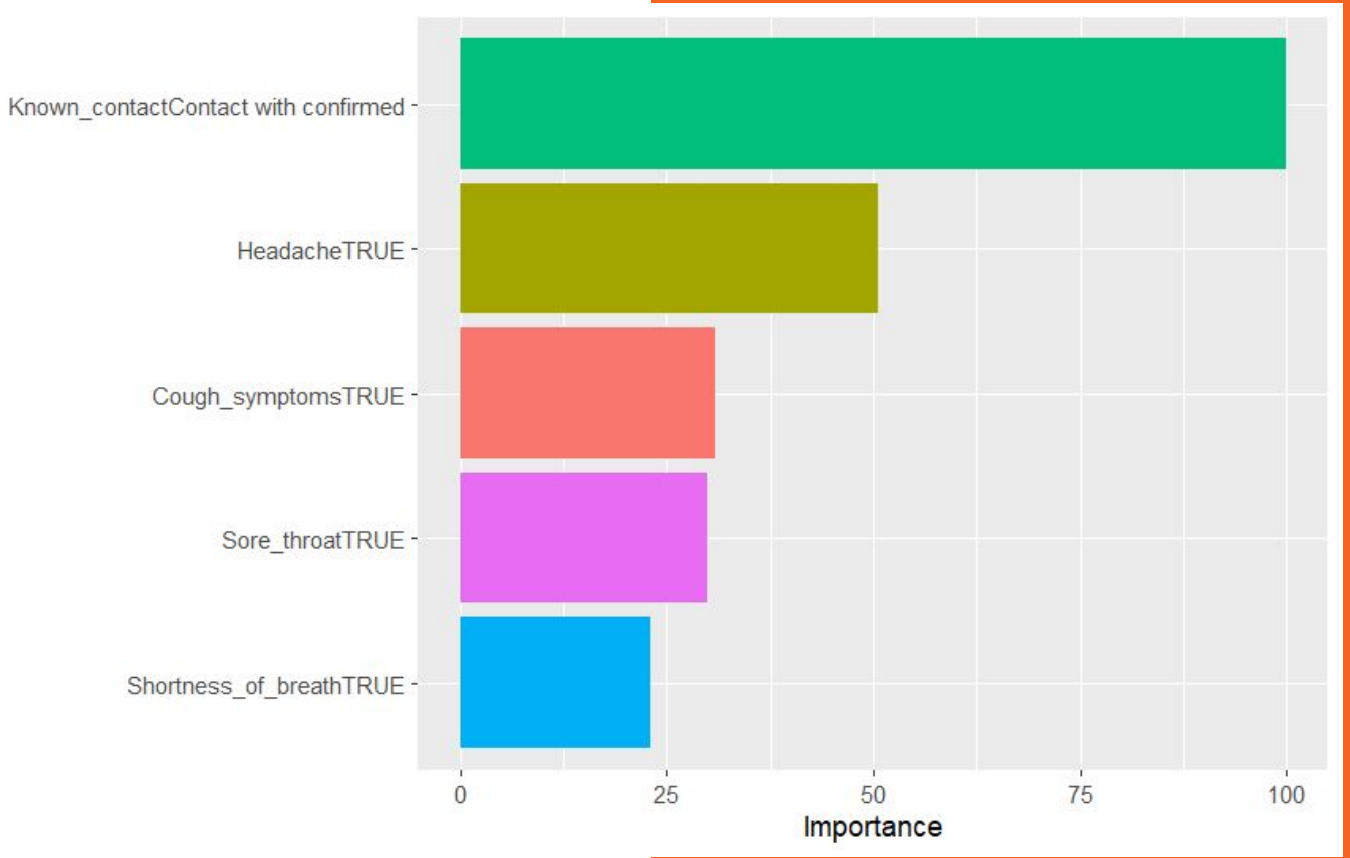
Data collected from March 22nd - April 14th, 2020.

