



[This Photo](https://www.groundup.org.za/article/covid-19-worlds-unprecedented-experiment/) by Unknown Author is licensed under [CC BY-ND](https://creativecommons.org/licenses/by-nd/3.0/)



[This Photo](https://bib.us.es/content/mdpi) by Unknown Author is licensed under [CC BY](https://creativecommons.org/licenses/by/3.0/)

*Article*

***"Tracking COVID-19 Variants in Malaysia with Machine Learning"***

**VADIJE VISHISHTA- 2203A52186**

1 **Affiliation;** [**2203A52186@sru.edu.in**](mailto:2203A52137@sru.edu.in)

**\* Section - AIML AA**

**† SR UNIVERSITY**

**Abstract**: COVID-19 has become the most important global threat by 2020. Among other countries, Malaysia has been affected by the epidemic. In Malaysia, where the number of cases is relatively low, the demographic impact has been significant. Unfortunately, there have been many deaths in many states in Malaysia, although some countries have been able to minimize the impact by following safety measures and precautions and fewer deaths have been reported in Malaysia than in other countries .This progress can be attributed to the significant role played by the Government of Malaysia in implementing precautions and ensuring the safety of its people in various countries. In this work, we aim to use an algorithm to predict mortality in 11 states in Malaysia. These predictions can provide valuable insight into the magnitude of the impact of COVID-19. While it is important to note that the work may not have complete conclusions, it may yield valuable pre-dictions . The main focus of this project is to assess the impact of COVID-19 on mortality rates in the 11 countries under consideration mainly in Malaysia. In this work we are going to use few algorithms (logistic, perceptron, svm- classification, KNN) to predict the results. The best part of the business is that it is located in less than one country. of death than others.

**Keywords:** classification models; backward logistic regression ; support vector machine ; K-nearest neighbors; root mean square error (RMSE) ; Original means complete Error (MAE) .

**Introduction**

**Feature description:** This work helps us identify countries with early COVID-19 mortality in Malaysia by identifying a few important indicators. In today’s world, where so many people are educated and informed, Malaysia’s mortality situation can easily be understood. The project has two main objectives: to help people learn about the deaths of affected people in Malaysian states during COVID-19 and how people took precautions to survive the pandemic situation. And they can address future generations to protect them from pandemic COVID-1 They can also take care of themselves and reduce their chances of being infected with COVID-19.

The Coronavirus Disease 2019 (COVID-19) pandemic was declared by the World Health Organization (WHO) in early 2020 million people Live in the world. Through this project, people will be able to see the number of countries affected by COVID-19 in Malaysia. Additional precautions can be taken for the future. This project demonstrates how we can use knowledge and technology to improve health in the modern world.

**Present Generation Uses:** Malaysia's COVID-19 death toll is lower than the global average; This is due to early preparation and planning, effective implementation of public health and clinical systems, increased monitoring of communications, prompt case discovery and aggressive escalation of MCOs.

The COVID-19 epidemic is the largest outbreak of infectious disease in Malaysia since the Spanish flu of 1918 which killed 34,644 people or 1% of British Malays at the time In the 1999 Nipah virus outbreak 105 Malaysia -People a are in the country dead, with only 2 people losing their lives in the 2003 SARS epidemic. The COVID-19 pandemic has killed more than 100 people in Malaysia so far. Malaysia has had two waves of COVID-19.

**Literature review:**

Research has investigated the epidemiological characteristics of COVID-19 in Malaysia, including the course of the disease, the susceptibility of the disease and the factors that contribute to its spread Studies typically report on infection rates, testing and contact tracing methods, and the role of asymptomatic vectors in infection.

The study focuses on Malaysia’s public health response, including the implementation of measures such as curfew periods (lockdown), mask orders, social distancing and vaccination campaigns and improved

Research frequently examines the impact of these factors on caseloads and health care.

The literature has examined the readiness and capacity of the Malaysian healthcare system to deal with COVID-19 cases, including ICU beds, ventilators and healthcare professional supplies

Research could also explore strategies for optimizing health care during pandemics.

The recent study focuses on COVID-19 vaccination in Malaysia, including vaccine distribution, vaccine delays and the impact of vaccination on cases, hospitalizations and mortality reduction

|  |  |
| --- | --- |
| **Algorithm** | **Accuracy** |
| **Project 1:**  SVM (support vector machine)  **Project 2:**  Logistic Regression  KNN (K-nearest neighbor)  SVM (Support Vector Machine) | 0.9954268292682927  0.166215121  **0.16621951219512193**  **0.9954268292682927** |

