PLTHIYAGU (HTTPS://ID.ANALYTICSVIDHYA.COM/ACCOUNTS/PROFILE/)

tics
a

(https://www.analyticsvidhya.com/blog/)

ALGORITHM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/CATEGORY/ALGORITHM/)

BLOG/CATEGORY/MACHINE-LEARNING/)

TEGORY/TECHNIQUE/)

ion Tree in Machine Learning

COM/BLOG/AUTHOR/ABHISHEK-SHRM/), JUNE 30, 2020

Overview

- How do you split a decision tree? What are the different splitting criteria when working with decision trees?
- Learn all about decision tree splitting methods here and master a popular machine learning algorithm

Introduction

Decision trees are simple to implement and equally easy to interpret. I often lean on decision trees as my go-to machine learning algorithm, whether I'm starting a new project or competing in a hackathon.

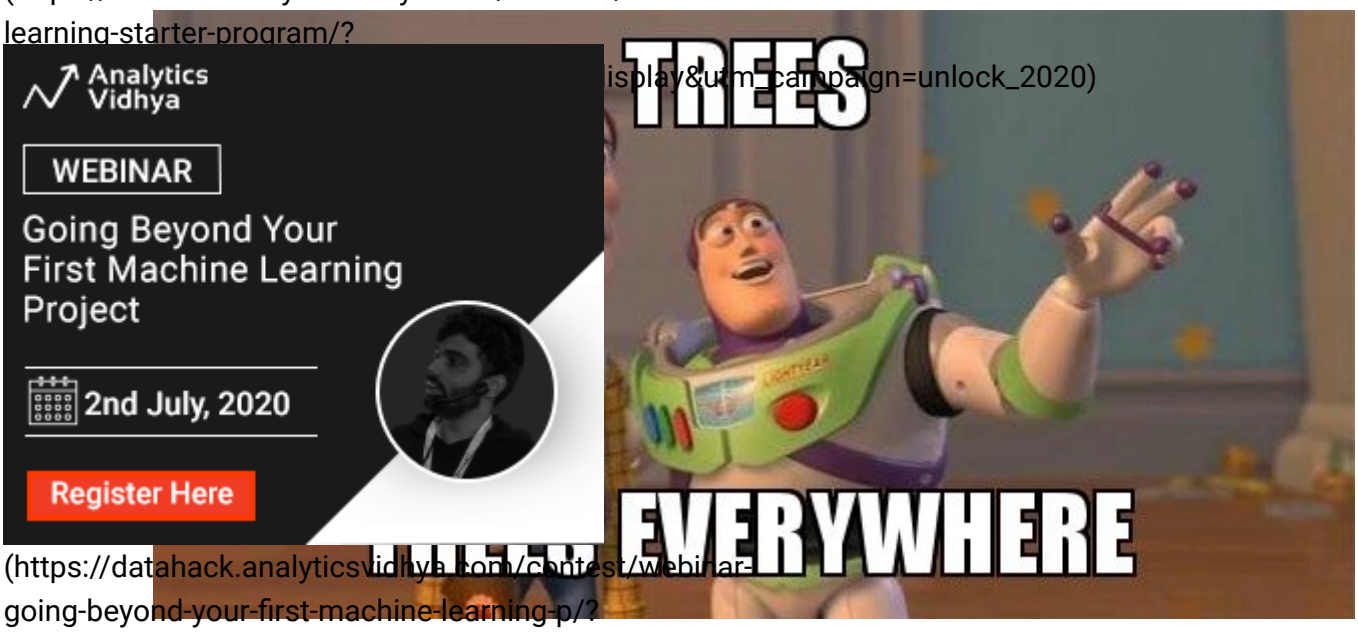
Unlock 2020

And decision trees are idea for machine learning newcomers as well! But the questions you should ask (and should know the answer to) are:

d Information Gain?

estions, you've come to the right place! Decision Tree is a powerful
as the building block for other widely used and complicated
est (https://www.analyticsvidhya.com/blog/2018/12/building-a-
al-world-data-products-ml-for-programmers-part-3/), XGBoost
/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/),
com/blog/2020/02/4-boosting-algorithms-machine-learning/). You
this topic!

(https://courses.analyticsvidhya.com/bundles/machine-learning-starter-program/?utm_source=Sticky_banner1&utm_medium=display&utm_campaign=unlock_2020)



(https://datahack.analyticsvidhya.com/contest/webinar-going-beyond-your-first-machine-learning-p/?utm_source=Sticky_banner2&utm_medium=display&utm_campaign=webinar)

Modern-day programming libraries have made using any machine learning algorithm easy, but this comes at the cost of hidden implementation, which is a must-know for fully understanding an algorithm. Another reason for this infinite struggle is the availability of multiple ways to split decision tree nodes adding to further confusion.

