

**A PROJECT REPORT**

**ON**

**HEART DISEASE CLASSIFICATION USING**

**DECISION TREE CLASSIFIER ALGORITHIM**

Submitted in partial fulfillment for the requirement of the award of

TRAINING

IN

Data Analytics, Machine Learning and AI using Python



*Submitted By*

**Ayan Roychoudhury (Meghnad Saha Institute of Technology,Kolkata)**

*Under the guidance of*

**Mr. Bipul Shahi**



I

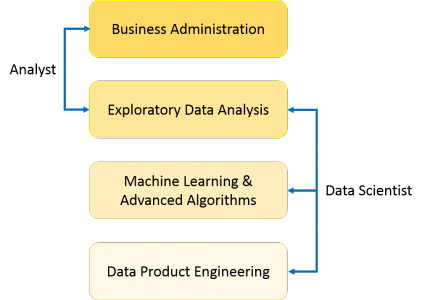
**ACKNOWLEDGEMENT**

My sincere gratitude and thanks towards my project paper guide Bipul Shahi, Corporate Trainer, Developer . IOT, Artificial Intelligence, Robotics, Cloud Computing, Android Apps.

It was only with his backing and support that I could complete the report. He provided me all sorts of help and corrected me if ever seemed to make mistakes. I have no such words to express my gratitude. I would also like to thank my College professors Dr.Swapnadip De (H.O.D. of ECE Department) and Dr. C Pani Madam to inspire me in such activities and guide me whenever I needed. I acknowledge my sincere gratitude to the lecturers, research scholars and the lab technicians for their valuable guidance and helping attitude even in their very busy schedule. And at last but not the least, I acknowledge my dearest parents for being such a nice source of encouragement and moral support that helped me tremendously in this aspect. I also declare to the best of my knowledge and belief that the Project Work has not been submitted anywhere else.

**INTRODUCTION**

Machine learning is the science of getting computers to act without being explicitly programmed. In the past decade, machine learning has given us self-driving cars, practical speech recognition, effective web search, and a vastly improved understanding of the human genome. As the world entered the era of big data, the need for its storage also grew. It was the main challenge and concern for the enterprise industries until 2010. The main focus was on building framework and solutions to store data.



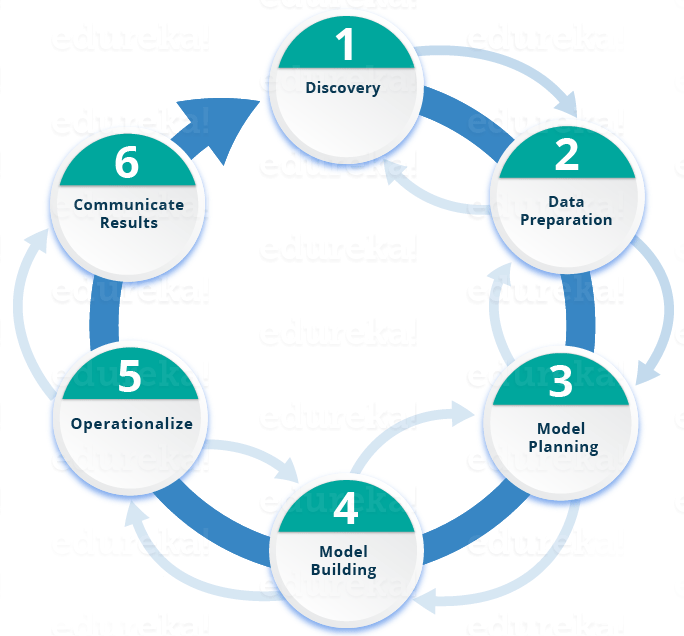
A Data Analyst usually explains what is going on by processing history of the data. On the other hand, Data Scientist not only does the exploratory analysis to discover insights from it, but also uses various advanced machine learning algorithms to identify the occurrence of a particular event in the future.

So, Data Science is primarily used to make decisions and predictions making use of predictive causal analytics, prescriptive analytics (predictive plus decision science) and machine learning.