**Few results on sam****e:**

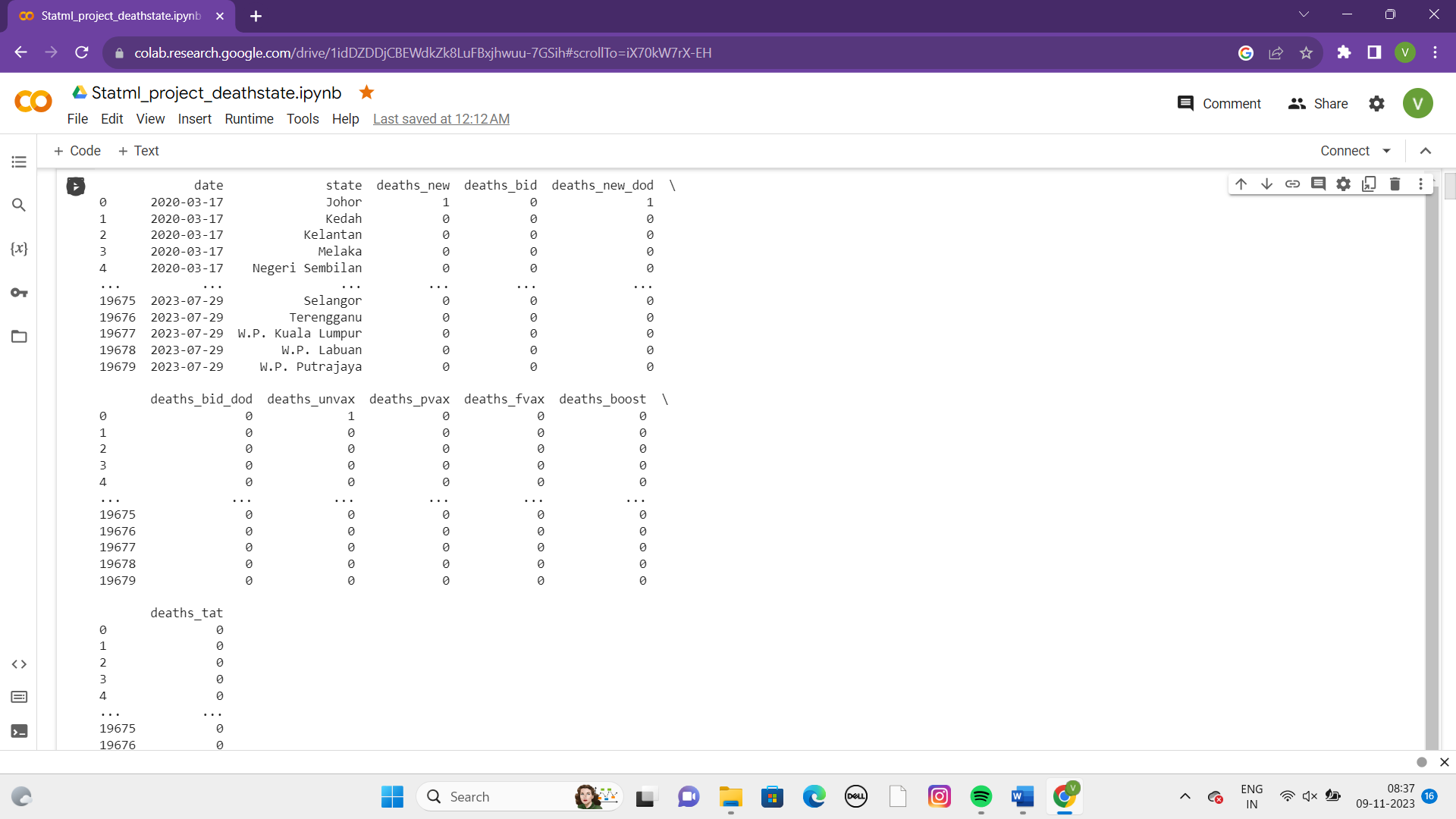
**1.Methodology**

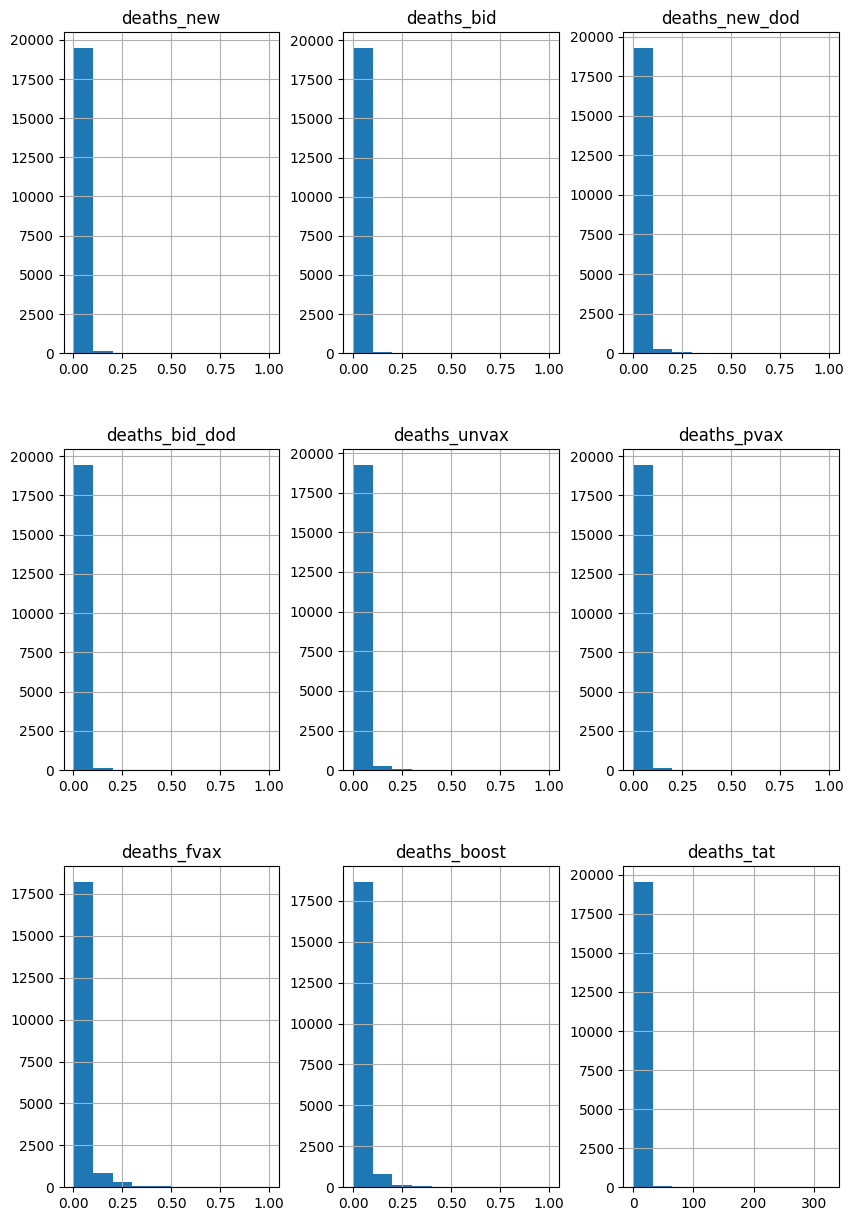
* 1. **Dataset:** This data set consists of 19680 rows and 11 columns.

