**Survivorship in Titanic**

Report submitted in partial fulfillment of the requirement for the degree of

**Bachelor of Technology**

In

**Computer Science & Engineering**

By

**DEVESH GUPTA**

To

Maharaja Surajmal Insitute of Technology

Affiliated to Guru Gobind Singh Indraprastha University

Janakpuri, New Delhi-58

September 2018

**ABSTRACT**

This project is named as Survivorship in Titanic which received its name on the tragic historical case of sinking Titanic Ship in Pacific Ocean and for the counting of its survivor. This is completely a history but it does have the effect on our today’s life style. As the number of such accident is no more a story of our life, such cases has been increased and getting the prediction right of its survivor is as too difficult.

In such scenarios, getting the correct number of deaths and survivors are as important as to take the necessary preventive actions. Till that time, we are only left with the predictions to get the numbers right until the situations are under control.

To get the prediction right of the number of saved person, I have tried my best to utilize the power of coding to predict the same. There are certain factors which are taken into consideration for evaluation which are further defined in the report.

**CERTIFICATE**

This is to certify that the project report entitled “Survivorship in titanic” is a record of the bonafide work done by DEVESH GUPTA.

This report is submitted to MSIT as per requirement of the syllabus prescribed by GURU GOBIND SINGH INDRAPRASTH UNIVERSITY, NEW DELHI for the degree of Bachelor of Technology (Computer Science and Engineering), during the academic year 2018-19.

Mr. Sushil Kumar Ms. Koyal Dutta Gupta

(Project Coordinator) Head of Department

Computer Science and Engineering

Ms. Kavita Sheoran

(Project Coordinator)

**ACKNOWLEDGEMENT**

I am extremely thankful and grateful to MR. ASHUTOSH SINGH, my project guide. They being my guide have taken interest in the progress of my project work by providing facilities and guidance. I am indebted to my guide for his inspiration, support and kindness showered on me through out the project work.

I owe special gratitude to our **Head of Department of Computer Science and Engineering** for her constant support and guidance through out the course of studies for the project. Her sincerity, thoroughness and perseverance have been a constant source of inspiration for me.

I take this opportunity to thanks the teaching and non-teaching staff of **Computer Science and Engineering** of **Maharaja Surajmal Institute of Technology** for their encouragement and support in my project work.

I also would like to acknowledge the contribution of our friends who helped me I various aspects ranging from their suggestions in the compilation of the Report.

**DEVESH GUPTA**

**TABLE OF CONTENT**

//A lost the final file this is incomplete sorry

|  |  |  |  |
| --- | --- | --- | --- |
| S. No. | Chapter | Page No. |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

**INTRODUCTION**

**Project Introduction:**

This project is named as Survivorship in Titanic which received its name on the tragic historical case of sinking Titanic Ship in Pacific Ocean and for the counting of its survivor. This is completely a history but it does have the effect on our today’s life style. As the number of such accident is no more a story of our life, such cases has been increased and getting the prediction right of its survivor is as too difficult.

In such scenarios, getting the correct number of deaths and survivors are as important as to take the necessary preventive actions. Till that time, we are only left with the predictions to get the numbers right until the situations are under control.

To get the prediction right of the number of saved person, I have tried my best to utilize the power of coding to predict the same. There are certain factors which are taken into consideration for evaluation which are further defined in the report.

**Project Objective**

The main objective is to make machine learn the target feature by just analyzing the given data and to implement the machine learning in the prediction scenario without any human interference.

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves.

The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide.

The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly.

**Problem statement:**

To predict the number of survivors in any natural or unnatural disaster happened and its influence on the human life.

**Scope of work:**

“Survivorship in Titanic”- This project enhances the chance of a machine to inbuilt its own characteristics to analyse the data on the basis of previous experience and learn from them.

The scope of this project is not limited to the prediction but cover a wide scope:

1. Supervised Learning
2. Un-supervised Learning
3. Semi-supervised Reinforcement Learning
4. Supervised learning

The majority of practical machine learning uses supervised learning.

Supervised learning is where you have input variables (x) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output.

Y = f(X)

The goal is to approximate the mapping function so well that when we have new input data (x) that we can predict the output variables (Y) for that data. It is called supervised learning because the process of an algorithm learning from the training dataset can be thought of as a teacher supervising the learning process. We know the correct answers, the algorithm iteratively makes predictions on the training data and is corrected by the teacher. Learning stops when the algorithm achieves an acceptable level of performance. supervised learning problems can be further grouped into regression and classification problems.

1. Unsupervised Machine Learning

Unsupervised learning is where we only have input data (X) and no corresponding output variables. The goal for unsupervised learning is to model the underlying structure or distribution in the data in order to learn more about the data. These are called unsupervised learning because unlike supervised learning above there is no correct answers and there is no teacher. Algorithms are left to their own devises to discover and present the interesting structure in the data. Unsupervised learning problems can be further grouped into clustering and association problems.