* Predictive causal analytics
* Prescriptive analytics
* Machine learning for making predictions
* Machine learning for pattern discovery

## Lifecycle of Data Science

Here is a brief overview of the main phases of the Data Science Lifecycle:



## Technology and Concepts

## 1.Supervised Machine 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 you have new input data (x) that you 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.

* **Classification**: A classification problem is when the output variable is a category, such as “red” or “blue” or “disease” and “no disease”.
* **Regression**: A regression problem is when the output variable is a real value, such as “dollars” or “weight”.

Some common types of problems built on top of classification and regression include recommendation and time series prediction respectively.

Some popular examples of supervised machine learning algorithms are:

* Linear regression for regression problems.
* Random forest for classification and regression problems.
* Support vector machines for classification problems.

**2.Unsupervised Machine Learning**

Unsupervised learning is where you 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.

* **Clustering**: A clustering problem is where you want to discover the inherent groupings in the data, such as grouping customers by purchasing behavior.
* **Association**:  An association rule learning problem is where you want to discover rules that describe large portions of your data, such as people that buy X also tend to buy Y.

Some popular examples of unsupervised learning algorithms are:

* k-means for clustering problems.
* Apriori algorithm for association rule learning problems.

## ****Classification Definition****

We use the training dataset to get better boundary conditions which could be used to determine each target class. Once the boundary conditions are determined, the next task is to predict the target class. The whole process is known as classification.

**Basic Terminology in Classification Algorithms**

* **Classifier:** An algorithm that maps the input data to a specific category.
* **Classification model:**A classification model tries to draw some conclusion from the input values given for training. It will predict the class labels/categories for the new data.
* **Feature:** A feature is an individual measurable property of a phenomenon being observed.
* **Binary Classification:** Classification task with two possible outcomes. **Eg: Gender classification (Male / Female)**
* **Multi-class classification:** Classification with more than two classes. In multi-class classification, each sample is assigned to one and only one target label. **Eg: An animal can be a cat or dog but not both at the same time.**
* **Multi-label classification:**Classification task where each sample is mapped to a set of target labels (more than one class). **Eg: A news article can be about sports, a person, and location at the same time.**

**Applications of Classification Algorithms**

* Email spam classification
* Bank customers loan pay willingness prediction.
* Cancer tumor cells identification.
* Sentiment analysis
* Drugs classification
* Facial key points detection
* Pedestrians detection in an automotive car driving.