**Target Variable:** death\_ tat (total number of deaths)

**Features:** deaths\_ new, deaths\_ bid, deaths\_ new\_ dod , deaths\_ bid \_ dod, deaths\_

unvax, deaths\_ pvax, deaths\_ fvax, deaths\_ boost, deaths\_ tat





**Fig 1: Visualization of data**

**Preprocessing**

**Normalization:** This data set requires normalization i.e making the values in the form of 0’s and 1’s .This process includes many methods min max scalar is one of them.

**Filling nan values:** The data set contains few missing values these are filled with 0’s by implementing the below code.

**Drop:** When we want to eliminate any column we use this process**.**

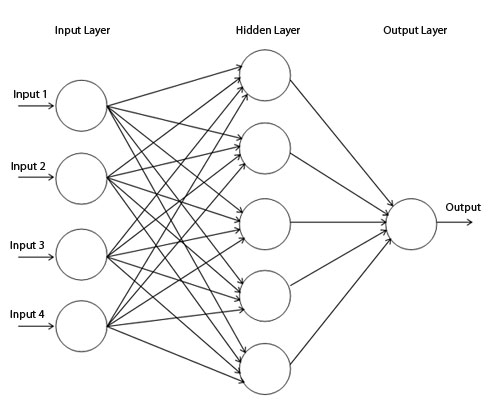
**Implementation:**

**Logistic Regression:** Logistic regression resembles a selection tool in many areas. Imagine you want to know if an email is spam or not. Logistic regression helps you with that. It calculates the probability that something is in one of two categories such as 0 or 1, yes or no, true or false. Then, it makes a decision based on the rule, e.g., if the probability is greater than 0.5, it says "yes", otherwise it says "no". People use the tool to diagnose diseases in medicine, filter spam in email systems, and analyze credit risk in the financial sector. It is a simple but practical way of making decisions with only two possible outcomes.

**Perceptron:** A perceptron is a supervised machine learning tool that can be used to quantify the adequacy of predictions. It basically consists of four steps and is commonly used for binary classification, which determines whether a given input belongs to a particular class or not in the first step, it fetches input data from the dataset, and afterwards is initialized with random weighting.

In the next step, the bias setting is applied, and then the activation function is applied. The result of the activation function is the final prediction. The default activation function is sigmoid, and results range from 0 to 1. The perceptron calculates the accuracy of the data set and performs a supervised learning process

Although it includes a hidden layer, it is also known as a monolayer neural network. Please note that some grammatical errors have been corrected in this answer.



**Support Vector Machine (SVM):** A support vector machine (SVM) is a supervised learning algorithm for assessing the accuracy of a model. In this method, the dependent variable and the target variable are plotted on a graph. Of all the lines, one of them is considered the best fit line, and is known as the "maximum margin over-plane", not the "maximum". Points adjacent to this hyperplane are known as "support vectors,". When the data set contains three variables, the result is usually a 2D model. SVMs are powerful tools for classification and regression applications, and they excel in finding optimal decision boundaries by maximizing the differences between classes.

**K-Nearest Neighbor:** k-Nearest Neighbors (KNN) is a simple but effective technique that can be used to predict diabetes by classification

KNN is a non-parametric algorithm that classifies data points based on their similarity to other data points in the training set. It uses distance measures (e. g. Euclidean distance) to measure similarity between data points.

Set the value of "k", which represents the number of nearest neighbors to consider. Choosing the right "k" value is important; You can experiment with different values ​​and use techniques like cross-validation to find the best results.