Symptom Prevalence and COVID-19 Infection Outcomes by Demographics

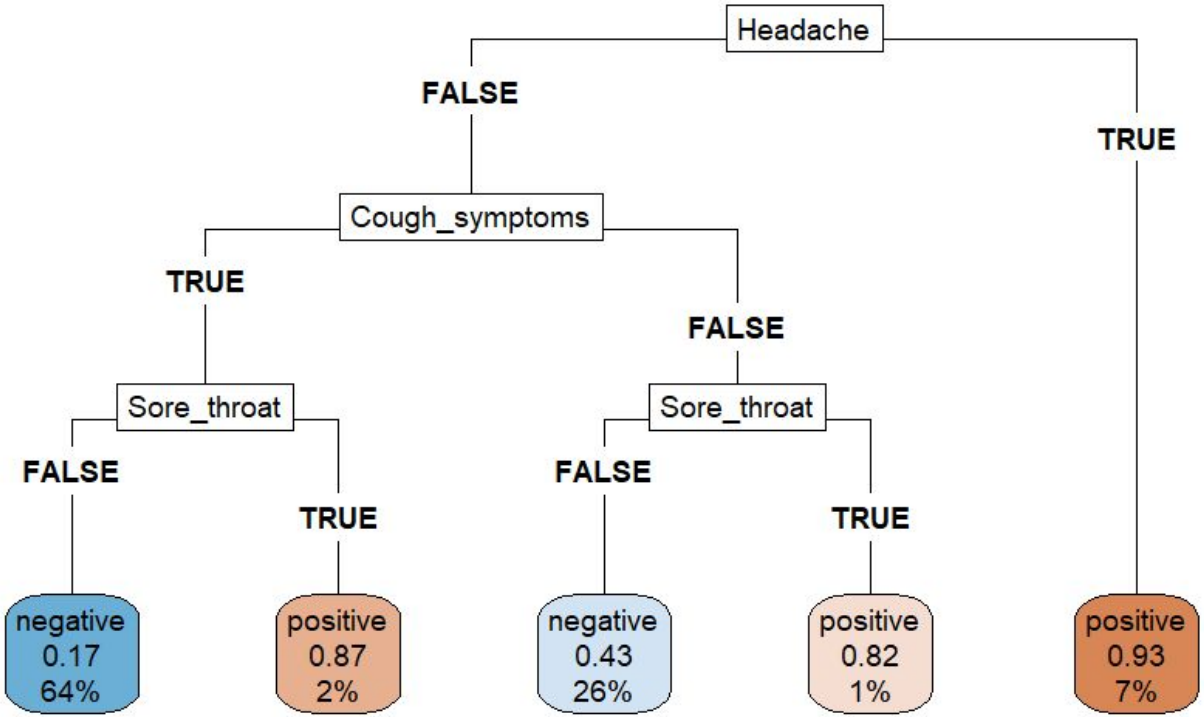


Feature Importance

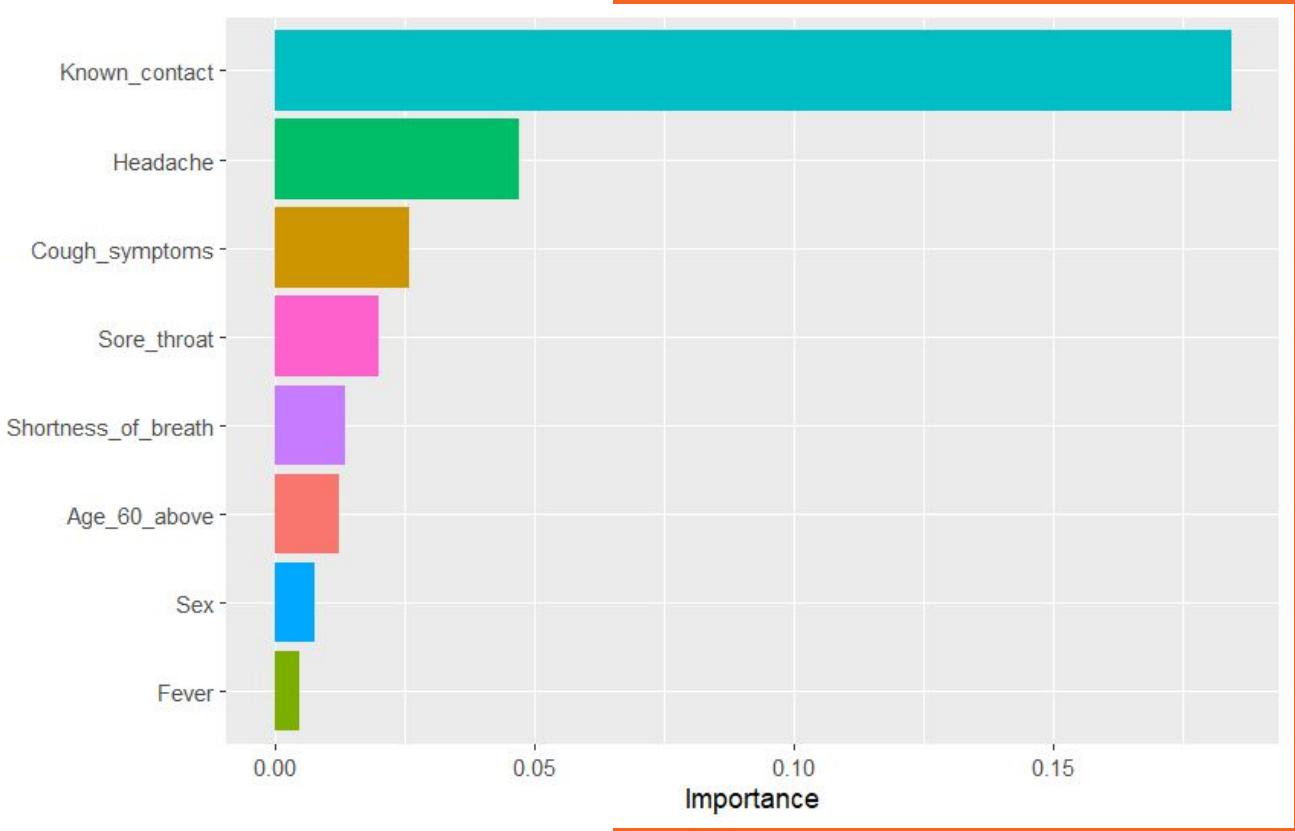
Feature Important using Decision Tree Classifier