1. Semi -supervised Reinforcement Learning

Problems where we have a large amount of input data (X) and only some of the data is labeled (Y) are called semi-supervised learning problems. These problems sit in between both supervised and unsupervised learning. A good example is a photo archive where only some of the images are labeled, (e.g. dog, cat, person) and the majority are unlabeled. Many real world machine learning problems fall into this area. This is because it can be expensive or time-consuming to label data as it may require access to domain experts. Whereas unlabeled data is cheap and easy to collect and store.

We can use unsupervised learning techniques to discover and learn the structure in the input variables.

We can also use supervised learning techniques to make best guess predictions for the unlabeled data, feed that data back into the supervised learning algorithm as training data and use the model to make predictions on new unseen data.

**MAIN TEXT**

**Software used**

The Jupyter Notebook

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

**Algorithm used**

Decision Tree Classifier

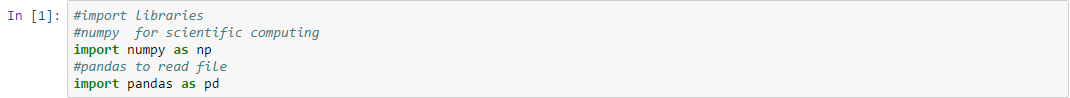
Decision tree learning uses a decision tree (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). It is one of the predictive modelling approaches used in statistics, data mining and machine learning. Tree models where the target variable can take a discrete set of values are called classification trees; in these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees.

In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. In data mining, a decision tree describes data (but the resulting classification tree can be an input for decision making). This page deals with decision trees in data mining.

Advantages of decision tree classifier over others

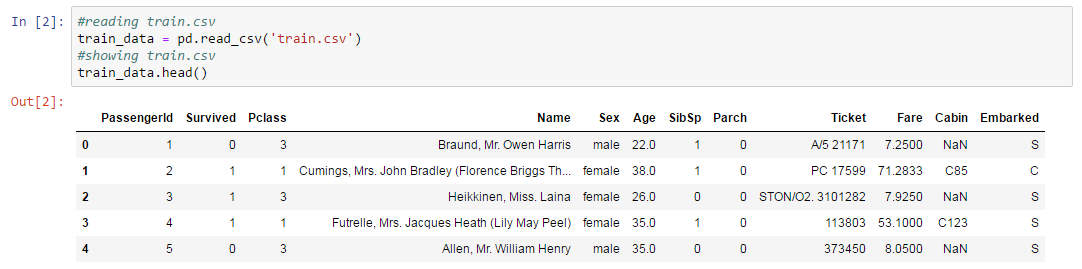
* Amongst other data mining methods, decision trees have various advantages:
* Simple to understand and interpret. People are able to understand decision tree models after a brief explanation. Trees can also be displayed graphically in a way that is easy for non-experts to interpret.[16]
* Able to handle both numerical and categorical data.[16] Other techniques are usually specialised in analysing datasets that have only one type of variable. (For example, relation rules can be used only with nominal variables while neural networks can be used only with numerical variables or categoricals converted to 0-1 values.)
* Requires little data preparation. Other techniques often require data normalization. Since trees can handle qualitative predictors, there is no need to create dummy variables.[16]
* Uses a white box model. If a given situation is observable in a model the explanation for the condition is easily explained by boolean logic. By contrast, in a black box model, the explanation for the results is typically difficult to understand, for example with an artificial neural network.
* Possible to validate a model using statistical tests. That makes it possible to account for the reliability of the model.
* Non-statistical approach that makes no assumptions of the training data or prediction residuals; e.g., no distributional, independence, or constant variance assumptions
* Performs well with large datasets. Large amounts of data can be analysed using standard computing resources in reasonable time.
* Mirrors human decision making more closely than other approaches.[16] This could be useful when modeling human decisions/behavior.
* Robust against co-linearity, particularly boosting
* In built feature selection. Additional irrelevant feature will be less used so that they can be removed on subsequent runs
* .Decision trees can approximate any Boolean function eq. XOR.[17]

**Code with results and conclusions**



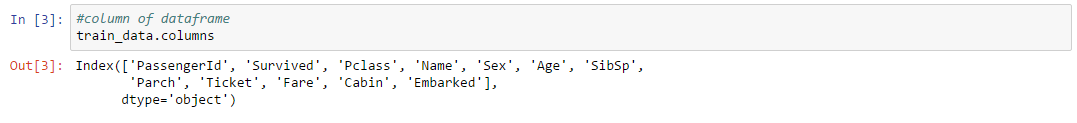
NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.

Pandas is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language.



(.csv) Extension: A comma-separated values (CSV) file is a delimited text file that uses a comma to separate values. A CSV file stores tabular data (numbers and text) in plain text. Each line of the file is a data record. Each record consists of one or more fields, separated by commas. The use of the comma as a field separator is the source of the name for this file format.