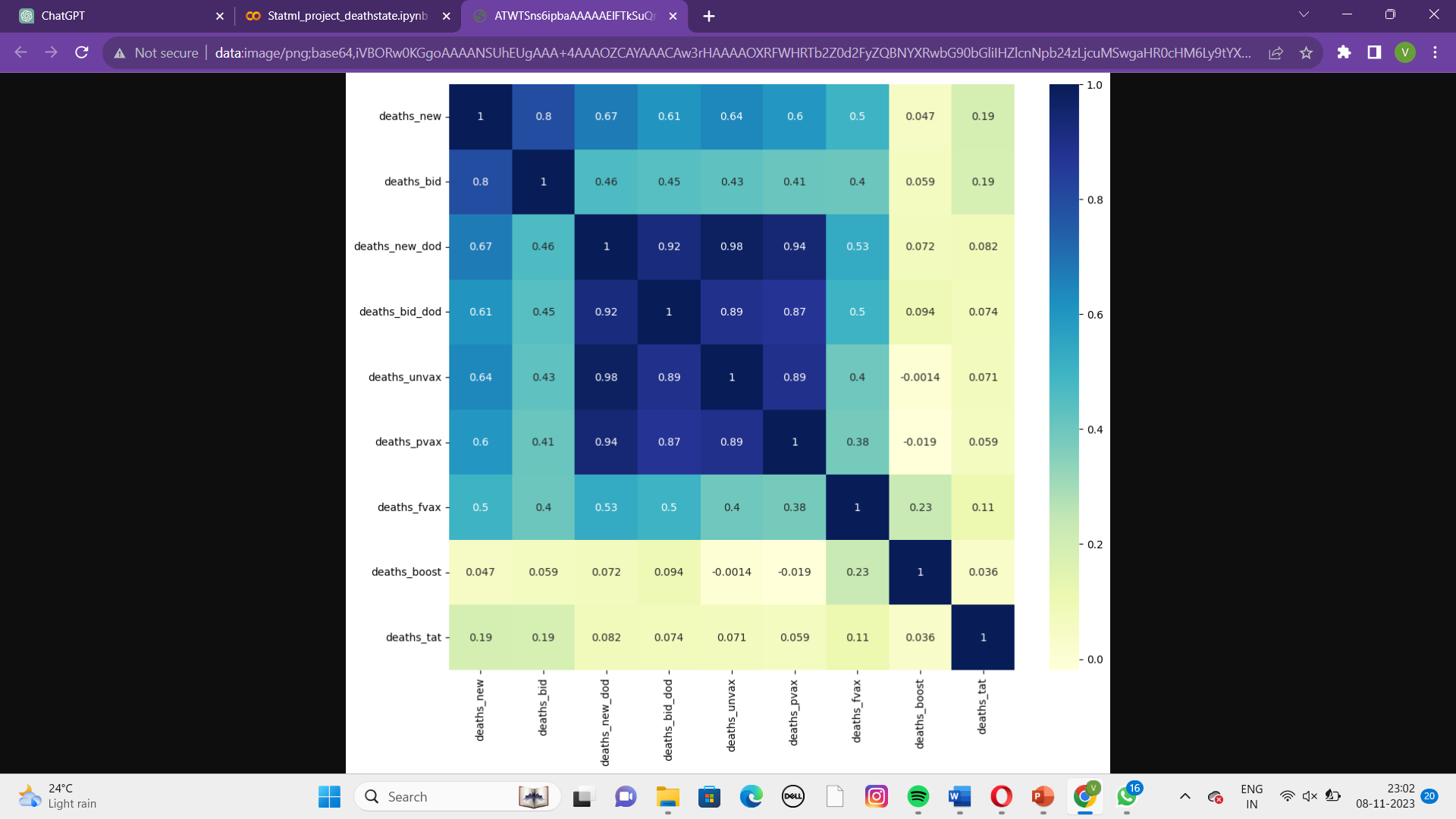
KNN models define decision boundaries based on the similarity of data points. Visualizing these boundaries can help to understand how the model separates the two groups (diabetic and nondiabetic).

The trained KNN model can be incorporated into clinical decision support systems to help healthcare professionals assess an individual's risk of diabetes and determine the need for further screening or intervention.

KNN is a versatile and easy-to-interpret system that can be used for diabetes prediction. However, choosing an appropriate "k" value and dealing with class imbalances are important considerations for good results in practice.

**Confusion matrix for KNN:** In K-Nearest Neighbors (KNN) classification, you can construct a confusion matrix to evaluate the performance of your model. The confusion matrix is ​​a table that helps you understand how well the model classifies observations as true positives, true negatives, false positives, and false negatives. You can use libraries in Python to calculate and build the confusion matrix, such as scikit-learn.

The confusion matrix is ​​a valuable tool to evaluate the performance of classification models, as it provides insights into accuracy, precision, recall, and F1-score among other metrics Helps you understand how the model distinguishes good learning from sins well below and any biases or issues in its forecasts or if so



**Boot strap:**

Bootstrapping provides a way to measure the variability and robustness of your models. It helps in assessing how the performance of the model may change when applied to different subsets of data. Additionally, bootstrapping can be particularly useful when dealing with limited data or when you want to quantify the uncertainty associated with your predictions.

Although bootstrapping offers advantages, it is important to thoroughly examine and validate the results of the bootstrapped model and ensure that the chosen method of collection (e.g., polling, averaging) is you the specific nature of the diabetes prognosis problem corresponds.

**Bootstrapping for Logistic Regression:**

Bootstrap samples are generated from the training data set by replacing randomly selected data points.

Training multiple logistic regression models on these bootstrapped samples yields a range of models.

The final prediction is usually obtained by pooling predictions from these models, which can reduce overfitting and improve overall model effectiveness.

**Bootstrapping for SVM Classification:**

Bootstrapping can be used to generate multiple subsamples of the training data for SVMs.

Each subset is used to train the SVM model separately.

Combining predictions from these models, such as by analysis or averaging, can lead to more robust predictions, especially when dealing with noisy or unbalanced data

**Bootstrapping for Perceptron:**

When using Perceptron, bootstrapping can be incorporated by sampling repeated data sets.