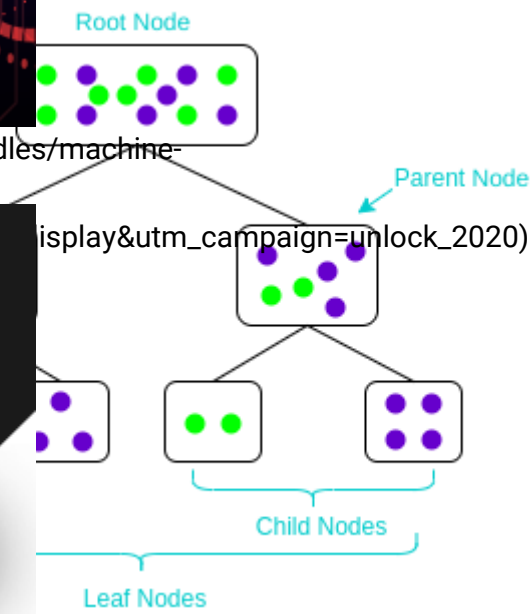
Have you ever encountered this struggle? Failed to find a solution? In this article, I will explain 4 simple methods for splitting a node in a decision tree.

I assume familiarity with the basic concepts in regression and decision trees. Here are two free and popular courses to quickly learn or brush up on the key concepts:

Unlock 2020

- *Fundamentals of Regression Analysis (https://courses.analyticsvidhya.com/courses/Fundamentals-of-Regression-Analysis?utm_source=blog&utm_medium=4-ways-split-decision-tree)*
- *Getting started with Decision Trees* (https://courses.analyticsvidhya.com/courses/getting-started-with-decision-trees?utm_source=blog&utm_medium=4-ways-split-decision-tree)

ted to decision trees which I'll be using throughout the article.



s divided into sub-nodes is known as Parent Node, and these sub-nodes are known as Child Nodes. Since a node can be divided into multiple sub-nodes, therefore a node can act as a parent node of numerous child nodes

- **Root Node:** The top most node of a decision tree. It does not have any parent node. It represents the entire population or sample
- **Leaf / Terminal Nodes:** Nodes that do not have any child node are known as Terminal/Leaf Nodes

What is Node Splitting in a Decision Tree & Why is it Done?

Before learning any topic, I believe it is essential to understand why you're learning it. That helps in understanding the goal of learning a concept. So let's understand why to learn about node splitting in decision trees.

Unlock 2020

Since you all know how extensively decision trees are used, there is no denying the fact that learning about decision trees is a must. A decision tree makes decisions by splitting nodes into sub-nodes. This process is performed multiple times during the training process until only homogenous nodes are left. And it is the only ell. Therefore, node splitting is a key concept that everyone should

ess of dividing a node into multiple sub-nodes to create relatively this, which can be broadly divided into two categories based on the

plitting method in detail. Let's start with the first method of splitting

Reduction in Variance

g the node used when the target variable is continuous, i.e., t uses variance as a measure for deciding the feature on which node is split into child nodes

$$Variance = \frac{\sum (X - \mu)^2}{N}$$

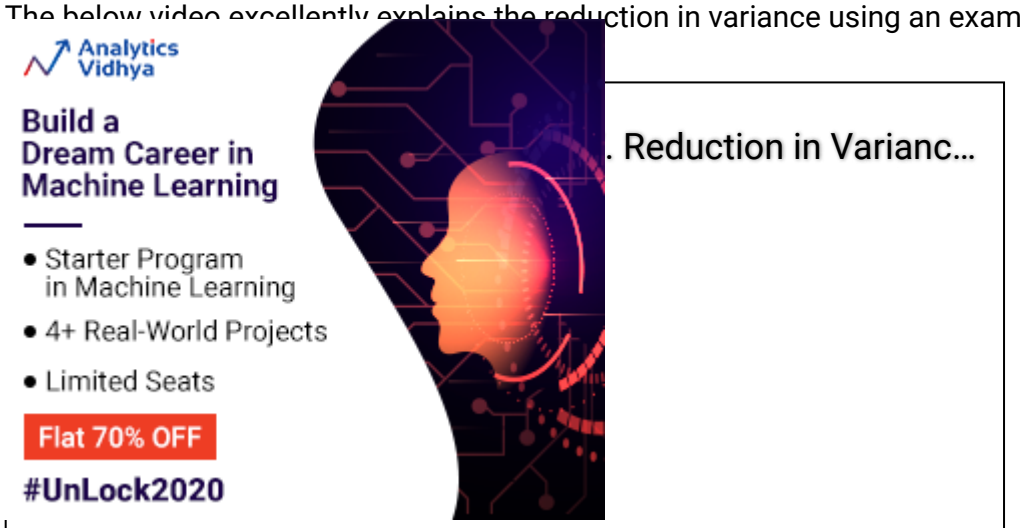
Variance is used for calculating the homogeneity of a node. If a node is entirely homogeneous, then the variance is zero.