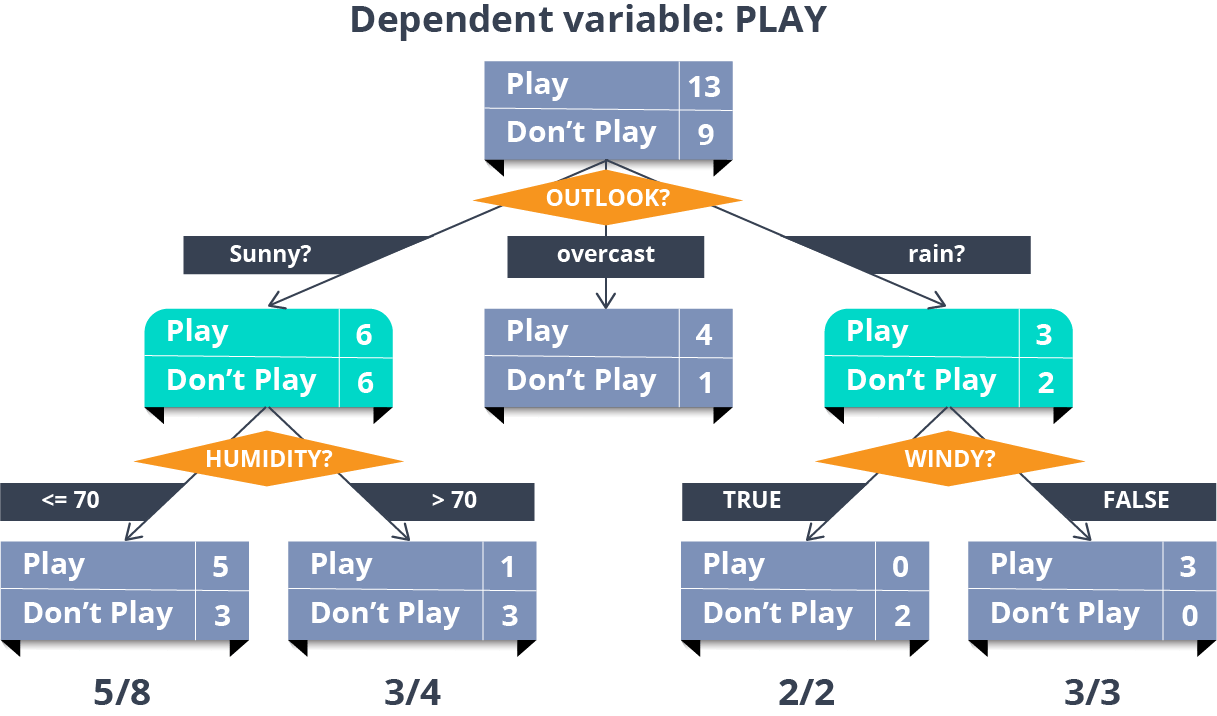
**Types of Classification Algorithms**

Classification Algorithms could be broadly classified as the following:

* ***Linear Classifiers***
  + Logistic regression
  + Naive Bayes classifier
  + Fisher’s linear discriminant
* ***Support vector machines***
  + Least squares support vector machines
* ***Quadratic classifiers***
* ***Kernel estimation***
  + k-nearest neighbor
* ***Decision trees***
  + Random forests
* ***Neural networks***
* ***Learning vector quantization***

## ****Decision Trees****

The Decision Tree is by far, my most favorite algorithm. With versatile features helping actualize both categorical and continuous dependent variables, it is a type of supervised learning algorithm mostly used for classification problems. What this algorithm does is, it splits the population into two or more homogeneous sets based on the most significant attributes making the groups as distinct as possible.



# The Two Type of Decision Trees

**Classification Trees:** When the decision tree has categorical target variable. The above tree is an example of a classification tree because we know that there are two options for the result.

**Regression Trees: When the decision tree has a continuous target variable. For example, a regression tree would be used for the price of a newly launched product because price can be anything depending on various constraints.**

Both types of decision trees fall under the Classification and Regression Tree (CART) designation.

## Terminology of Regression Trees

**Root:** This is the beginning of decision tree, which also represents the population sample. For example, say you want to decide the best performing employee in an organization based on various criteria, such as attendance of the employee, number of successful projects, number of employees he/she mentored, etc. So here, the entire population of employees is at the root of the decision tree.

**Leaf:** The terminal node is called the leaf node. In our example, the final best employee would be the leaf node or the terminal node.

**Decision Node:** Here the other nodes are divided into the further categories. In our example, the various criteria would determine a decision node.

**Child Node:** When a node is divided into other subparts, subparts are called child nodes. And the node which is divided is called the **parent node**.

**Problem Statement**

Find out the best accuracy of your model for the case of the Classification of the Heart Disease.You can use any classifier algorhithm.

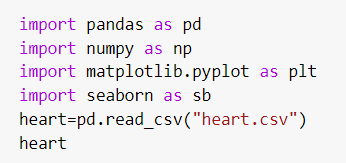
Problem Description of the Dataset

In this problem we have used the Heart disease csv file , obtained from the Kaggle website. Here , in this dataset sheet, there are about seven columns.Each column describing the age , sex , blood pressure bp and numerous other coloums , signifuing cholesterol level and others , which are the defying roles responsible for the heart diseases in human beings.

