.

Department Of Computer Engineering

**STES’S SINHGAD ACADEMY OF ENGINEERING**

**KONDHWA BK, PUNE 411048**

**2022-2023**

**“Titanic Survivor Prediction Model”**

**Submitted to the**

**Savitribai Phule Pune University**

**In partial fulfillment for the award of the Degree of**

**Bachelor of Engineering**

**in**

**Computer Engineering**

**By**

1. **Cedrick Andrade COBA006**
2. **Aishwarya Diwane COBA015**
3. **Pranav Chavan COBA109**
4. **Tanmay Dabhade COBA088**

Under the guidance of

**Prof. A. Shapurkar**



**CERTIFICATE**

This is to certify that the mini project report entitled **“Titanic Survivor Prediction Model”** being submitted by **Cedrick Andrade COBA006, Aishwarya Diwane COBA015, Pranav Chavan COBA109, Tanmay Dabhade COBA088** is a record of bonafide work carried out by him/her under the supervision and guidance of **Prof. A. Shapurkar** in partial fulfillment of the requirement for **BE (Computer Engineering) – 2019 course** of Savitribai Phule Pune University, Pune in the academic year 2022-2023.

Date:

Place: Pune

Subject Coordinator Head of the Department

Principal

This Mini Project report has been examined by us as per the Savitribai Phule Pune University, Pune requirements at **SINHGAD ACADEMY OF ENGINEERING** Pune – 411048 on . . . . . . . . . .

Internal Examiner External Examiner

**ACKNOWLEDGEMENT**

First and foremost, praises and thanks to the God, the Almighty, for showers of blessings throughout my project work to complete the research successfully.

I would like to express my deep and sincere gratitude to my subject teacher Prof. A. Shapurkar for giving us the opportunity to do this project and provide invaluable guidance throughout this project. Her dynamism, vision, sincerity and motivation have deeply inspired us. He has taught us the methodology to carry out the research and to present the project works as clearly as possible. It was a great privilege and honor to work and study under her guidance. We are extremely grateful for what she has offered us. We would also like to thank him for his friendship, empathy, and great sense of humor.

We are extremely grateful to all group members Cedrick Andrade, Aishwarya Diwane, Pranav Chavan, Tanmay Dabhade for the dedication and consistency towards this mini project. And also thankful for all the resources which are provided by each group member and which played a very crucial role in the accomplishment of this project.

(Students Name & Signature)

**CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Sr. No** | **TITLE** | **Page no** |
| **1.** | Abstract | **5** |
| **2.** | Introduction | **6** |
| **3.** | Problem Statement | **7** |
| **4.** | Motivation | **7** |
| **5.** | Objectives | **7** |
| **6.** | Theory | **8** |
| **7.** | Conclusion | **18** |
| **8.** | References | **18** |

**Abstract**

On April 15, 1912, during her maiden voyage, the widely considered “unsinkable” RMS Titanic sank after colliding with an iceberg. Unfortunately, there weren’t enough lifeboats for everyone onboard, resulting in the death of 1502 out of 2224 passengers and crew.

While there was some element of luck involved in surviving, it seems some groups of people were more likely to survive than others.

In today’s digital world, using Machine Learning we can implement models that are capable of predicting outcomes based on past events. Using the Titanic Training Dataset we predict if a Passenger in the Testing Dataset will survive using Supervised Learning and Classification Techniques to build the model.

**Introduction**

The data has been split into two groups: training set (train.csv) test set (test.csv)

The training set should be used to build your machine learning models. For the training set, the outcome has been provided(also known as the “ground truth”) for each passenger. The model will be based on “features” like passengers’ gender and class.

The test set should be used to see how well your model performs on unseen data. For the test set, the ground truth is not provided for each passenger. It is our job to predict these outcomes. For each passenger in the test set, we use the model we trained to predict whether or not they survived the sinking of the Titanic.

We use Decision Tree Classification for our Machine Learning Model. It has an Accuracy of 77.52% from the training data. Other techniques like Label Encoding and Feature Selection are also used.

We follow the Data Science Lifecycle to build the model.

**Problem Statement**

Build a machine learning model that predicts the type of people who survived the Titanic shipwreck using passenger data (i.e. name, age, gender, socio-economic class, etc.).

Dataset Link: https://www.kaggle.com/competitions/titanic/data

**Motivation**

The ability to predict the future based on past data and events has always been an interesting topic in todays world. Using ML classification techniques and the Titanic dataset, we were confident that we could build a prediction model to predict the survivors the Titanic Disaster. Succeeding in this would allow applications of our model to predict and maybe prevent future disasters with slight tweaks to the source code.

**Objective**

* To predict the survivors of the Testing Dataset by building a model using the Training Dataset of the Titanic Event.

**Theory**

**Data Analytics Lifecycle :**

The Data analytic lifecycle is designed for Big Data problems and data science projects. The cycle is iterative to represent real project. To address the distinct requirements for performing analysis on Big Data, step – by – step methodology is needed to organize the activities and tasks involved with acquiring, processing, analyzing, and repurposing data.

**Phase 1: Discovery:**-

The data science team learn and investigate the problem.

Develop context and understanding.

Come to know about data sources needed and available for the project.

The team formulates initial hypothesis that can be later tested with data.

**Phase 2: Data Preparation:-**

Steps to explore, preprocess, and condition data prior to modeling and analysis.

It requires the presence of an analytic sandbox, the team execute, load, and transform, to get data into the sandbox.

Data preparation tasks are likely to be performed multiple times and not in predefined order.

**Phase 3: Model Planning:-**

Team explores data to learn about relationships between variables and subsequently, selects key variables and the most suitable models.

In this phase, data science team develop data sets for training, testing, and production purposes.

Team builds and executes models based on the work done in the model planning phase.

**Phase 4: Model Building:-**

Team develops datasets for testing, training, and production purposes.

Team also considers whether its existing tools will suffice for running the models or if they need more robust environment for executing models.

**Phase 5: Communication Results:-**

After executing model team need to compare outcomes of modeling to criteria established for success and failure.

Team considers how best to articulate findings and outcomes to various team members and stakeholders, taking into account warning, assumptions.

Team should identify key findings, quantify business value, and develop narrative to summarize and convey findings to stakeholders.

**Phase 6: Operationalize:-**

The team communicates benefits of project more broadly and sets up pilot project to deploy work in controlled way before broadening the work to full enterprise of users.

This approach enables team to learn about performance and related constraints of the model in production environment on small scale , and make adjustments before full deployment.

The team delivers final reports, briefings, codes.