Several perceptron models are trained on these resampled data sets.

A final prediction can be made by considering the output of each perceptron, which can improve the performance of the model.

**Bootstrapping for k-Nearest Neighbors (KNN):**

Bootstrapping in KNN can help in assessing the stability and improvement of the robustness of the model.

Several KNN models can be trained on bootstrapped datasets.

**RESULT ANALYSIS:**

The result accuracies we finally got are:

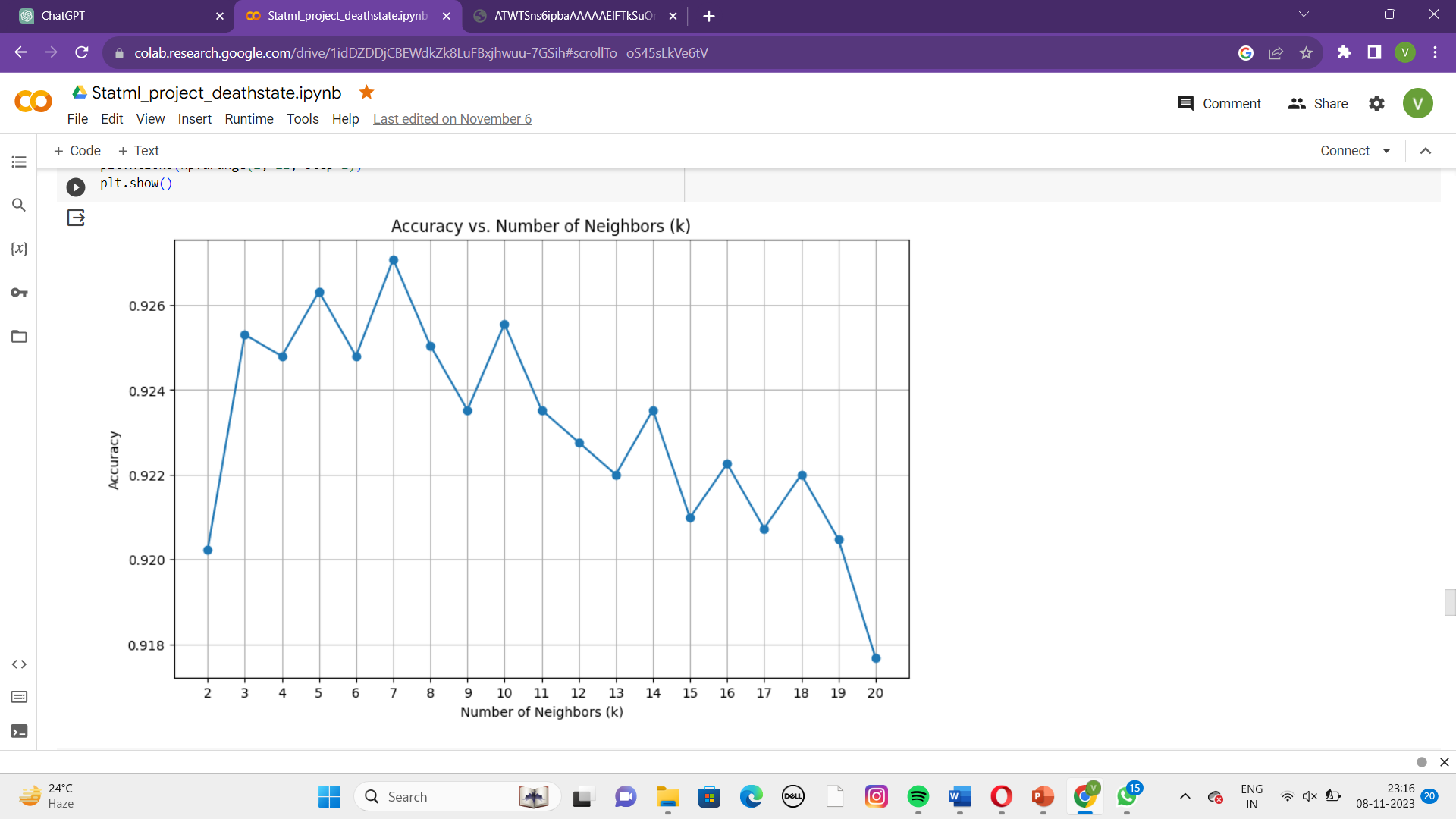
Logistic Regression: 0.166215121

SVM: **0.9954268292682927**

Perceptron Learning: 0.9924589682

KNN: k= 2 is the most accurate value for the given dataset.