Dataframe.head(): Return the first n rows. This function returns the first n rows for the object based on position. It is useful for quickly testing if your object has the right type of data in it.



Dataframe.columns: Tabular data structure with labeled axes (rows and columns). Arithmetic operations align on both row and column labels. Can be thought of as a dict-like container for Series objects. The primary pandas data structure

Variable Description:

Survival: Survival

(0 = No; 1 = Yes)

Pclass: Passenger Class

(1 = 1st; 2 = 2nd; 3 = 3rd)

Pclass is a proxy for socio-economic status (SES)

1st ~ Upper; 2nd ~ Middle; 3rd ~ Lower

Name: Name

Sex: Sex

Age: Age

Age is in Years

SibSp: Number of Siblings/Spouses Aboard

Sibling: Brother, Sister, Stepbrother, or Stepsister of Passenger Aboard

Spouse: Husband or Wife of Passenger Aboard Titanic

Parch: Number of Parents/Children Aboard

Parent: Mother or Father of Passenger Aboard Titanic

Child: Son, Daughter, Stepson, or Stepdaughter of Passenger Aboard Titanic

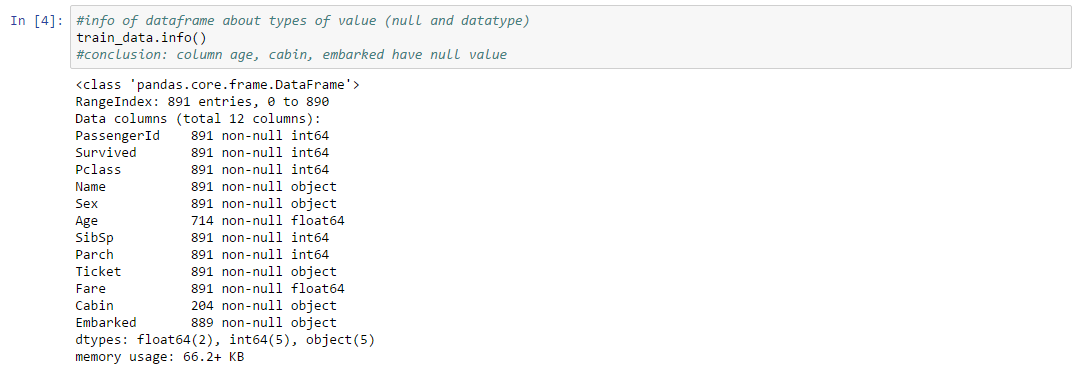
Ticket: Ticket Number

Fare: Passenger Fare

Cabin: Cabin Number

Embarked: Port of Embarkation

(C = Cherbourg; Q = Queenstown; S = Southampton)



info() is function used to get information about object like Ranged index, number of column, column name with no. of non null values and datatype of that column, number of column with certain datatype and memory usage.

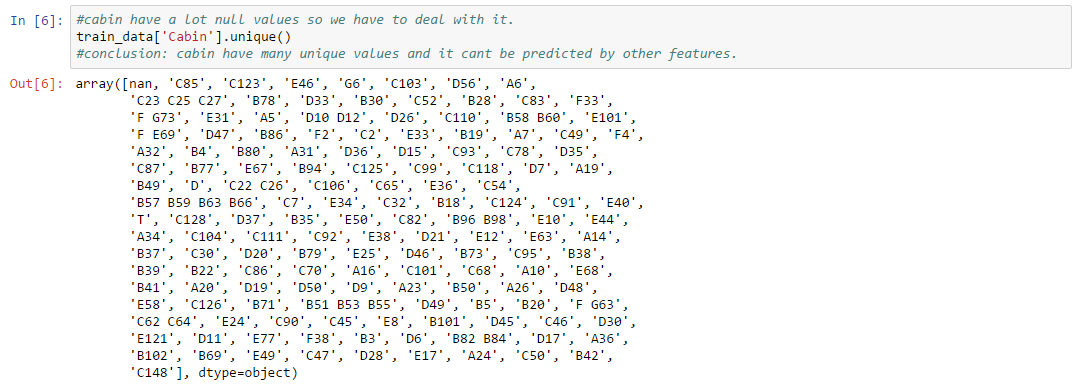
* Conclusion: Column age, cabin, embarked column contains null value.