**Libraries Used:**

**Pandas:**

Pandas is an open-source library that is built on top of NumPy library. It is a Python package that offers various data structures and operations for manipulating numerical data and time series. It is mainly popular for importing and analyzing data much easier. Pandas is fast and it has high-performance & productivity for users.

You can import it by the following code: import pandas as pd

We use Pandas for forming the data frame for the analysis in our project. We read the dataset via the following code: df = pd.read\_csv('covid\_vaccine\_statewise.csv',header=[0])

We also use other commands from this library to explore the dataset once its loaded onto the dataframe.

Example:- df.shape, df.columns, ,df.dtypes, df.head(), df.describe(include='all')

**Matplotlib:**

Matplotlib is an amazing visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack. It was introduced by John Hunter in the year 2002.

One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, histogram etc.

You can import it by the following code: import matplotlib.pyplot as plt

**Seaborn:**

Seaborn is an amazing visualization library for statistical graphics plotting in Python. It provides beautiful default styles and color palettes to make statistical plots more attractive. It is built on the top of matplotlib library and also closely integrated to the data structures from pandas.

Seaborn aims to make visualization the central part of exploring and understanding data. It provides dataset-oriented APIs, so that we can switch between different visual representations for same variables for better understanding of dataset.

You can import it by the following code: import seaborn as sns

We use the bar plot from the Seaborn library to visualize the covid-19 vaccination status statewise in our project. Example:- sns.barplot( x, y, data)

**Scikit:**

Scikitlearn is an open-source Python library that implements a range of machine learning, pre-processing, cross-validation, and visualization algorithms using a unified interface.

Important features of scikit-learn:

* Simple and efficient tools for data mining and data analysis. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means, etc.
* Accessible to everybody and reusable in various contexts.
* Built on the top of NumPy, SciPy, and matplotlib.
* Open source and commercially usable.

We use the following packages from Scikit Learn in our program.

* from sklearn.preprocessing import LabelEncoder
* from sklearn.model\_selection import train\_test\_split
* from sklearn.metrics import accuracy\_score, classification\_report, confusion\_matrix
* from sklearn.ensemble import DecisionTreeClassifier

**Preprocessing:-**

* **Label Encoder**

Label Encoding refers to converting the labels into a numeric form so as to convert them into the machine-readable form. Machine learning algorithms can then decide in a better way how those labels must be operated. It is an important pre-processing step for the structured dataset in supervised learning.

**Model Selection**

* **Train Test Split**

The train-test split is used to estimate the performance of machine learning algorithms that are applicable for prediction-based Algorithms/Applications. This method is a fast and easy procedure to perform such that we can compare our own machine learning model results to machine results. By default Test set is split into 30 % of actual data and the Training set is split into 70% of the actual data.

We need to split a dataset into train and test sets to evaluate how well our machine learning model performs. The train set is used to fit the model, the statistics of the train set are known. The second set is called the test data set, this set is solely used for predictions.

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X,y,test\_size=0.3,random\_state=10)

**Metrics**

* **Classification Report**

It is the report which explains everything about the classification. This is the summary of the quality of classification made by the constructed ML model. It comprises mainly 5 columns and (N+3) rows. The first column is the class label’s name and followed by Precision, Recall, F1-score, and Support. N rows are for N class labels and other three rows are for accuracy, macro average, and weighted average.

**Precision:** It is calculated with respect to the predicted values. For class-A, out of total predictions how many were really belong to class-A in actual dataset, is defined as the precision. It is the ratio of [i][i] cell of confusion matrix and sum of the [i] column.

precision = (TP) / (TP+FP)

**Recall:** It is calculated with respect to the actual values in dataset. For class-A, out of total entries in dataset, how many were actually classified in class-A by the ML model, is defined as the recall. It is the ratio of [i][i] cell of confusion matrix and sum of the [i] row.

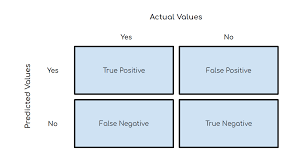
recall = (TP) / (TP+FN)

**F1-score:** It is the harmonic mean of precision and recall.

**Support:** It is the total entries of each class in the actual dataset. It is simply the sum of rows for every class i.

* **Confusion Matrix**

A much better way to evaluate the performance of a classifier is to look at the confusion matrix. The general idea is to count the number of times instances of class A are classified as class B.



**Decision Tree Classifier**

Decision Tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.



A decision tree for the concept PlayTennis.

**Construction of Decision Tree:**

A tree can be “learned” by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called recursive partitioning. The recursion is completed when the subset at a node all has the same value of the target variable, or when splitting no longer adds value to the predictions. The construction of a decision tree classifier does not require any domain knowledge or parameter setting, and therefore is appropriate for exploratory knowledge discovery. Decision trees can handle high-dimensional data. In general decision tree classifier has good accuracy. Decision tree induction is a typical inductive approach to learn knowledge on classification.

**Decision Tree Representation:**

Decision trees classify instances by sorting them down the tree from the root to some leaf node, which provides the classification of the instance. An instance is classified by starting at the root node of the tree, testing the attribute specified by this node, then moving down the tree branch corresponding to the value of the attribute as shown in the above figure. This process is then repeated for the subtree rooted at the new node.

The decision tree in above figure classifies a particular morning according to whether it is suitable for playing tennis and returns the classification associated with the particular leaf.(in this case Yes or No).

**Gini Index:**

Gini Index is a score that evaluates how accurate a split is among the classified groups. Gini index evaluates a score in the range between 0 and 1, where 0 is when all observations belong to one class, and 1 is a random distribution of the elements within classes. In this case, we want to have a Gini index score as low as possible. Gini Index is the evaluation metrics we shall use to evaluate our Decision Tree Model.