Decision Tree

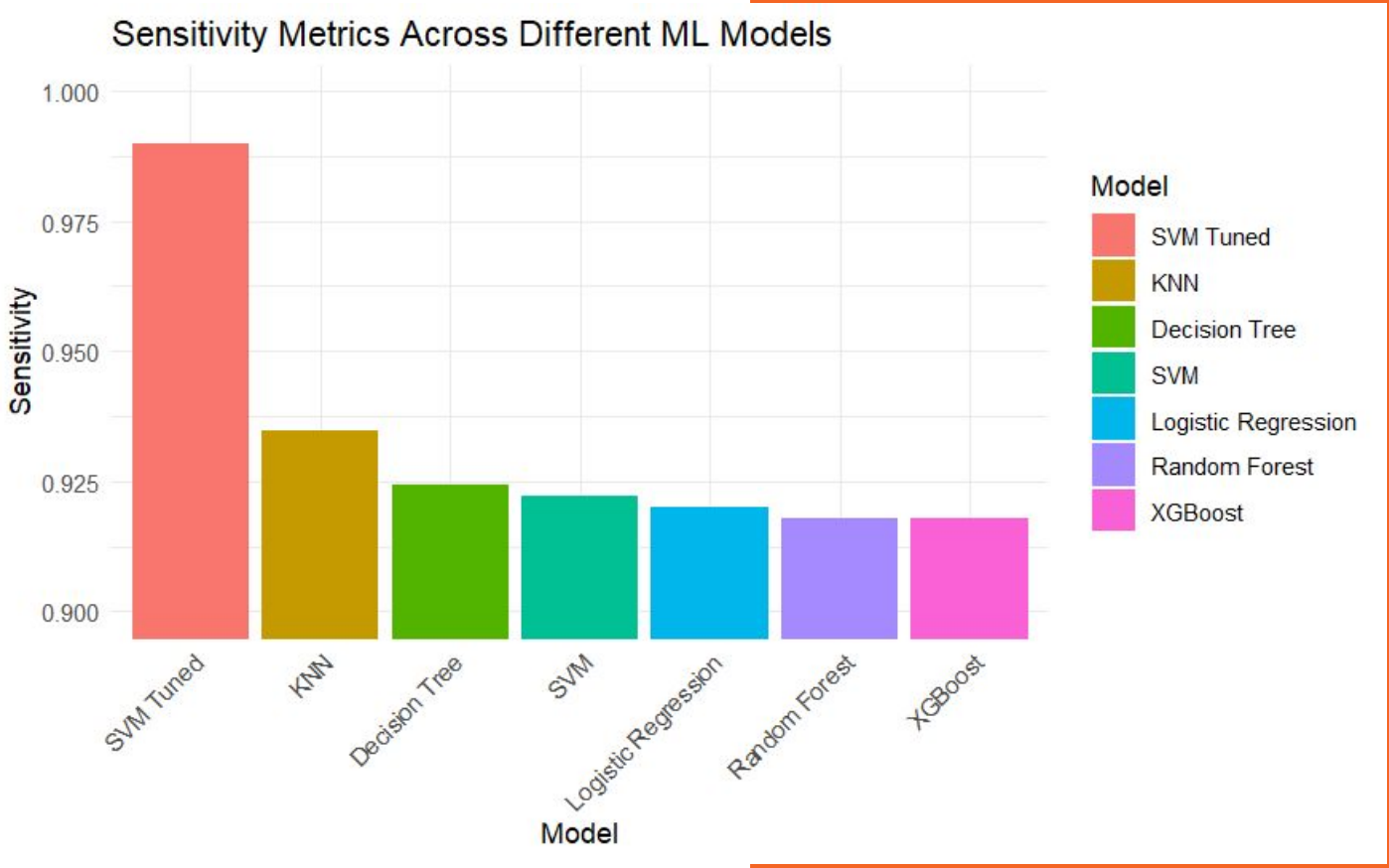


Feature Important using SVM Classifier

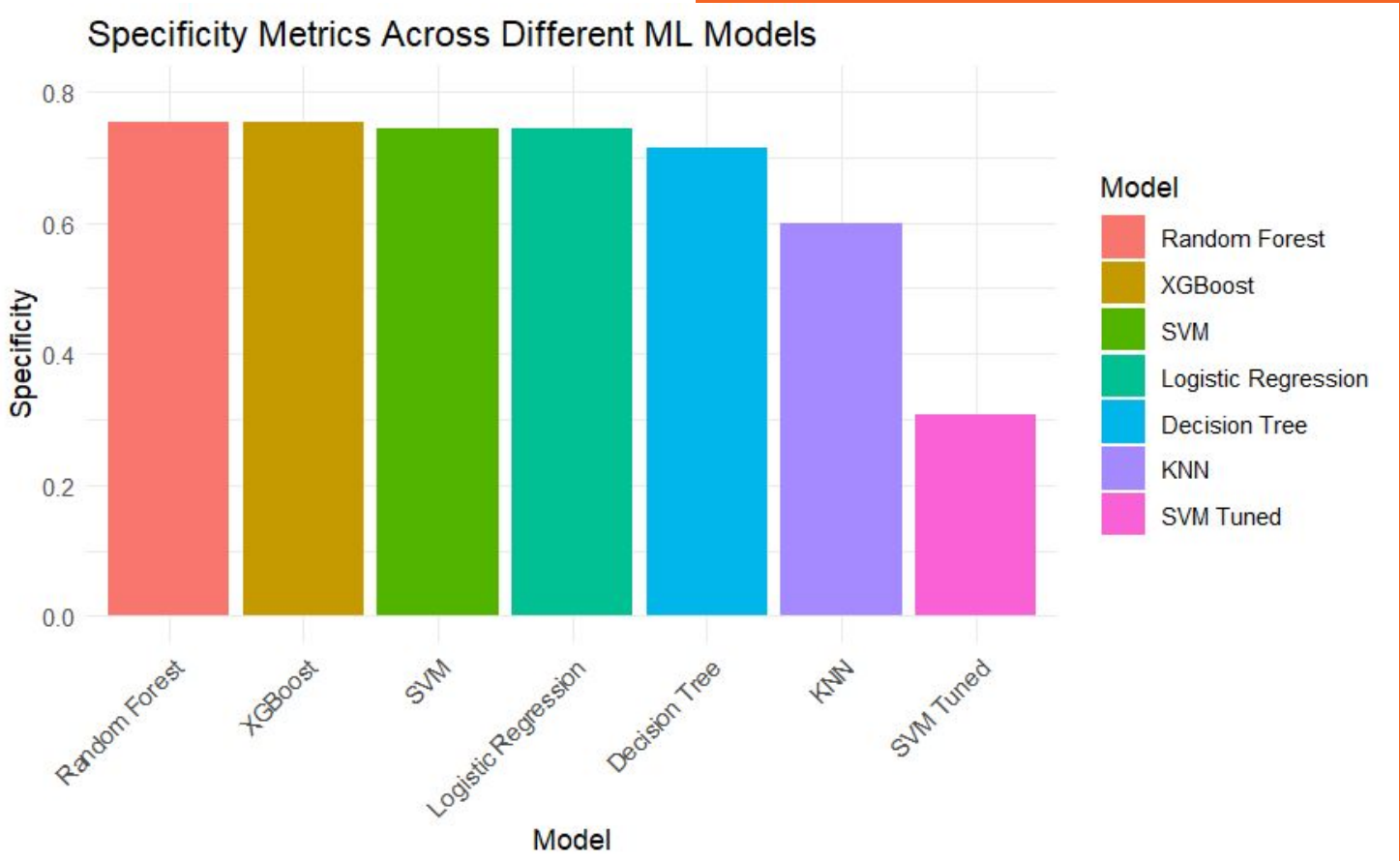


Model Performance Matrix

Sensitivity Metrics



Specificity Metrics



Sensitivity vs Specificity

What is more important?

Trade-Off in Medical Diagnostics: Ideal tools balance high sensitivity (detecting true cases) and specificity (avoiding false positives).

Priority in Infectious Diseases: For COVID-19, high sensitivity is essential to minimize false negatives, crucial for public health.

Balancing Act: The goal is to reduce both false negatives and positives for accurate diagnosis and effective disease management.

Model Efficacy: Tuned SVM model stands out for striking the best balance between sensitivity and specificity in COVID-19 diagnosis.

Public Health Implications: Accurate diagnosis is key for appropriate treatment and containment of infectious diseases like COVID-19.

What's Special about the Tuned SVM?