Problem Codes Steps

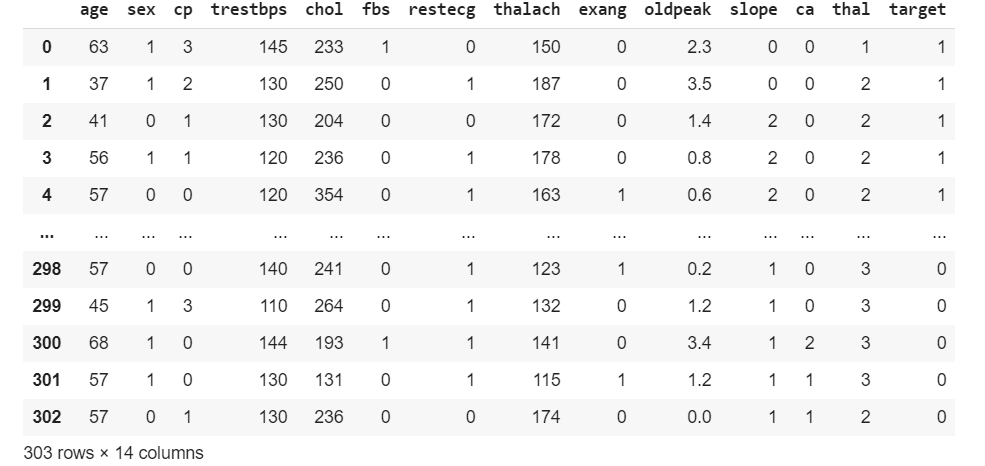
1.Importing Libraries –

We have imported the pandas , matplotlib and seaborn libraries to validate reading of the csv file to do the codes and the seaborn and matplotlib libraries are used to plot and draw the histograms.



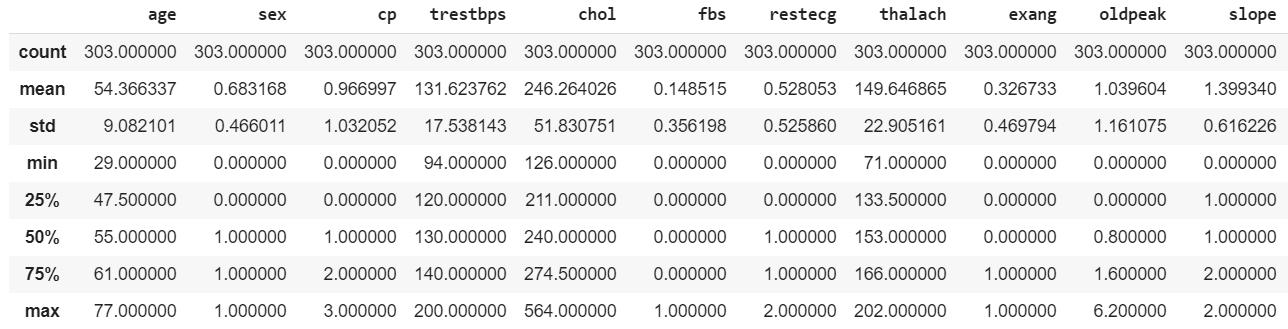
2.Reading of the Datasheet

This is the Datasheet of the heart disease we obtained from Kaggle.



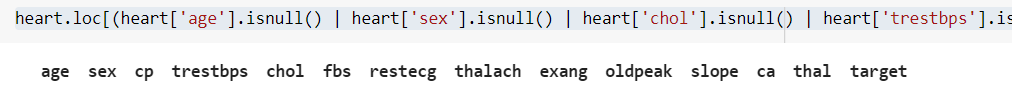
3.Checking for Null and mean

In this step , we are checking for Null values if present in the datasheet . After that , we checked for the mean std and max values if present in the datasheet.



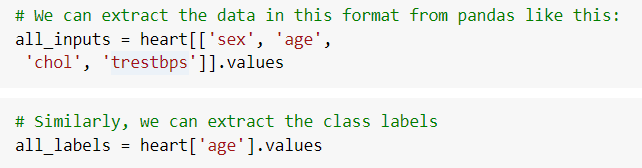
4.Tidying the data

Now that we've identified several errors in the data set, we need to fix them before we proceed with the analysis.