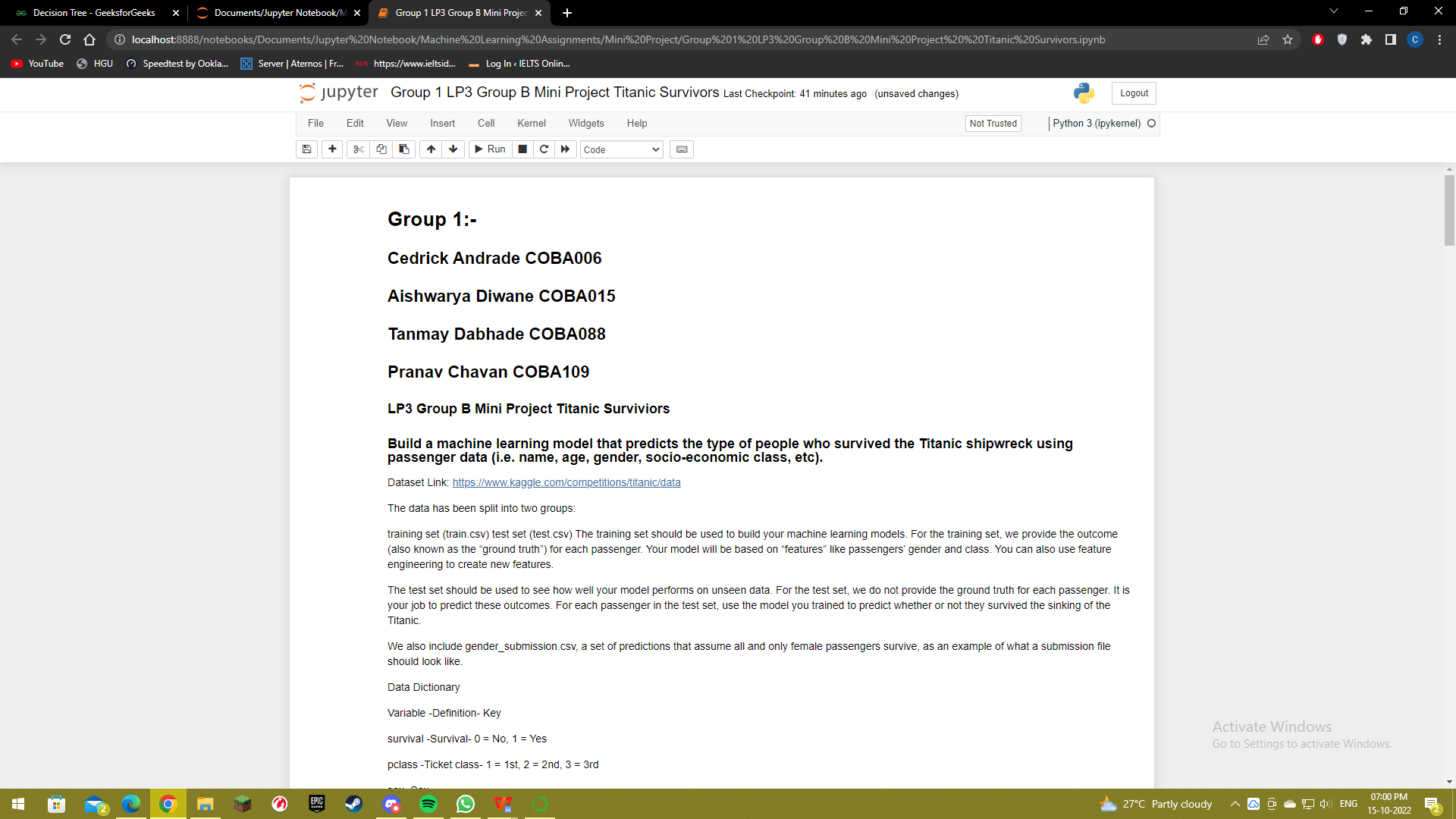
**The strengths of decision tree methods are:**

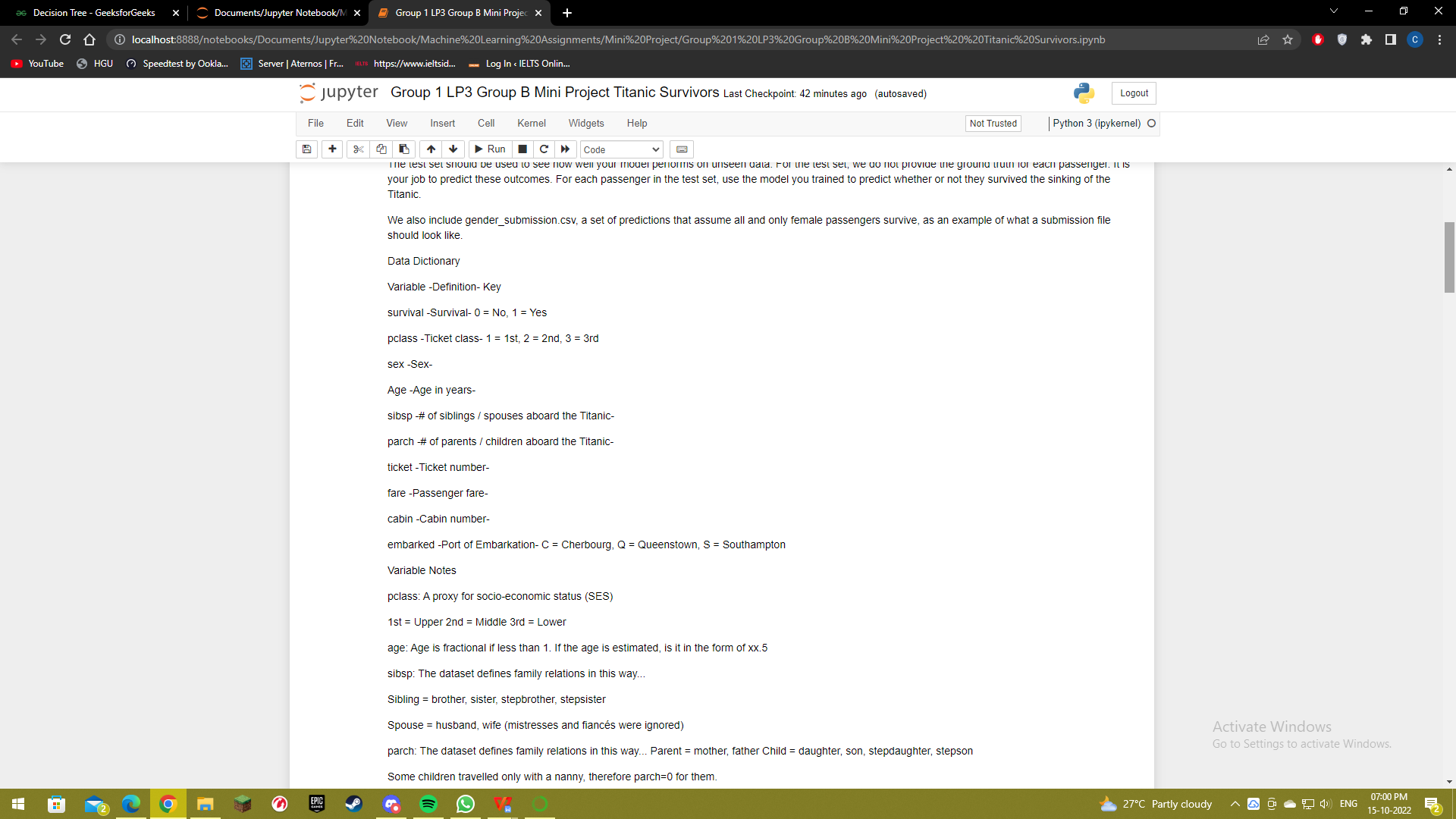
1. Decision trees are able to generate understandable rules.
2. Decision trees perform classification without requiring much computation.
3. Decision trees are able to handle both continuous and categorical variables.
4. Decision trees provide a clear indication of which fields are most important for prediction or classification.