Cost Parameter allows us to control sensitivity/specificity

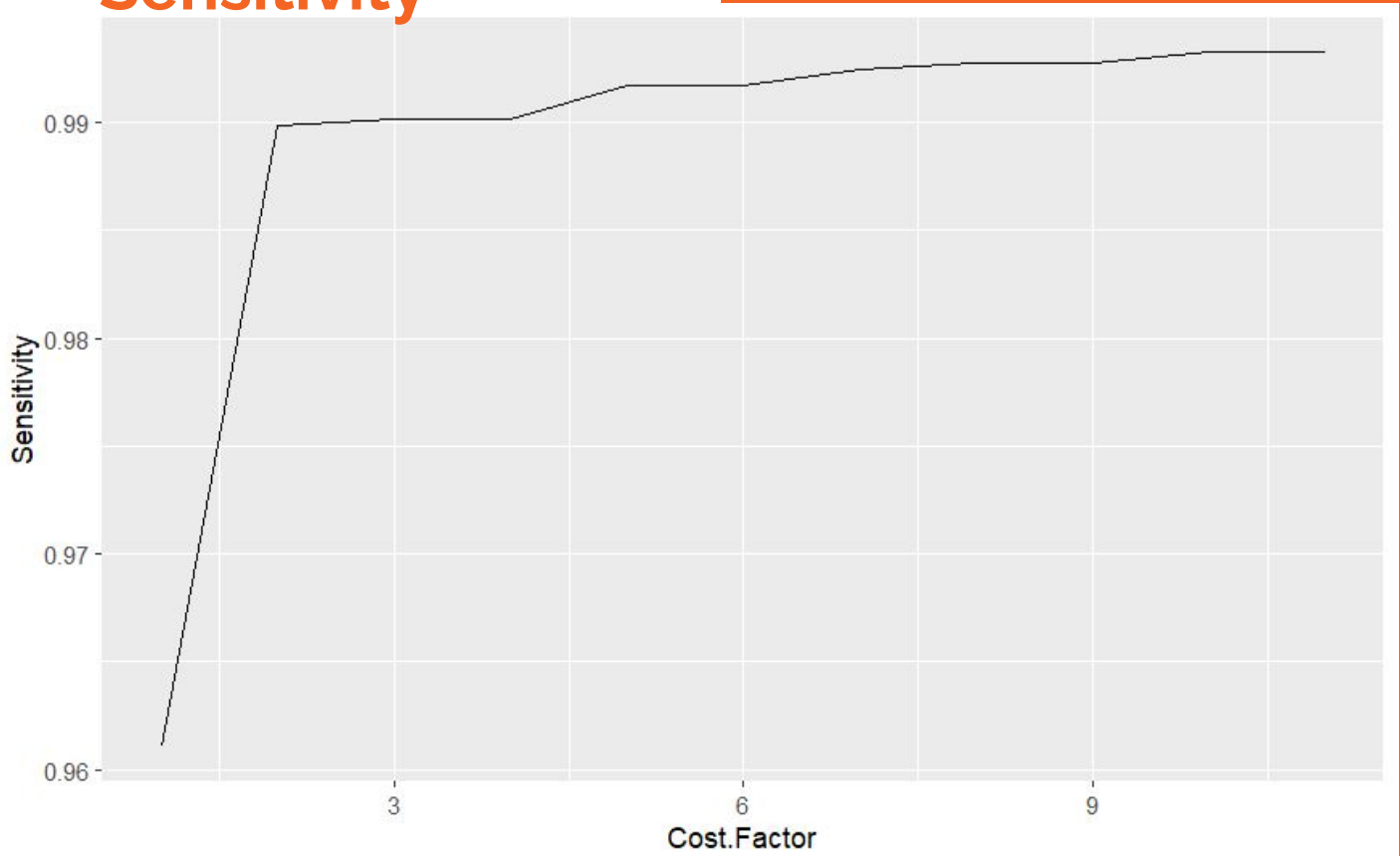
Role of Cost Parameter: Crucial in SVM models for balancing sensitivity and specificity.

Adjusting the Cost Parameter: Influences the model's emphasis on avoiding misclassifications.