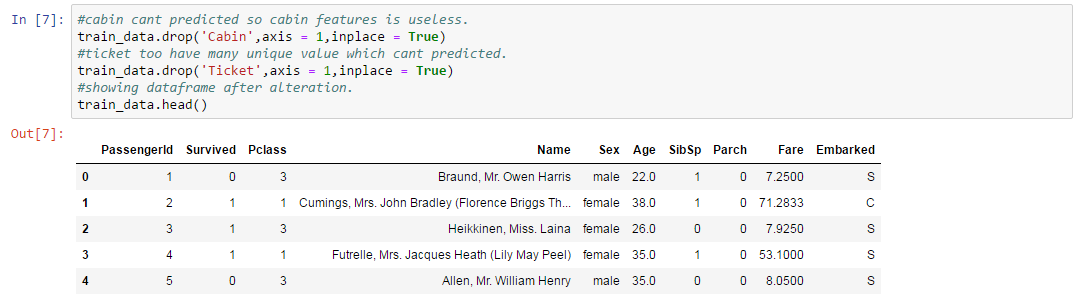
describe() is function used for numeric data, the results index will include count, mean,std, min, max as well as lower, so and upper percentiles. By default the lower percentile is 25 and the upper percentile is 75. The 50 percentile is the same as the median.

* Conclusion:

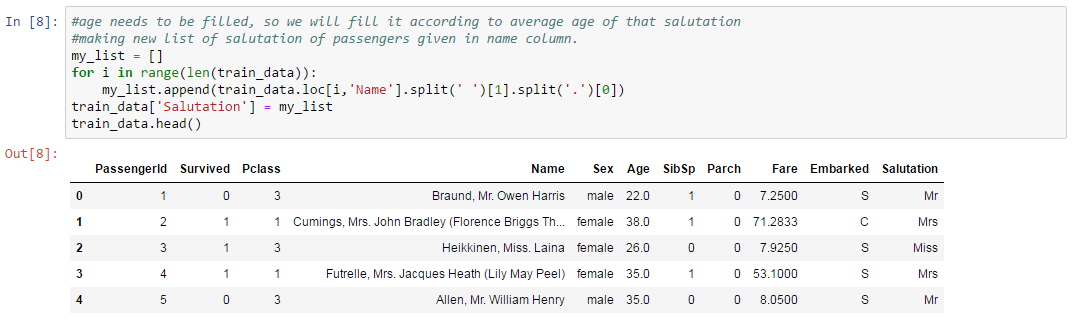
1. Passenger, age, fare contains many unique values. and survived, pclass, sibsp contains few unique values.
2. Remaining column contains non numeric values.



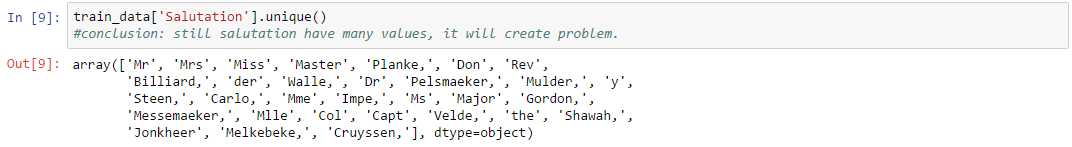
* Conclusion: Passenger have different unique cabin or no cabin. Null cabin can,t filled by using existing cabin because there is no pattern in cabin column. We can drop this column.



Dataframe.drop(): Drop specified labels from rows or columns. Remove rows or columns by specifying label names and corresponding axis, or by specifying directly index or column names. When using a multi-index, labels on different levels can be removed by specifying the level.

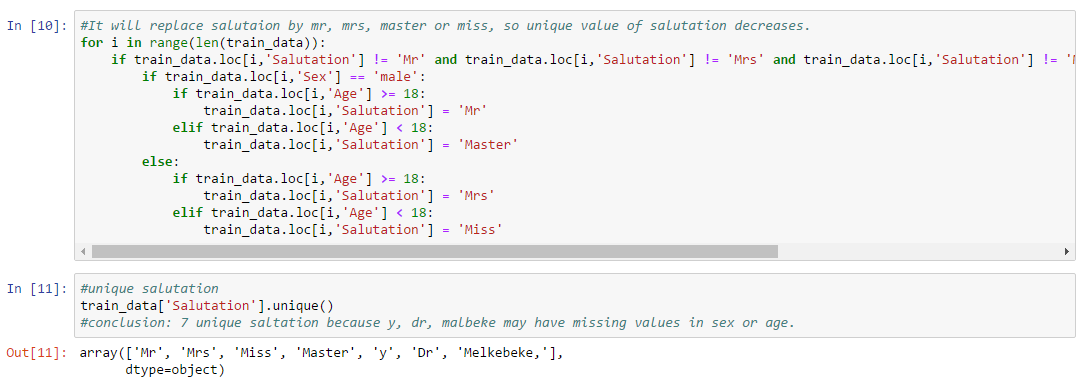


Dataframe.append(): The append() method appends an element to the end of the list.



* Conclusion: Still Salutation column contains null values.

Dataframe.unique(): Return unique values of Series object. Uniques are returned in order of appearance. Hash table-based unique, therefore does NOT sort.