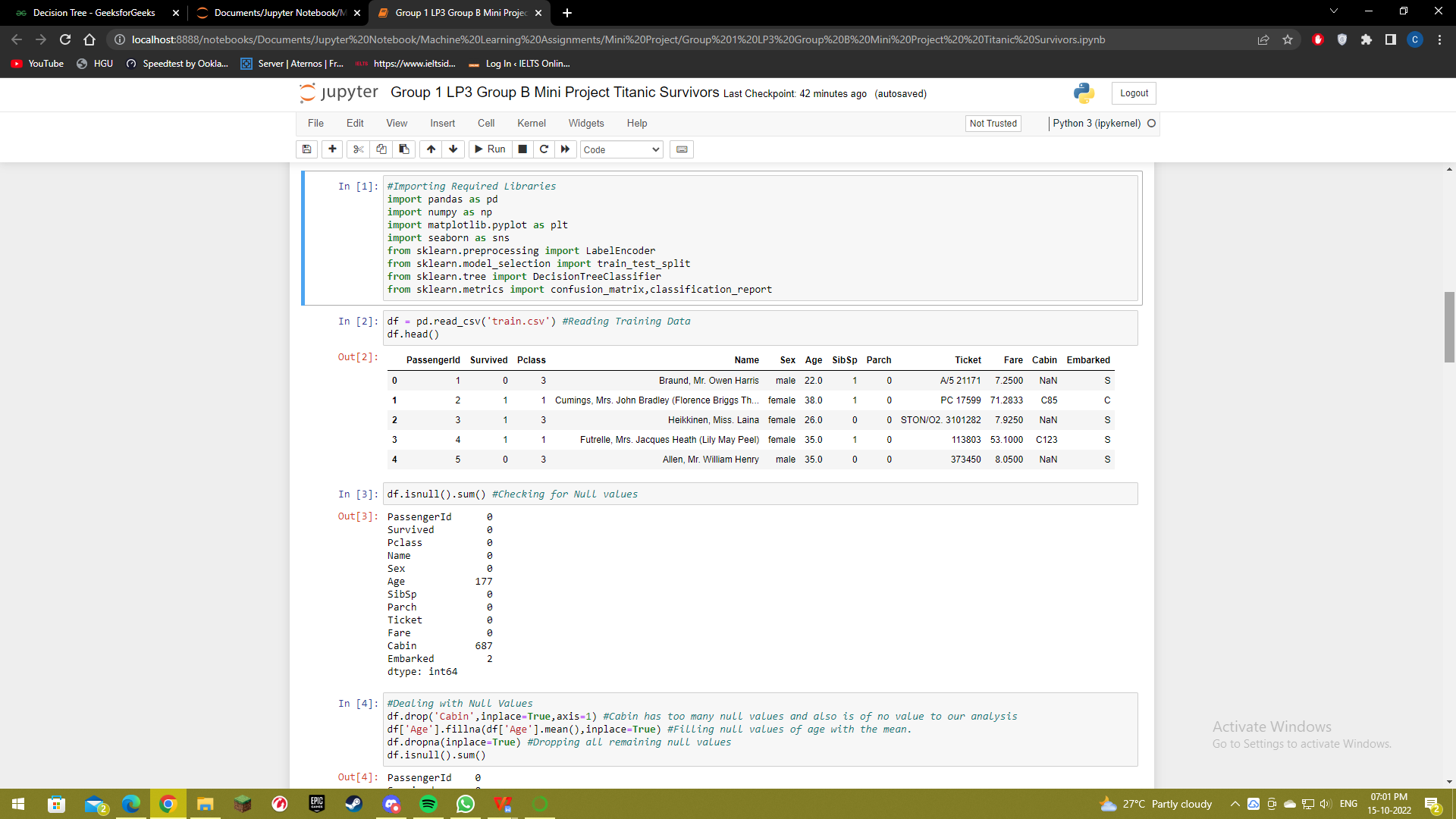
**The weaknesses of decision tree methods :**