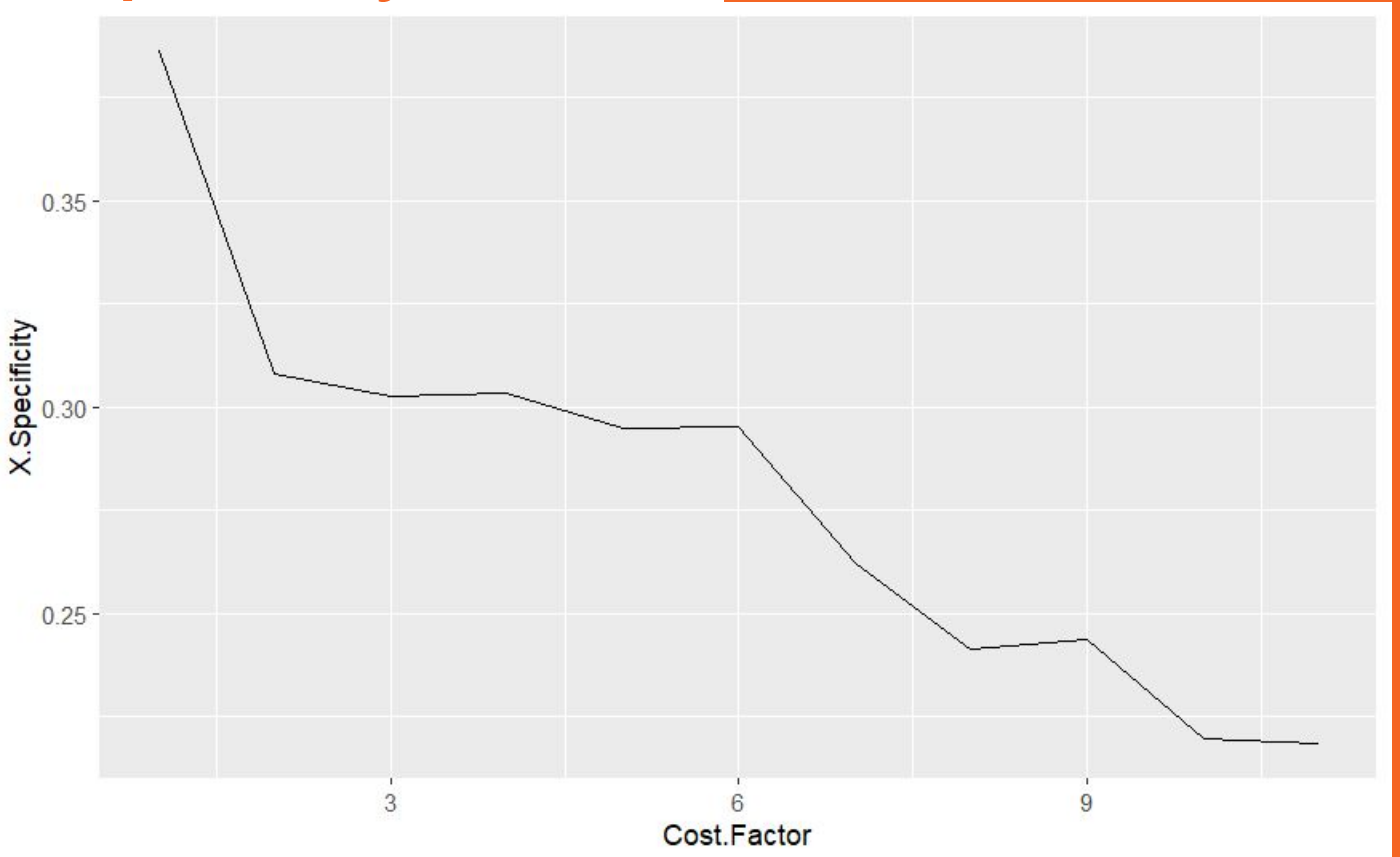
Increased Cost for Negative Class: Leads to more stringent positive predictions, reducing false positives but possibly increasing false negatives.

Decreased Cost for Negative Class: Results in less stringent predictions, decreasing false negatives but potentially increasing ~~false~~ positives.