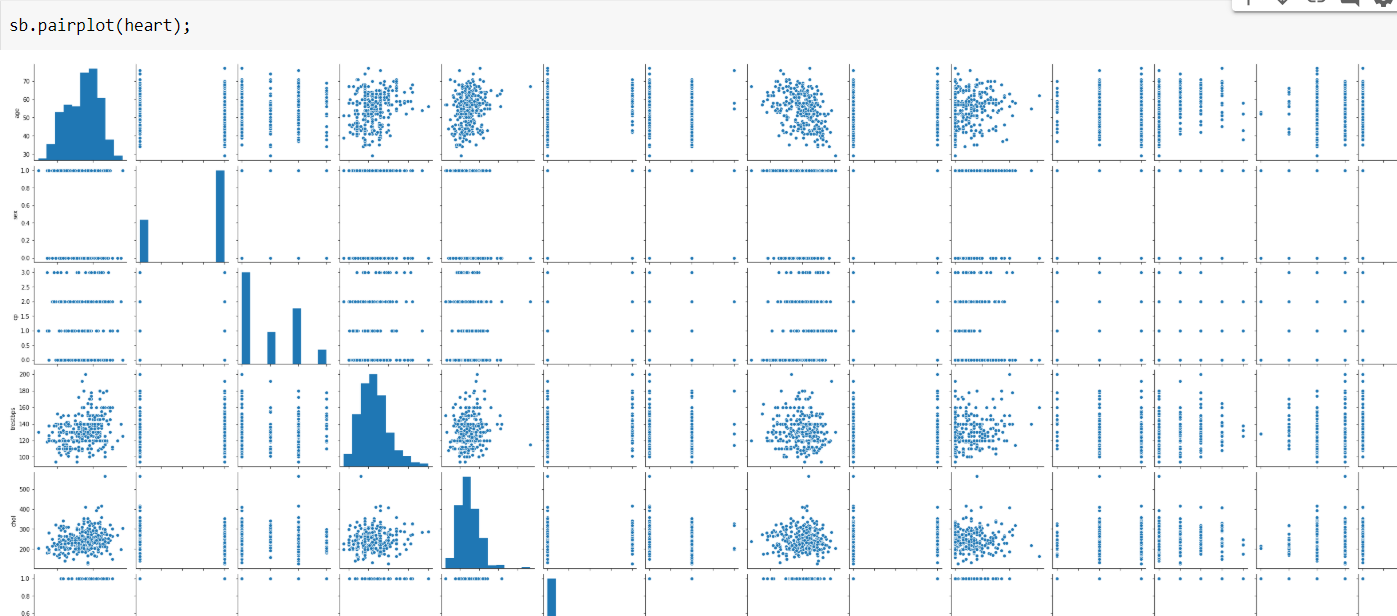
5.Testing our data

We can quickly test our data using assert statements: We assert that something must be true, and if it is, then nothing happens and the notebook continues running. However, if our assertion is wrong, then the notebook stops running and brings it to our attention.



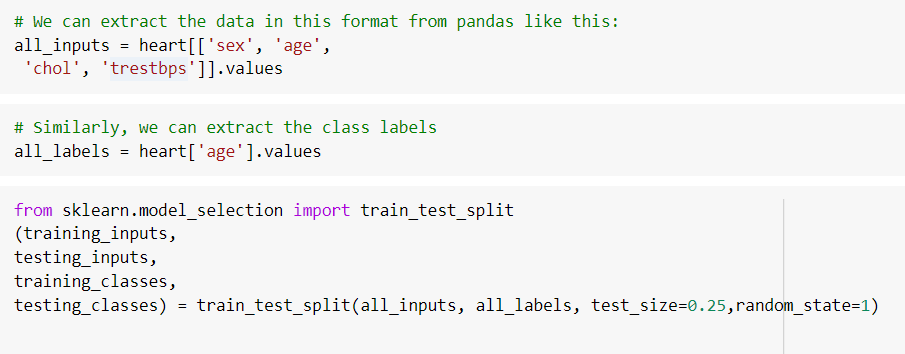
6.Exploratory analysis

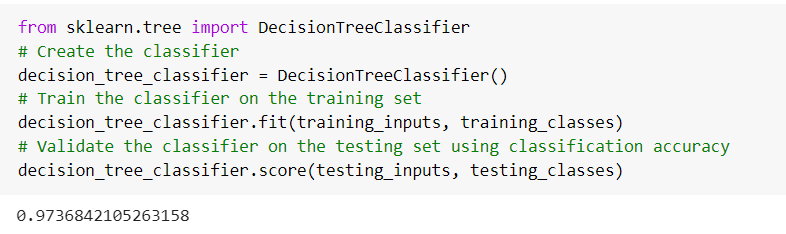
Exploratory analysis is the step where we start delving deeper into the data set beyond the outliers and errors. This is the stage where we plot all the data in as many ways as possible.