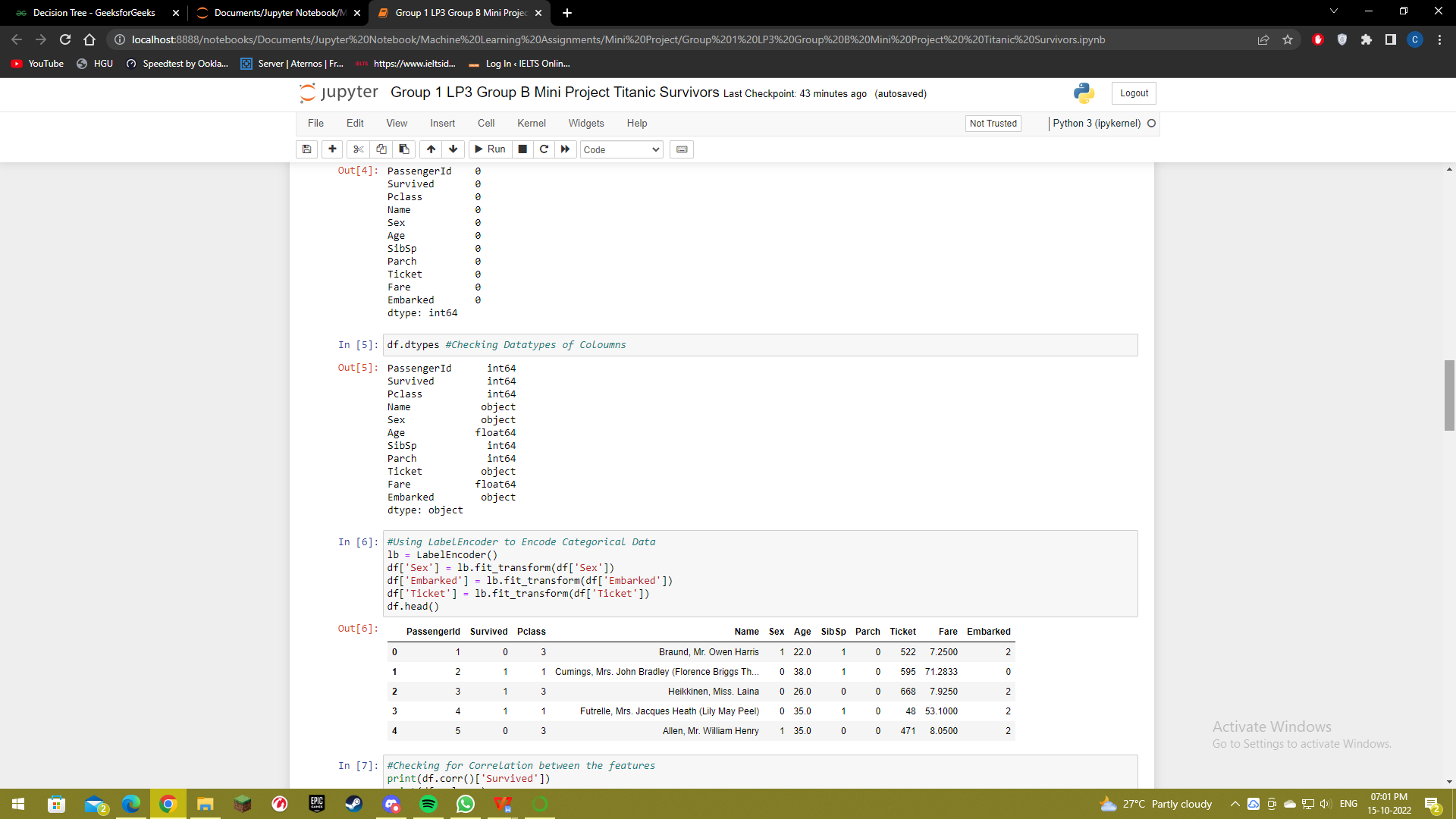
1. Decision trees are less appropriate for estimation tasks where the goal is to predict the value of a continuous attribute.
2. Decision trees are prone to errors in classification problems with many classes and a relatively small number of training examples.
3. Decision tree can be computationally expensive to train. The process of growing a decision tree is computationally expensive. At each node, each candidate splitting field must be sorted before its best split can be found. In some algorithms, combinations of fields are used and a search must be made for optimal combining weights. Pruning algorithms can also be expensive since many candidate sub-trees must be formed and compared.