* Conclusion: 7 unique values in Salutation column because Sex, Age column missing value

FLOW CHART IN [10]



Start

End

No

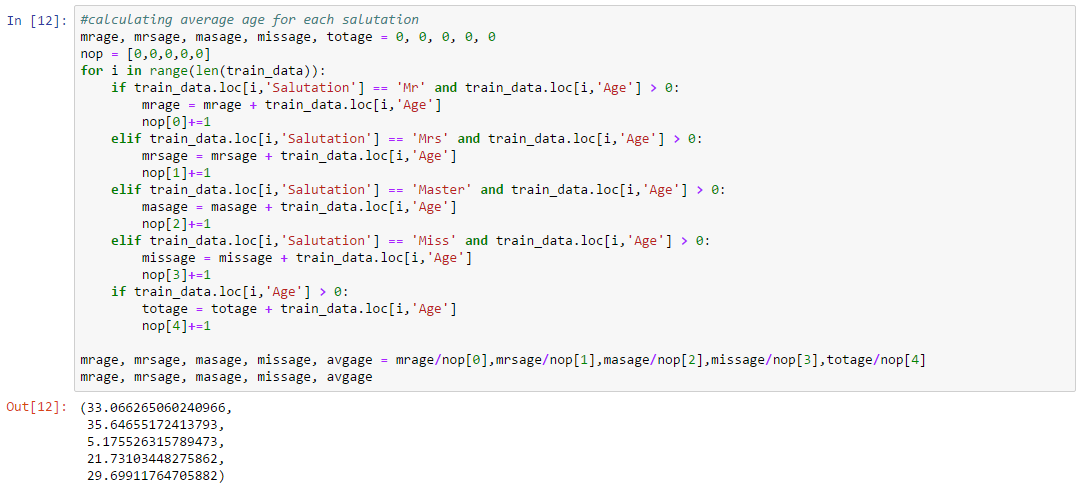
Yes

Yes

No

Yes No

Yes No Yes No





FLOW CHART IN[12]



Start

End

No

Yes

Yes

No

Yes

No

Yes

No

Yes

No

Yes

No

FLOW CHART IN[13]



Start

End

No

Yes

Yes

No

No

Yes Yes

No

No

Yes

Yes

No No

Yes

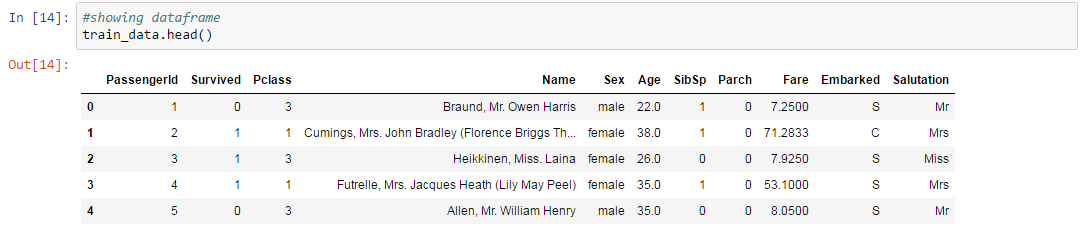
Yes

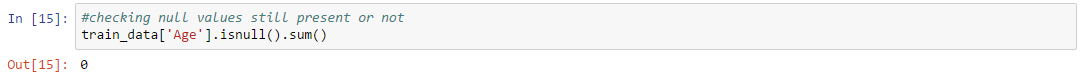
No

Yes No

Yes

No





Dataframe.isnull(): Detect missing values. Return a boolean same-sized object indicating if the values are NA. NA values, such as None or numpy.NaN, gets mapped to True values. Everything else gets mapped to False values. Characters such as empty strings '' or numpy.inf are not considered NA values.

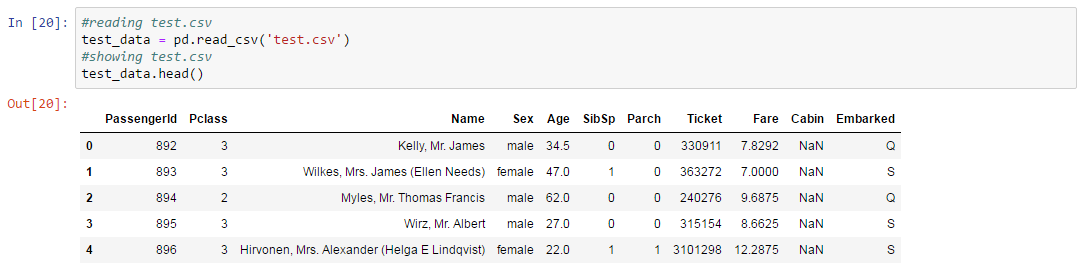
Dataframe.sum():Return the sum of the values for the requested axis