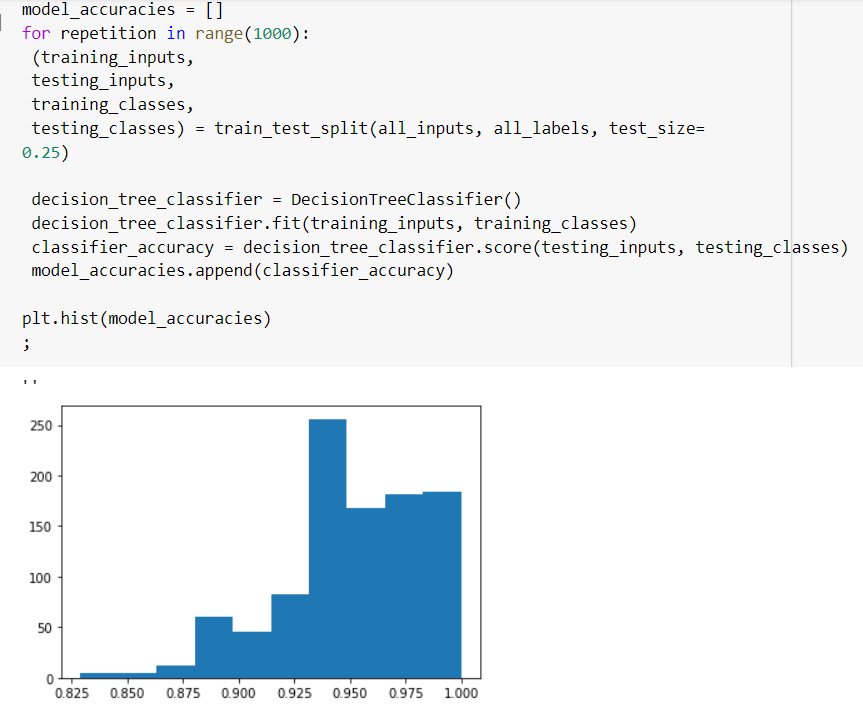
7.Classification

A training set is a random subset of the data that we use to train our models. A testing set is a random subset of the data (mutually exclusive from the training set) that we use to validate our models on unforseen data. Especially in sparse data sets like ours, it's easy for models to overfit the data: The model will learn the training set so well that it won't be able to handle most of the cases it's never seen before. This is why it's important for us to build the model with the training set, but score it with the testing set.





Here our model achieves 97% classification accuracy without much effort. However, there's a catch: Depending on how our training and testing set was sampled, our model can achieve anywhere from 80% to 100% accuracy.



This phenomenon is known as overfitting: The model is learning to classify the training set so well that it doesn't generalize and perform well on data it hasn't seen before.

# **Advantages:**

Compared to other algorithms decision trees requires less effort for data preparation during pre-processing.

* A decision tree does not require normalization of data.
* A decision tree does not require scaling of data as well.
* Missing values in the data also does NOT affect the process of building decision tree to any considerable extent.
* A Decision trees model is very intuitive and easy to explain to technical teams as well as stakeholders.

**Conclusions**

This technique of classification of heart Diseases uses different algorhithms to predict the accuracy of the model .First we use the seaborn to see the accuracy of the data , which is found to be about 97 percent. Now after that we used the decision tree classifier to obtain more accurate results which we truly found out to be almost nearly 98 percent to 99 percent. Hence our way of solving the problem is quite apt. This project is used to find the relations between various attributes , present within the table. This project at last finding the variance of age with other parametrs of heart like cholestrol , bp etc and represents in the form of Histograms.Thus at last our model achieves about 98 percent accuracy to find the analysis and thus helps in analysing the heart disease in a great way to predict its causes and factors affecting the disease so that we could understand the relations well and check in future.

Bibliography

I took the help to do this project from the websites of :

[www.scikitlearndocumentation.com](http://www.scikitlearndocumentation.com)

[www.kaggledatasheets.com](http://www.kaggledatasheets.com)

[www.analyticsvidhya.com](http://www.analyticsvidhya.com)