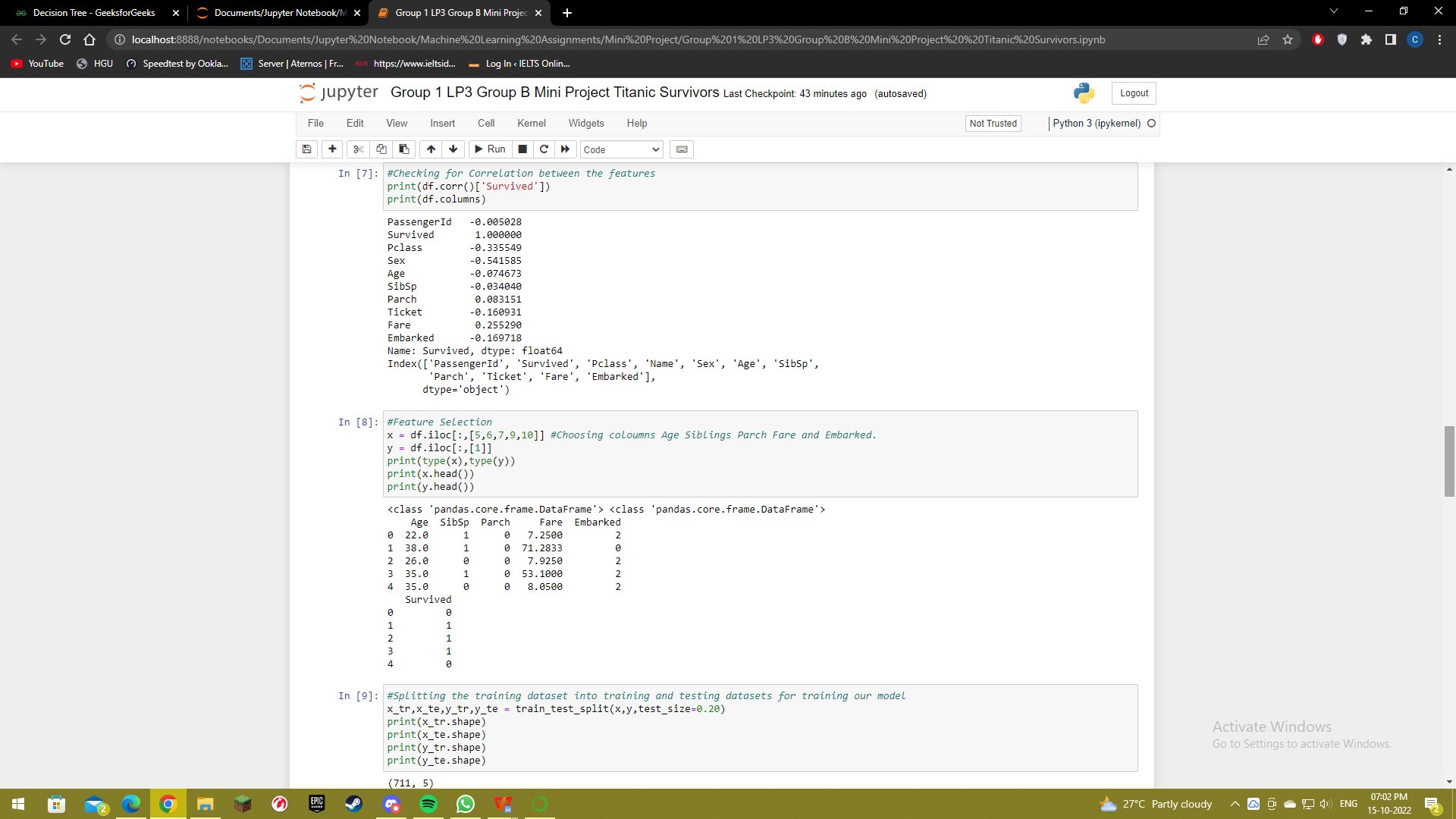
**Code and Output:-**

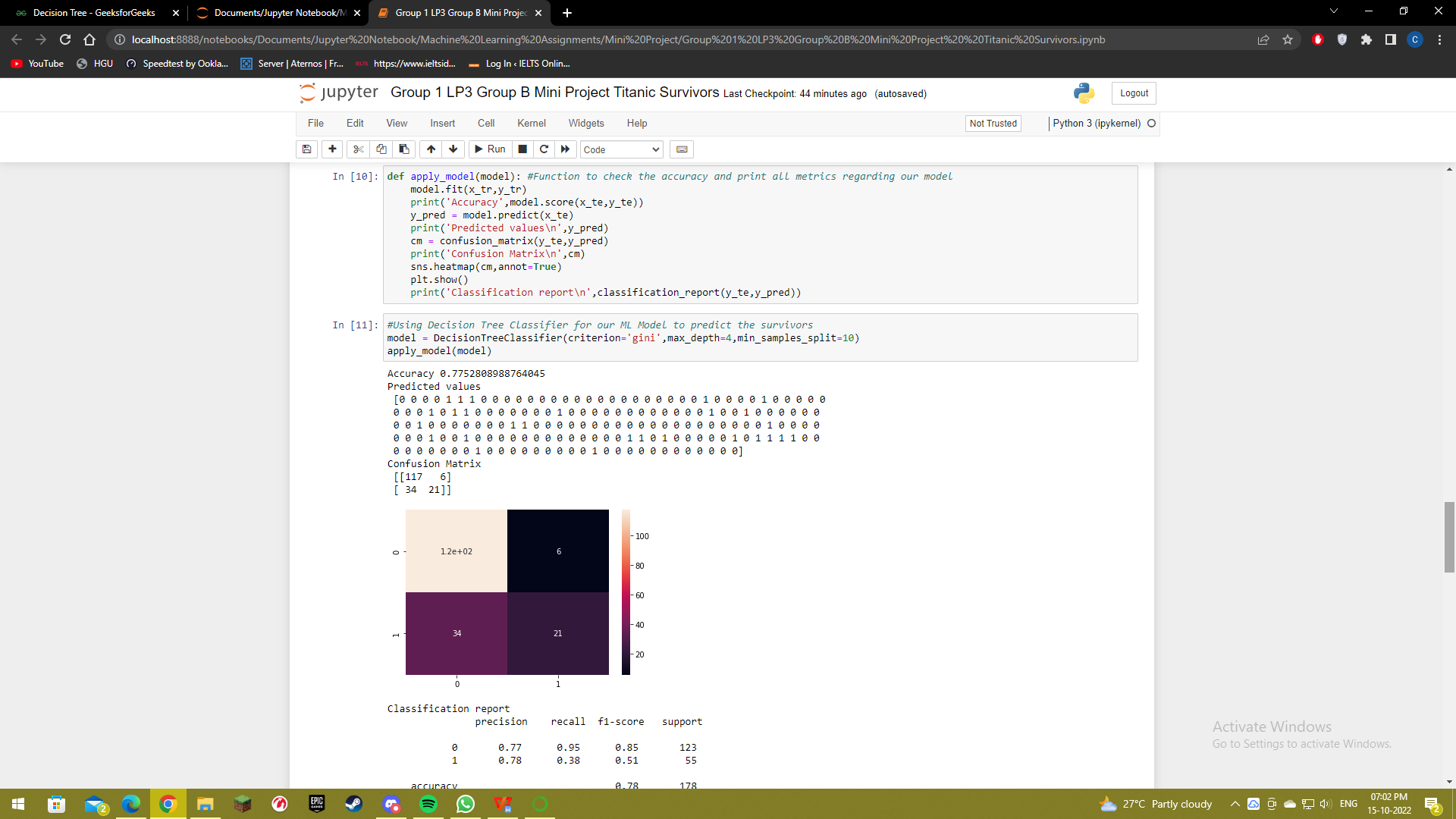


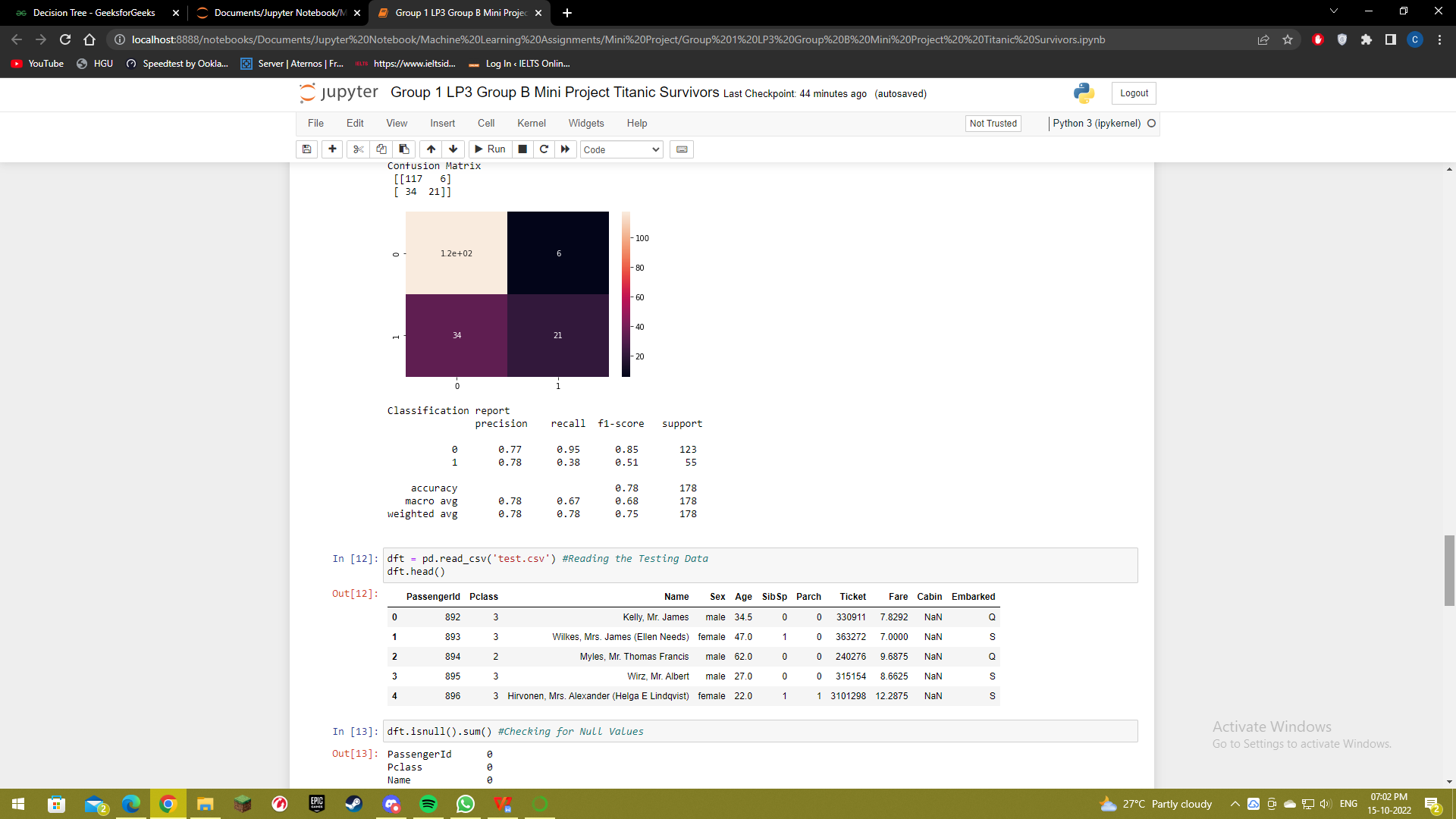


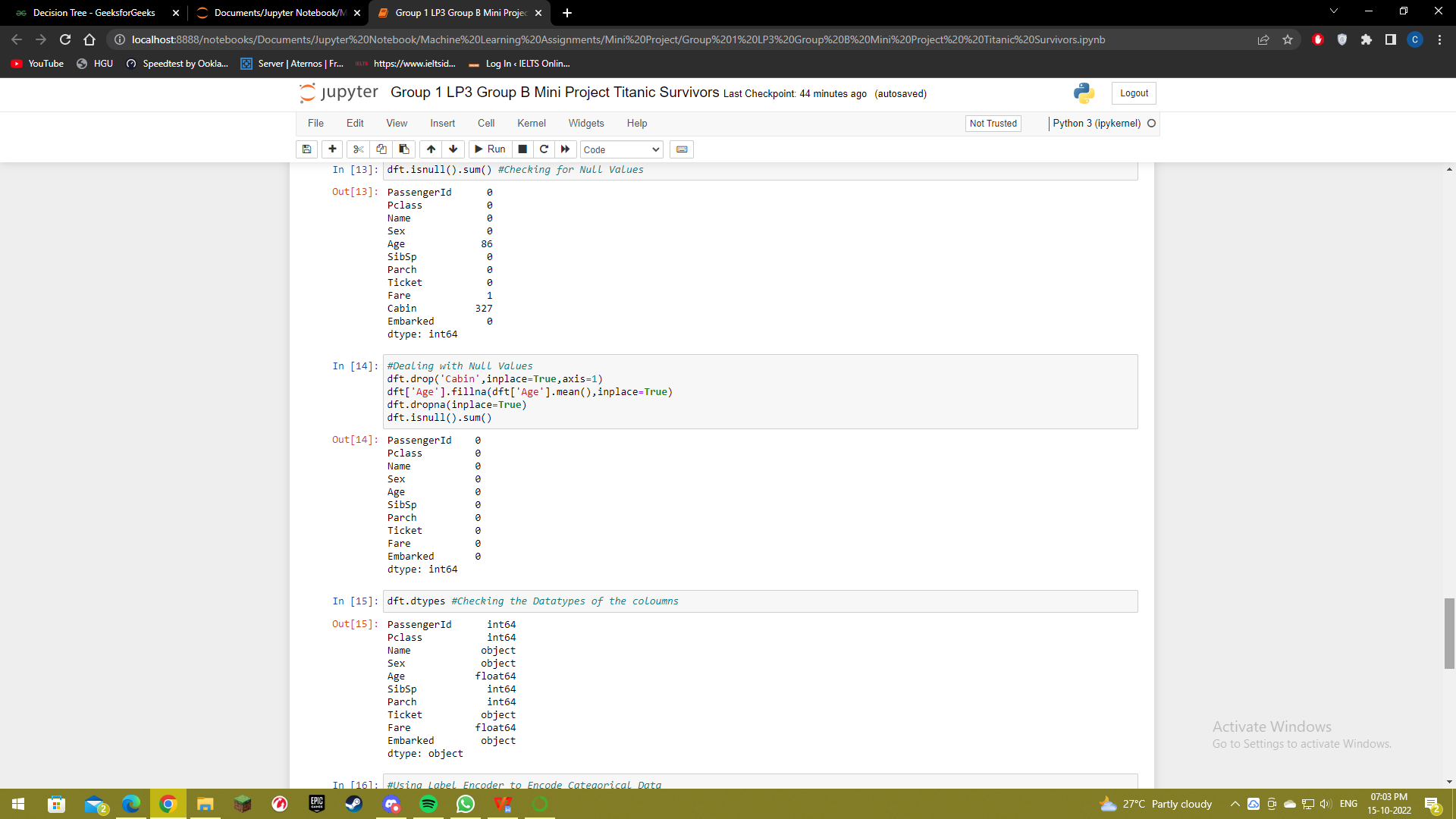