Fig 2 shows the graph for KNN k values accuracy



finally after bootstrapping is applied we got the following graphs for the 4 models:

Fig-3 Logistic Regression, Fig-4 SVM, Fig-5 Perceptron Learning, Fig-6 KNN

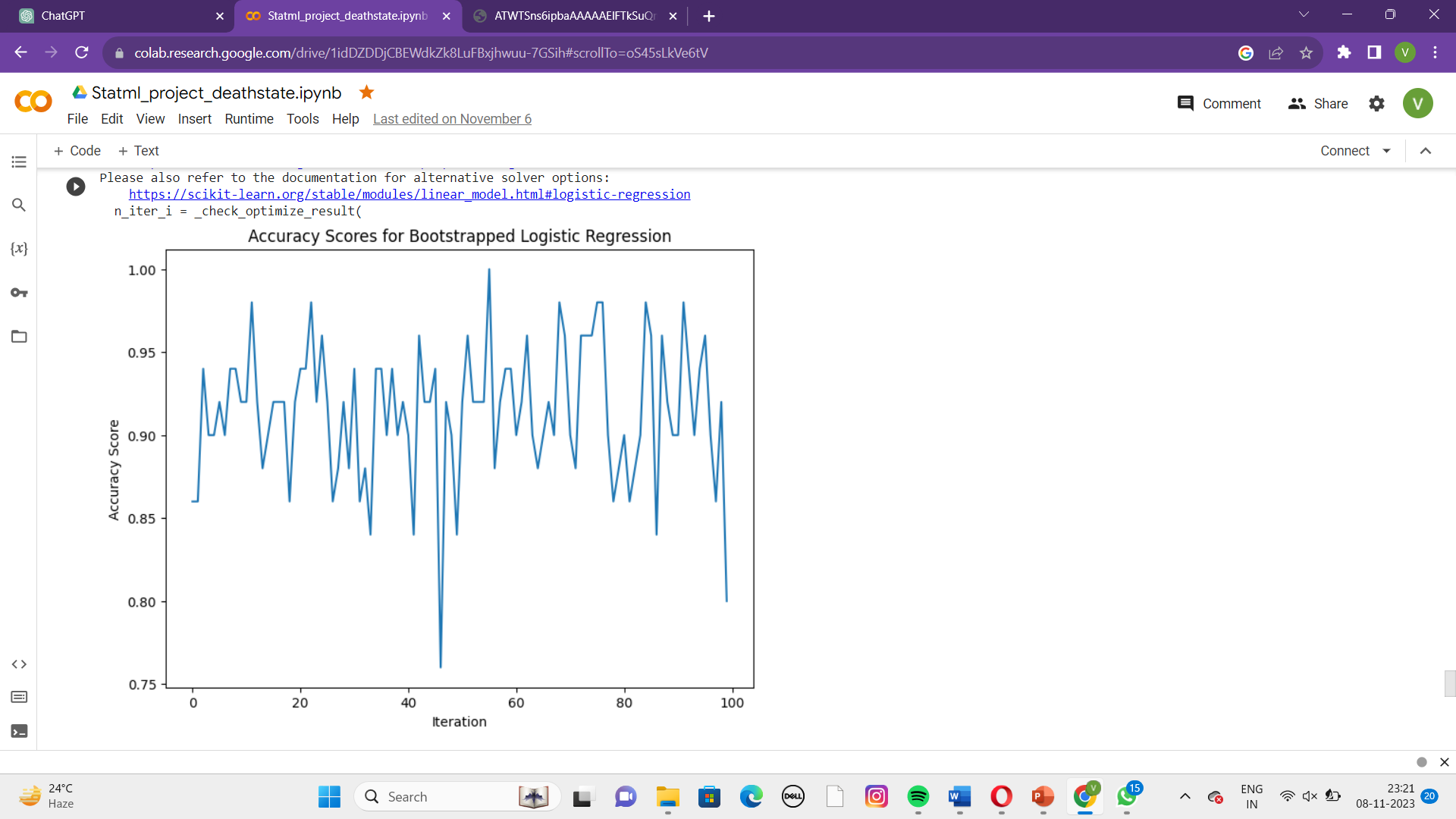


Fig 3

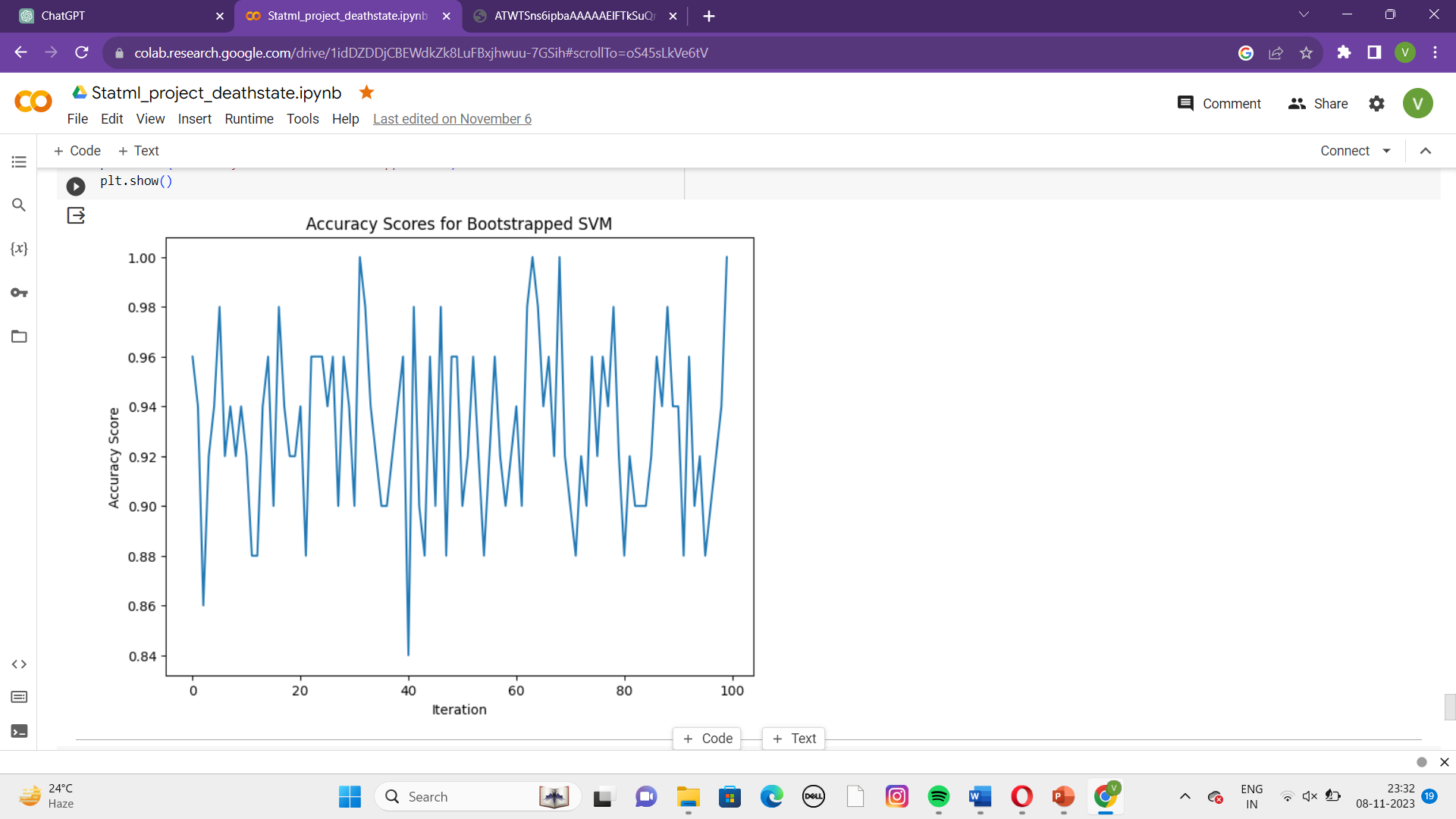


Fig 4

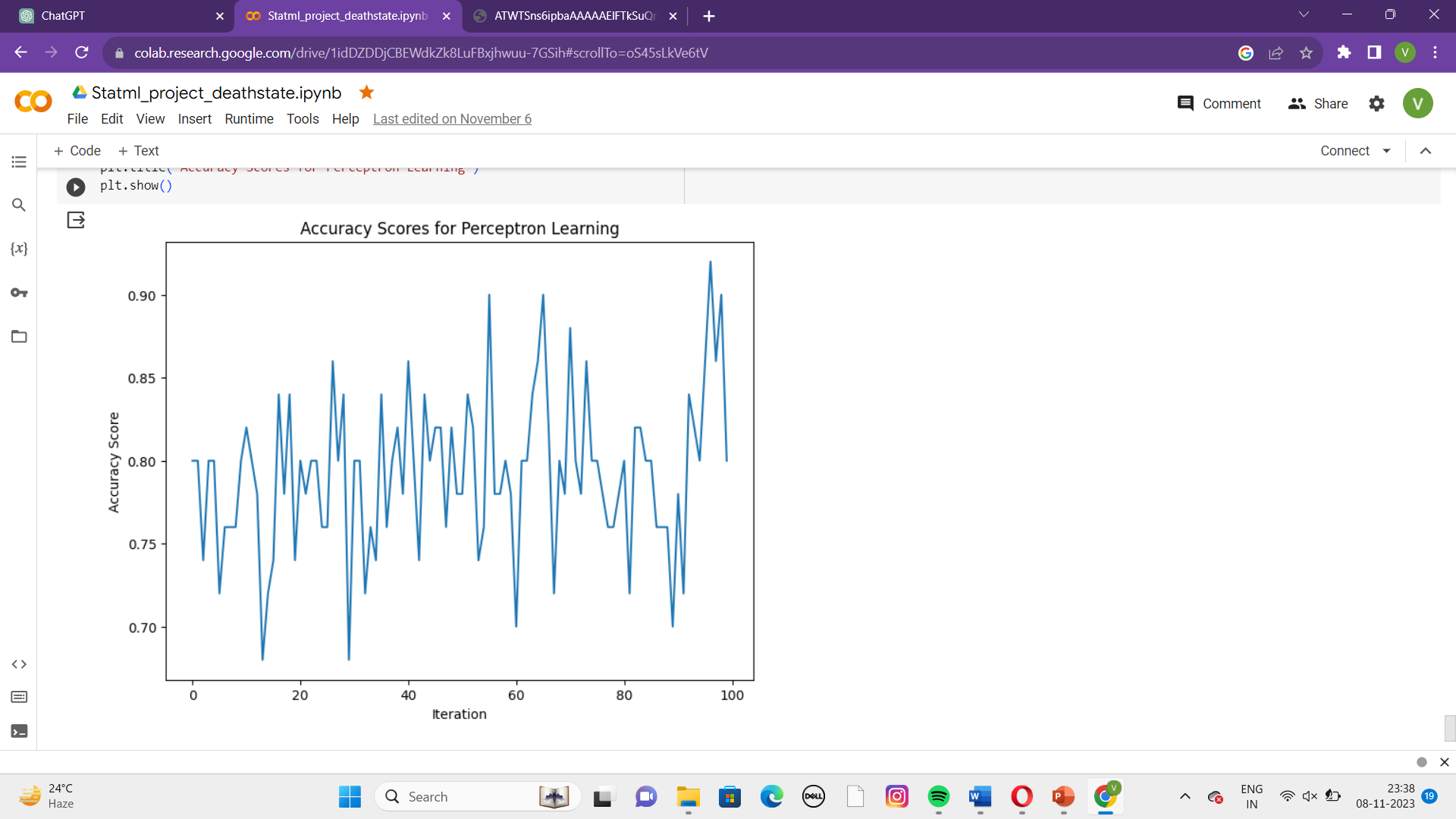
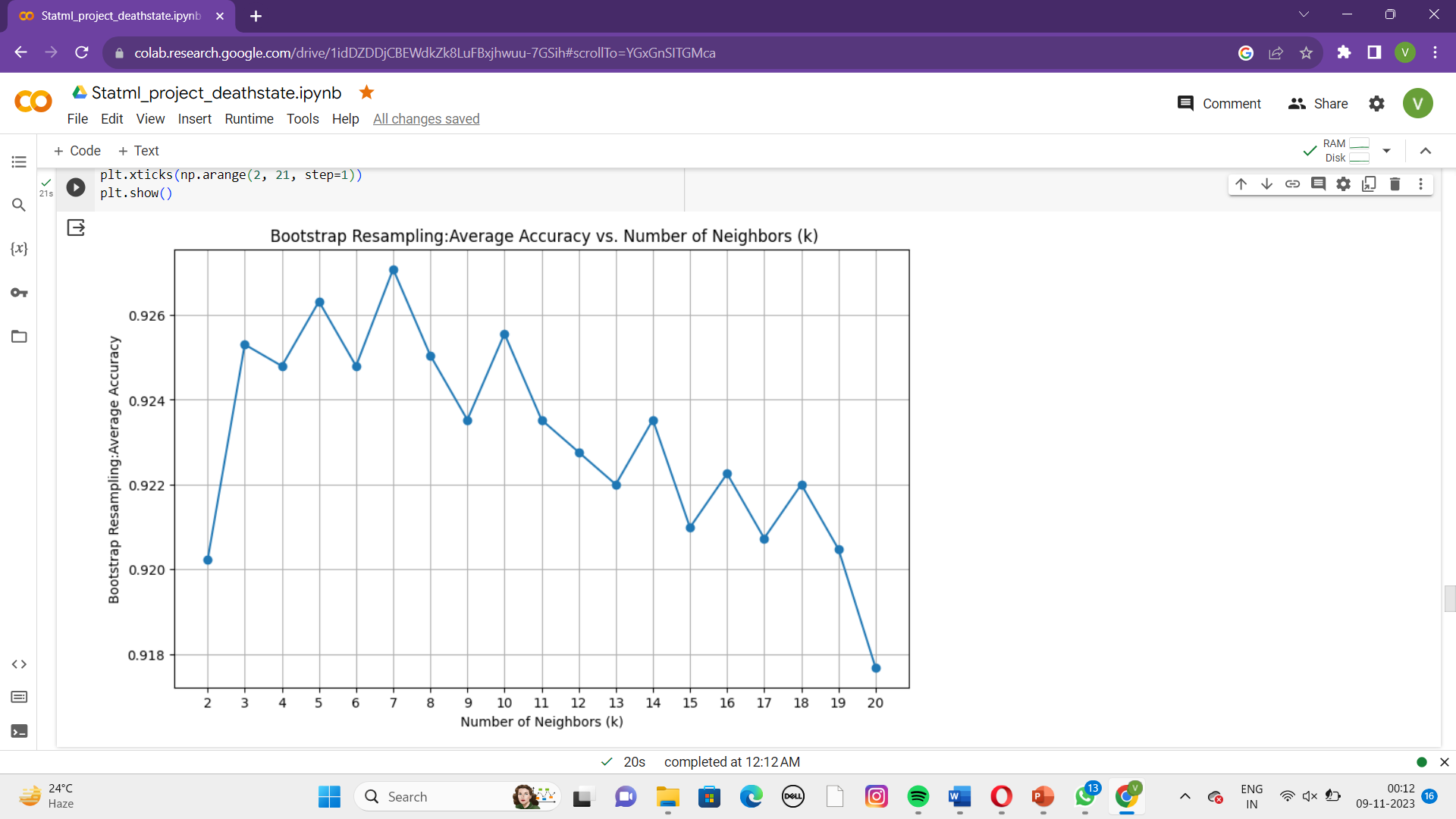


Fig 5

**

*Fig:6*

**Conclusion:**

*The algorithm provides the information about the death cases of the different states in Malaysia Country.Here,we can observe that the Government of Malaysia helped the people to take care of themselves from the COVID-19.Due to the precautions follwed by the people there are less no.of death cases in different states of Malaysia.In this dataset KNN outperforms other algorithms,making it the most efficient method.*

*It is important to note that the optimal algorithm may be different for different types of data. i.e the best method may change depending on the specific characteristics and requirements of each data system.*

***ACKNOWLEDGEMENTS:***

*I thanks SR University, Hassanparty, Faculty for collecting this.*

***Reference:***

1. <https://www.kaggle.com/datasets/iammustafatz/malaysia-covid-19-dataset>
2. <https://ieeexplore.ieee.org/document/9430222>
3. <https://medic.upm.edu.my/upload/dokumen/2022042811304814_MJMHS_0675.pdf>

d. <http://eprints.intimal.edu.my/1526/1/vol.2021_003.pdf>