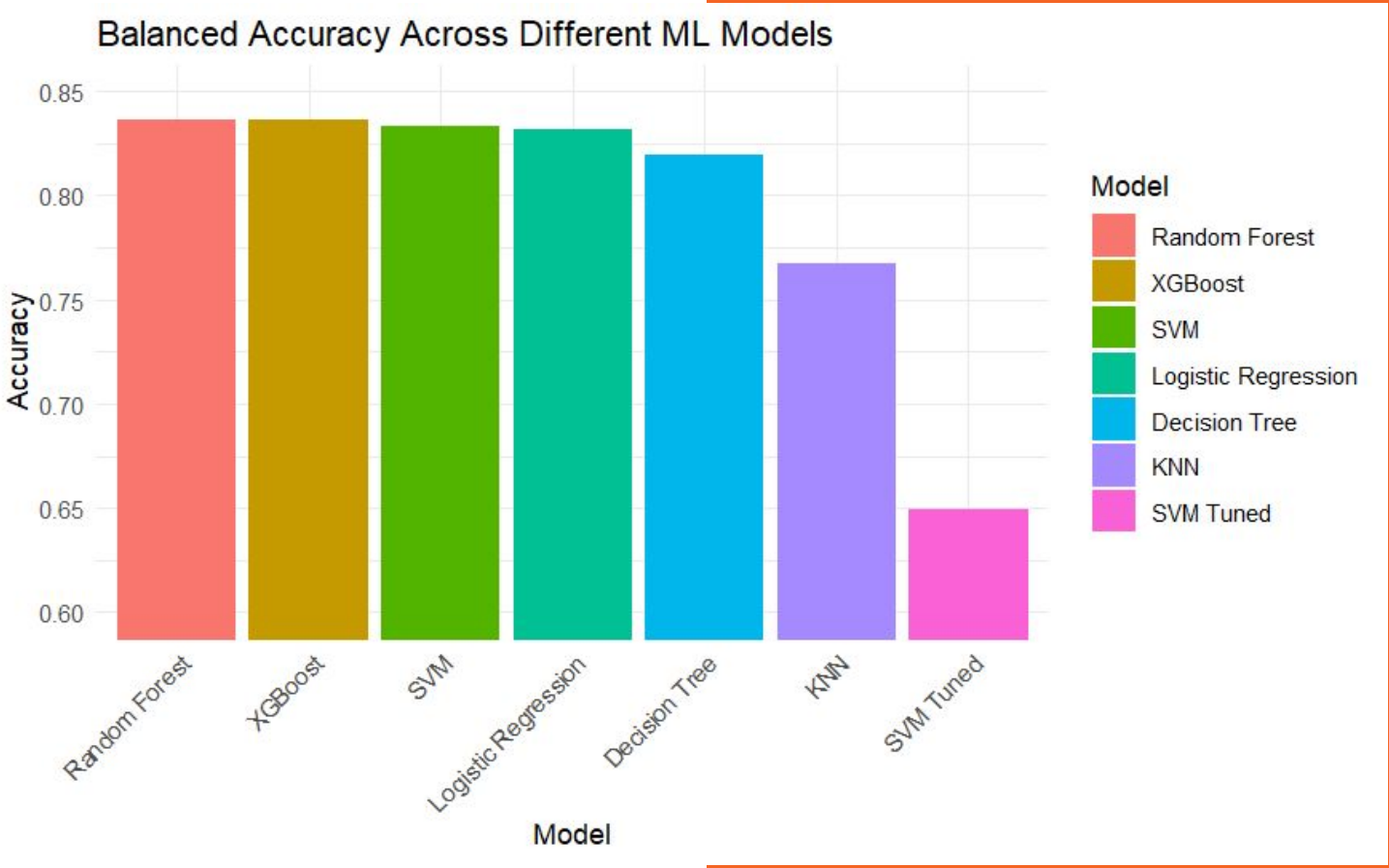
Cost-Factor Analysis: Sensitivity



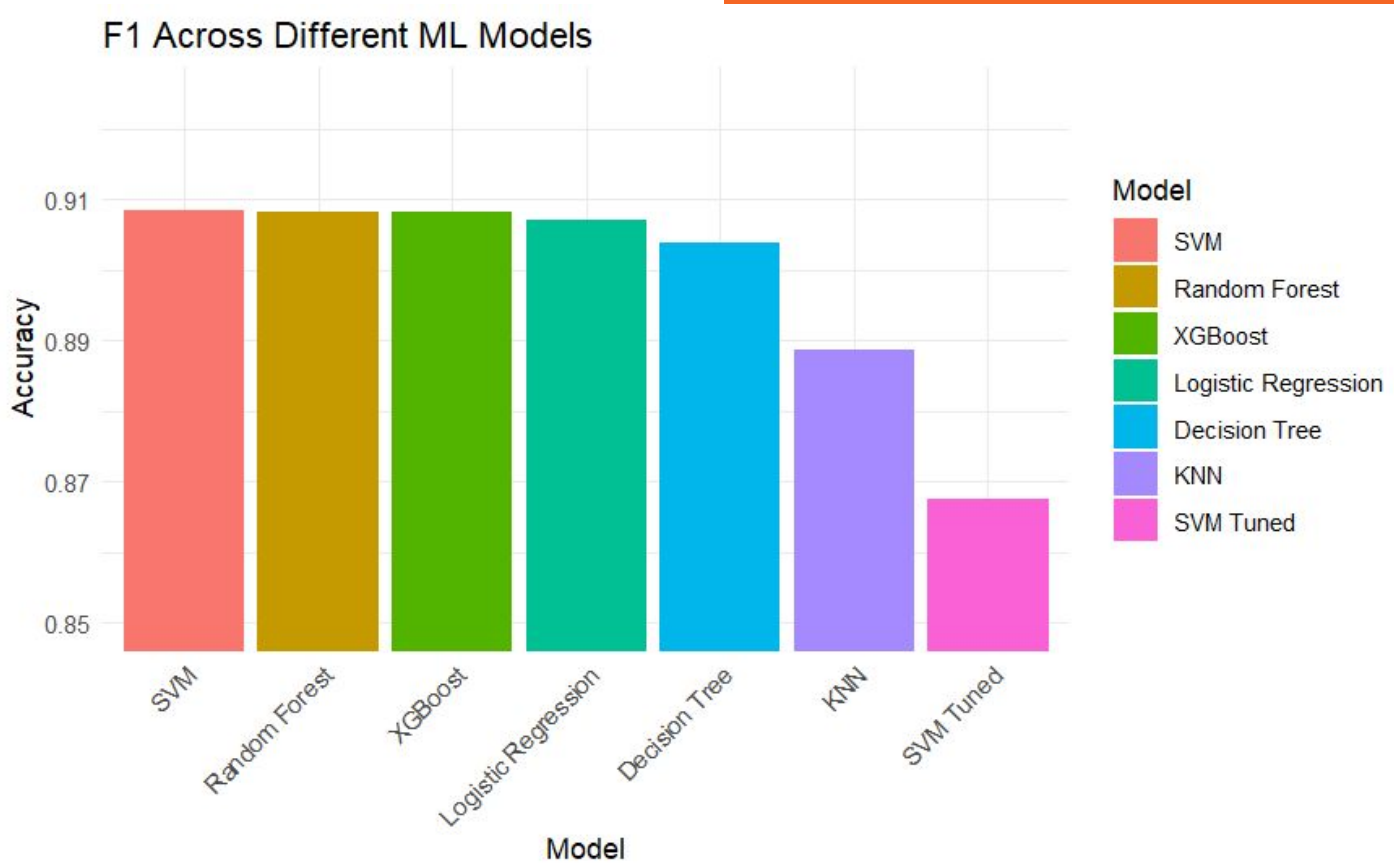
Cost-Factor Analysis: Specificity



Balanced Accuracy Matrics

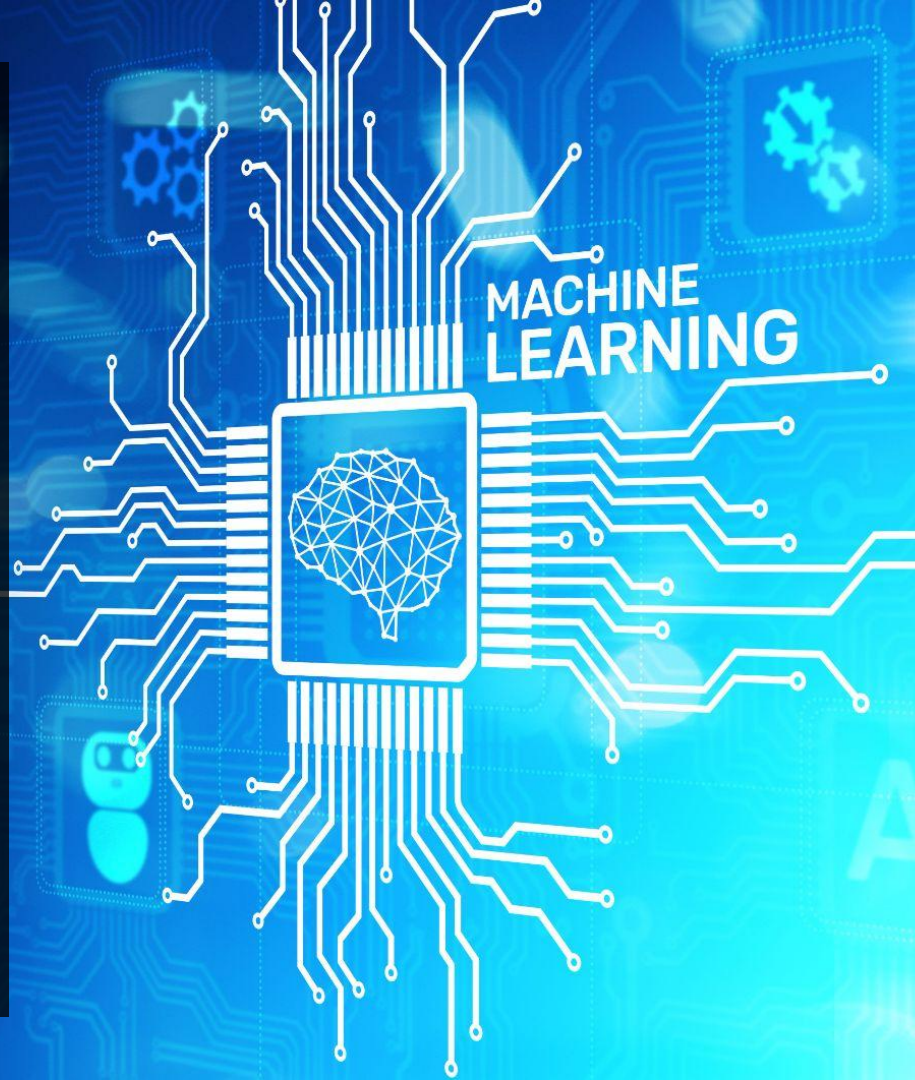


F1-score Matrics



Key Takeaway

Different machine learning models offer distinct advantages depending on the specific needs and circumstances.



What the objective is?

For High Sensitivity Needs:

The tuned SVM model is recommended. This model is highly sensitive and more likely to detect the presence of COVID-19, akin to a sensitive smoke alarm.

For High Specificity Needs:

The Random Forest model is advised. This model acts like a discerning gatekeeper, effectively identifying true cases while minimizing false alarms.

For Balanced Accuracy:

The standard SVM model, with its strong balance of precision and recall, is suitable for environments where both aspects of accurately identifying COVID-19 cases are equally important.

No single model is universally perfect.

Each has its strengths, which can be applied to different scenarios to align with the hospital's goals.

Note:

Different machine learning models offer distinct advantages depending on the specific needs and circumstances.