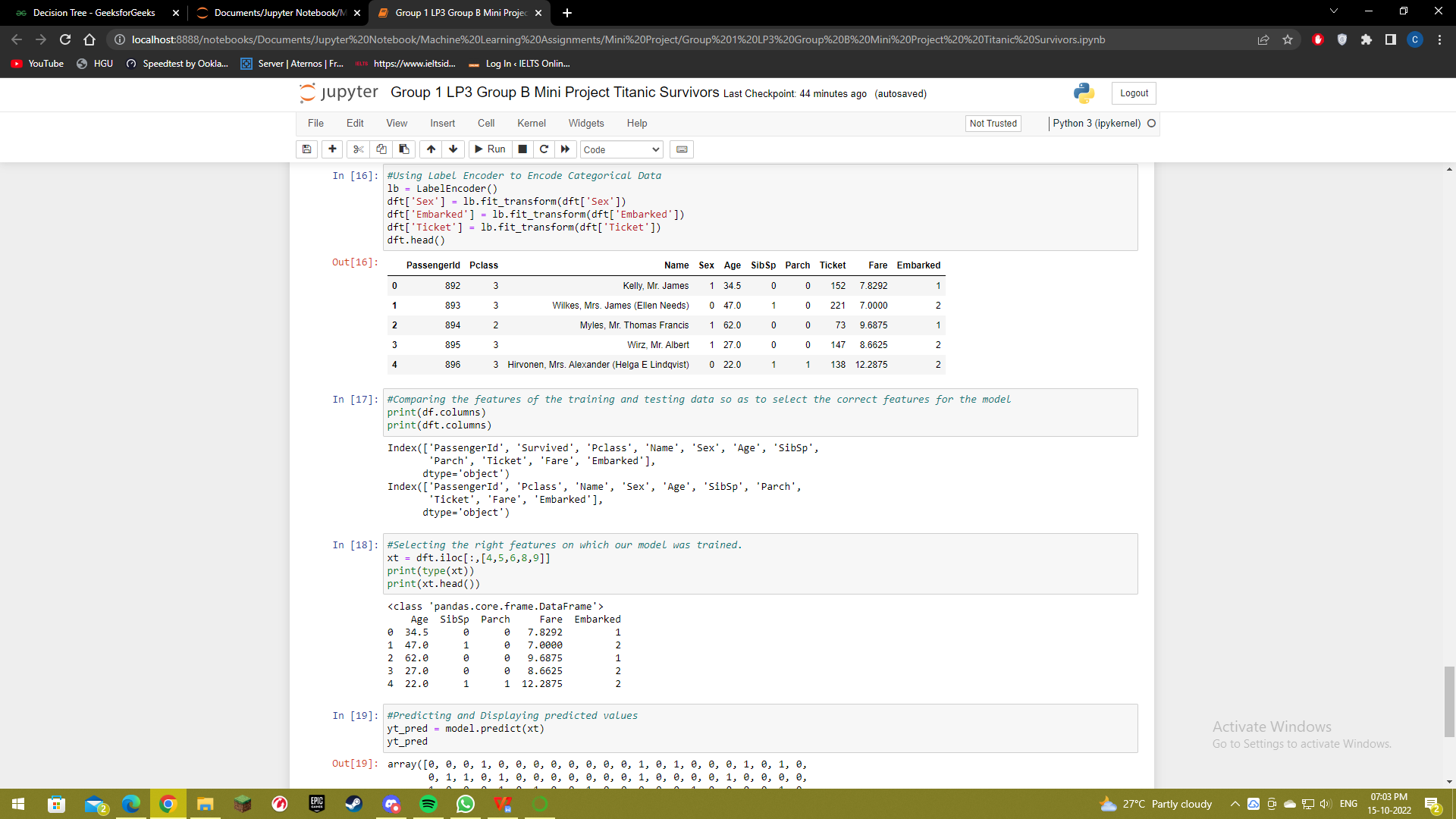


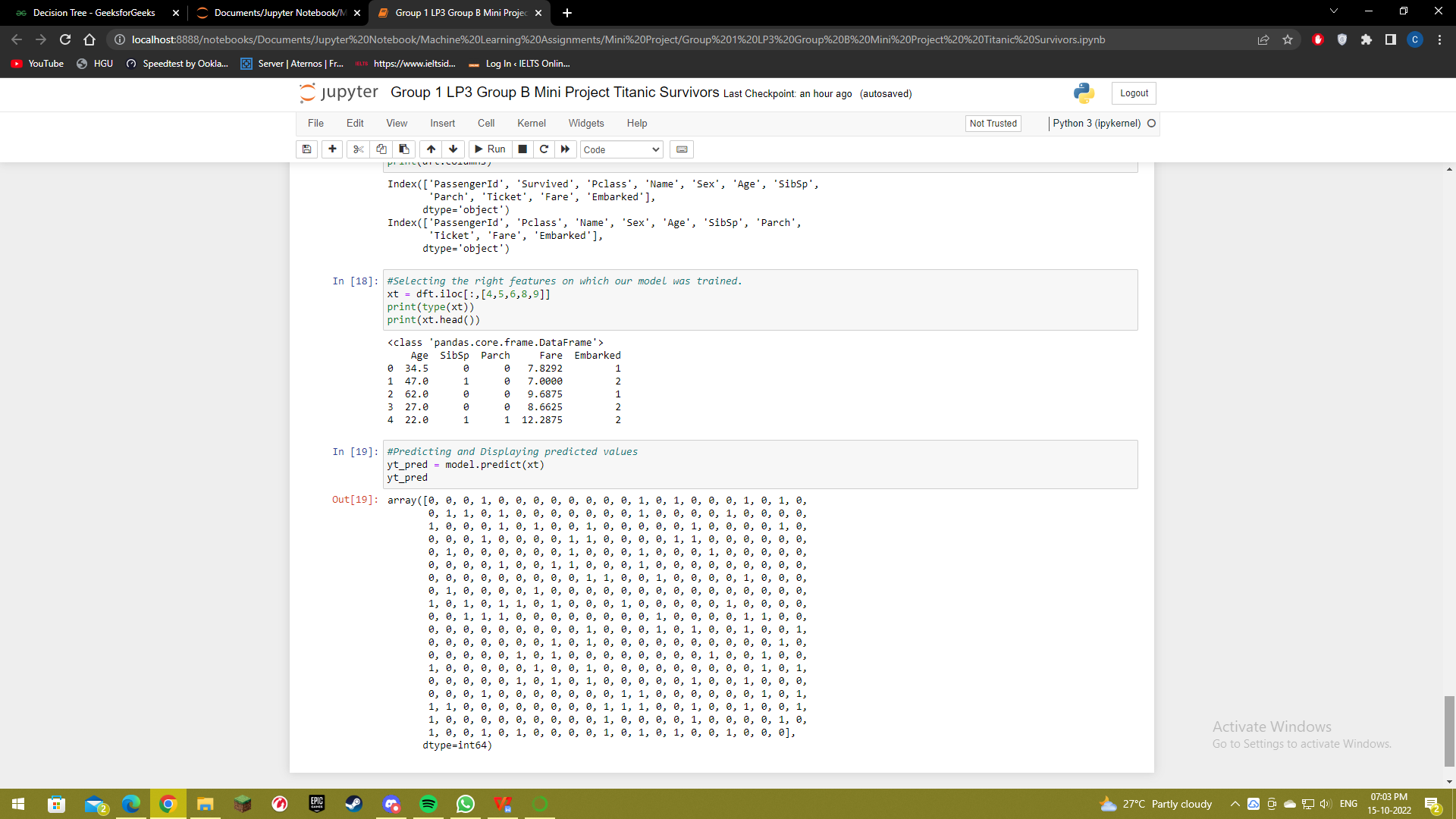












**Conclusion**

Thus we have implemented a Machine Learning model using Decision Tree Classification techniques to predict the survivors of the Titanic Dataset.

**References**

* [www.geekforgeeks.org](http://www.geekforgeeks.org)
* Techknowledge Machine Learning Textbook