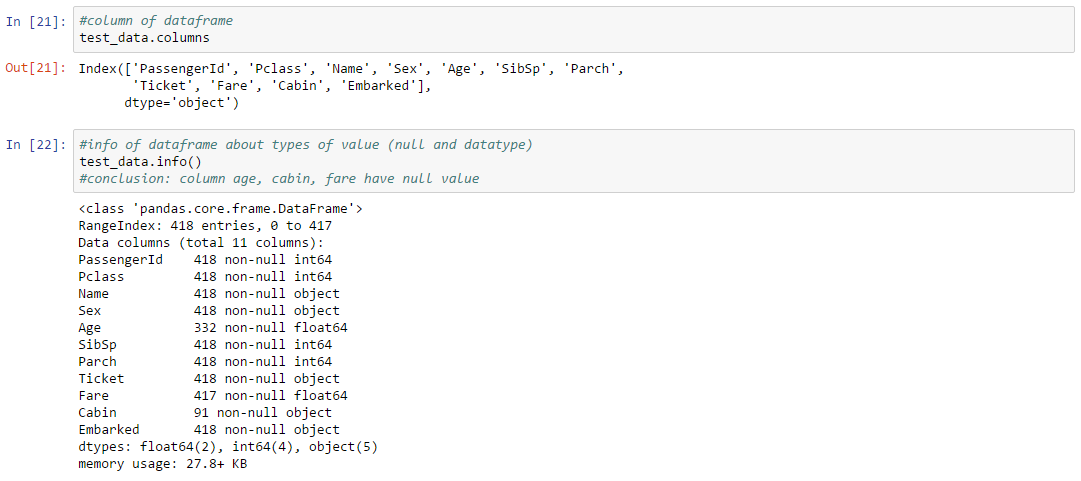


* Conclusion: ‘S’ is most common value in Embarked column.

Dataframe.count(): Count non-NA cells for each column or row. The values None, NaN, NaT, and optionally numpy.inf are considered NA.





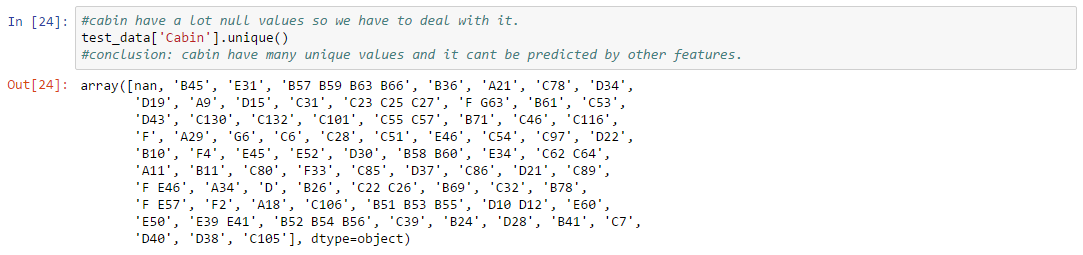


* Conclusion: Age, Cabin column may have null values.

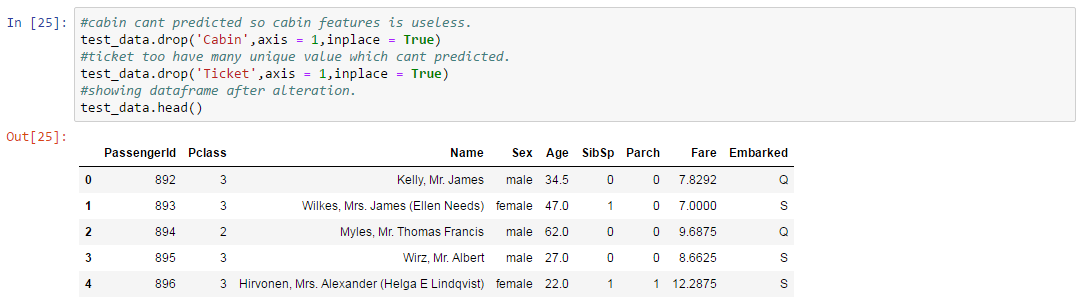


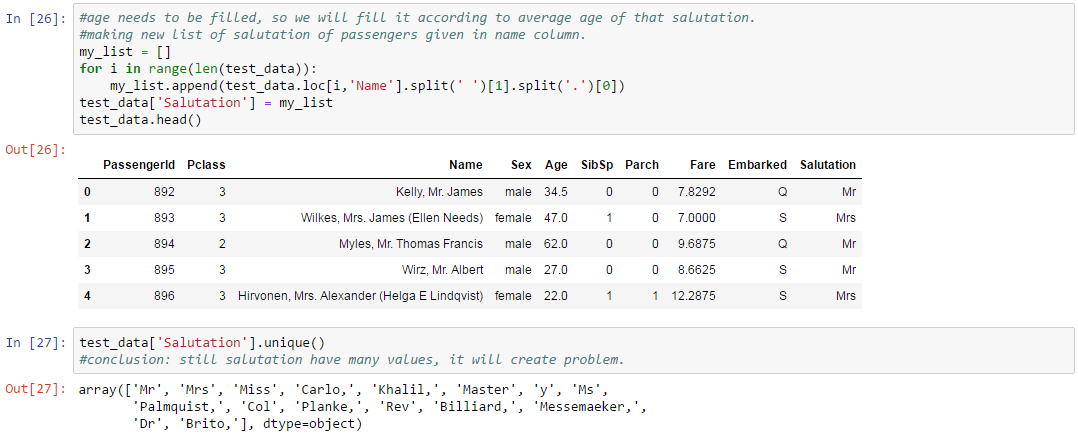
* Conclusion:

1. Passenger, Age, Fare contains many unique values and Survived, Pclass have few unique values.
2. Reamining column may contains non numeric.

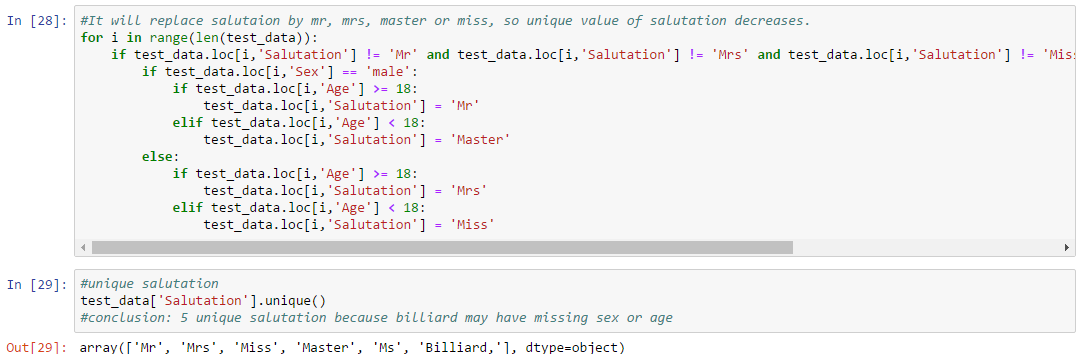


* Conclusion: Cabin contains unique values that can’t predicted.





* Conclusion: Salutation still contains large sets of values.



* Conclusion: 5 unique values in salutation column because billiard may have null.

FLOW CHART IN[28]

No

Yes

Yes

No

Yes No

Yes No Yes No



Start

End





FLOW CHART IN[30]

No

Yes

Yes 

Start

End

No

Yes

No

Yes

No

Yes

No

Yes

No

FLOW CHART IN[31]



Start

End

No

Yes

Yes

No

No

Yes Yes

No No

Yes Yes

No No

Yes

Yes

No

No

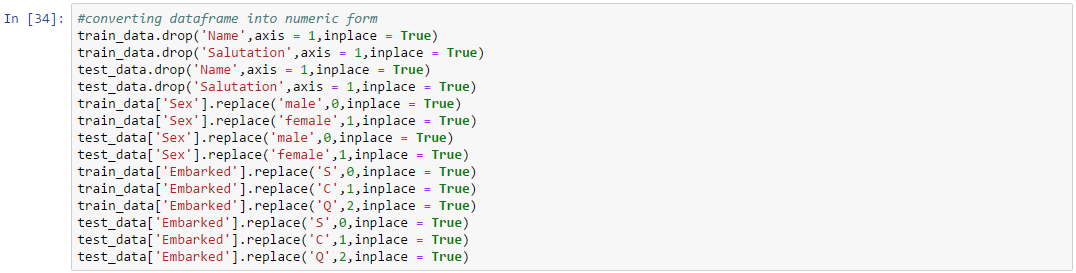
Yes

No

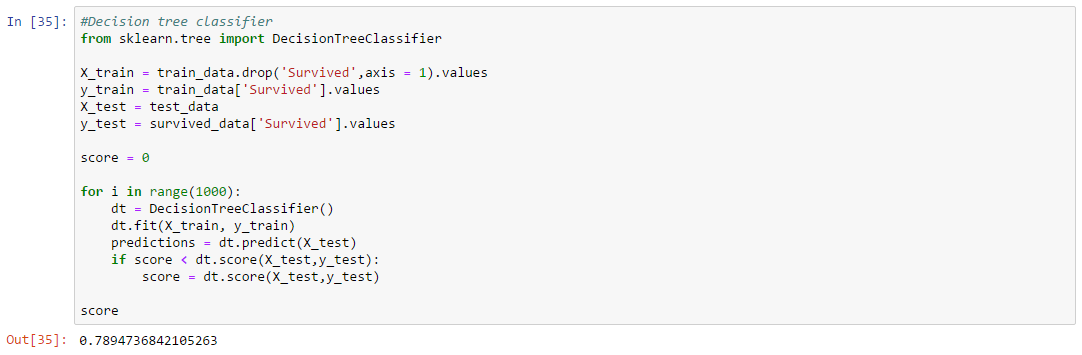
Yes

C:\Users\Devesh\Desktop\Snipings\in 32.PNG





Dataframe.replace(): Replace values given in to replace with value.Values of the DataFrame are replaced with other values dynamically.



Score is best score in 1000 test

Score means how many prediction given by our model is correct in part of 1.

**BIBLIOGRAPHY**

Kaggle, ’Titanic: Machine learning from disaster’ retrieved from <http://www.kaggle.com/c/titanic/data>

James Le, ‘Top algorithm for machine learning’ (2016 Jan 20), retrived from

[https://towardsdatac=science.com/a-tour-of-the-top-10-algorithms](about:blank)

Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. An Introduction to Statical Learning. Stanford: Springer, 2018. E-book

Shai Shalev-Shwartz, Shai Ben-David. Understanding Machine Learning From Theory to Algorithms. Cambridge, Cambridge University Press, 2014. E-Book

Wes Mckinney, ’Pandas’ (2018), retrived from

<https://pandas.pydata.org>

Scikit-developers, “Decision Tree” (2007), retrieved from

<https://scikit-learn.org/stable/modules/tree.html>

W3school developers, ‘Python lis append() method’ (2018), retrived from

<https://www.w3schools.com/python/ref_list_append.asp>

**APPENDICES**

**Source of the Project:**

From being the safest ship to the tragic story of sinking, Titanic gives us all the reason to explore more. With the number of people in hands and with limited choice whom to save, prediction plays an autonomous role in every phase. So, taking a great example from the history, let’s talk about few more technicality and the situation of that era which beautify the project details.

RMS Titanic was a British passenger liner that sank in the North Atlantic Ocean on 15 April 1912 after colliding with an iceberg during her maiden voyage from Southampton, England to New York City. The sinking of Titanic caused the deaths of 1,514 people in one of the deadliest peacetime maritime disasters in history. She was the largest ship afloat at the time of her maiden voyage. One of three Olympic class ocean liners operated by the White Star Line, she was built between 1909–11 by the Harland and Wolff shipyard in Belfast. She carried 2,224 people.

Her passengers included some of the wealthiest people in the world, as well as over a thousand emigrants from Britain, Ireland, Scandinavia and elsewhere seeking a new life in North America. The ship was designed to be the last word in comfort and luxury, with an on-board gymnasium, swimming pool, libraries, high-class restaurants and opulent cabins. She also had a powerful wireless telegraph provided for the convenience of passengers as well as for operational use. Though she had advanced safety features such as watertight compartments and remotely activated watertight doors, she lacked enough lifeboats to accommodate all of those aboard. Due to outdated maritime safety regulations, she carried only enough lifeboats for 1,178 people – a third of her total passenger and crew capacity.

After leaving Southampton on 10 April 1912, Titanic called at Cherbourg in France and Queenstown (now Cobh) in Ireland before heading westwards towards New York. On 14 April 1912, four days into the crossing and about 375 miles south of Newfoundland, she hit an iceberg at 11:40 pm (ship's time; UTC-3). The glancing collision caused Titanic's hull plates to buckle inwards in a number of locations on her starboard side and opened five of her sixteen watertight compartments to the sea. Over the next two and a half hours, the ship gradually filled with water and sank. Passengers and some crew members were evacuated in lifeboats, many of which were launched only partly filled. A disproportionate number of men – over 90% of those in Second Class – were left aboard due to a "women and children first" protocol followed by the officers loading the lifeboats. Just before 2:20 am Titanic broke up and sank bow-first with over a thousand people still on board. Those in the water died within minutes from hypothermia caused by immersion in the freezing ocean. The 710 survivors were taken aboard from the lifeboats by the RMS Carpathia a few hours later.

The disaster was greeted with worldwide shock and outrage at the huge loss of life and the regulatory and operational failures that had led to it. Public inquiries in Britain and the United States led to major improvements in maritime safety. One of their most important legacies was the establishment in 1914 of the International Convention for the Safety of Life at Sea (SOLAS), which still governs maritime safety today. Many of the survivors lost all of their money and possessions and were left destitute; many families, particularly those of crew members from Southampton, lost their primary bread-winners. They were helped by an outpouring of public sympathy and charitable donations. Some of the male survivors, notably the White Star Line's chairman, J. Bruce Ismay, were accused of cowardice for leaving the ship while people were still on board, and they faced social ostracism.

The wreck of the Titanic remains on the seabed, gradually disintegrating at a depth of 12,415 feet (3,784 m). Since its rediscovery in 1985, thousands of artefacts have been recovered from the sea bed and put on display at museums around the world. Titanic has become one of the most famous ships in history, her memory kept alive by numerous books, films, exhibits and memorials.