Here are the steps to split a decision tree using reduction in variance:

1. For each split, individually calculate the variance of each child node
2. Calculate the variance of each split as the weighted average variance of child nodes
3. Select the split with the lowest variance

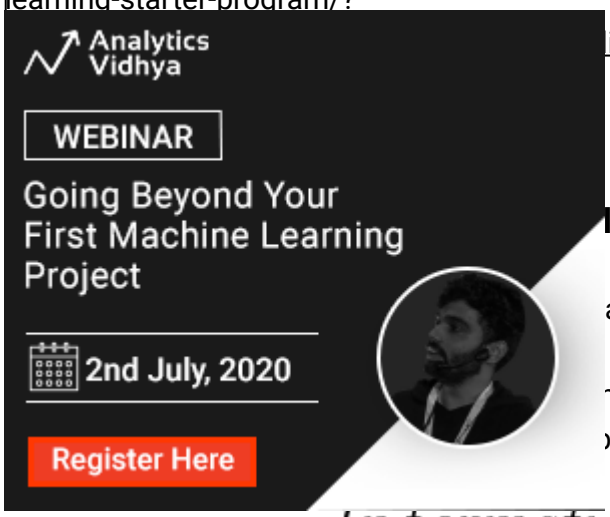
4. Perform steps 1-3 until completely homogeneous nodes are achieved

The below video excellently explains the reduction in variance using an example:



. Reduction in Varianc...

(https://courses.analyticsvidhya.com/bundles/machine-learning-starter-program/?utm_source=blog&utm_medium=display&utm_campaign=unlock_2020)



(https://datahack.analyticsvidhya.com/contest/webinar-going-beyond-your-first-machine-learning-p/?utm_source=blog&utm_medium=display&utm_campaign=webinar)

Information Gain

able? Reduction in variation won't quite cut it.

nformation Gain is used for splitting the nodes when the target
ot of the entropy and is given by:

$$Information\ Gain = 1 - Entropy$$

Entropy is used for calculating the purity of a node. Lower the value of entropy, higher is the purity of the node. The entropy of a homogeneous node is zero. Since we subtract entropy from 1, the Information Gain is higher for the purer nodes with a maximum value of 1. Now, let's take a look at the formula for calculating the entropy:

$$Entropy = -\sum_{i=1}^{n} p_i \log_2 p_i$$

Steps to split a decision tree using Information Gain:

1. For each split, individually calculate the entropy of each child node

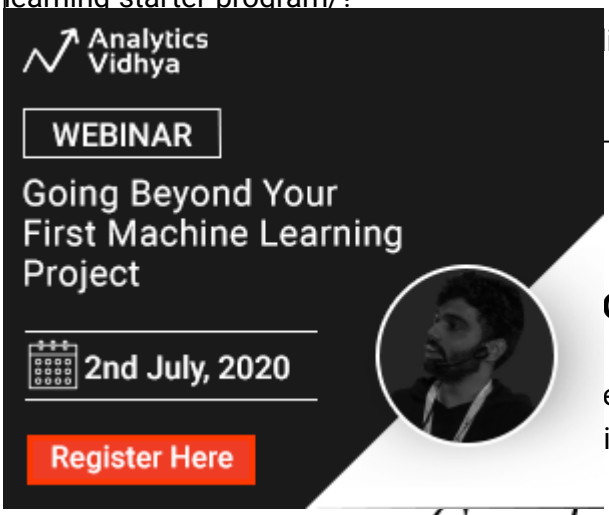
Unlock 2020

2. Calculate the entropy of each split as the weighted average entropy of child nodes
3. Select the split with the lowest entropy or highest information gain
4. Until you achieve homogeneous nodes, repeat steps 1-3

for splitting a decision tree:

. Understanding Infor...

(https://courses.analyticsvidhya.com/bundles/machine-learning-starter-program/?utm_source=banner2&utm_medium=display&utm_campaign=unlock_2020)



(https://datahack.analyticsvidhya.com/contest/webinar-going-beyond-your-first-machine-learning-p/?utm_source=banner2&utm_medium=display&utm_campaign=webinar)

Gini Impurity

es when the target variable is categorical. It is the most popular and ini Impurity value is:

$$Gini\ Impurity = 1 - Gini$$

Wait, what is Gini?

Gini is the probability of correctly labeling a randomly chosen element if it was randomly labeled according to the distribution of labels in the node. The formula for Gini is:

$$Gini = \sum_{i=1}^{n} p_i^2$$

And Gini Impurity is:

Unlock 2020

End Notes

Now you know about different methods of splitting a decision tree. In the next steps, you can watch our
e (https://www.youtube.com/playlist?list=PLdKd-
can take our free course on decision trees here
s/getting-started-with-decision-trees?
-decision-tree).

cles on decision trees below:

torial from Scratch (in R & Python)
/2016/04/tree-based-algorithms-complete-tutorial-scratch-in-
m=4-ways-split-decision-tree)
Weka (No Coding Required!)
(http://analyticsvidhya.com/blog/2020/05/decision-tree-weka-no-coding/?
utm_source=blog&utm_medium=4-ways-split-decision-tree)
dm Algorithm Should you Use?
/2020/05/decision-tree-vs-random-forest-algorithm/?
s-split-decision-tree)
Tree-Based Algorithms (Decision tree, Random Forests, XGBoost)
/2016/12/detailed-solutions-for-skilltest-tree-based-algorithms/?
s-split-decision-tree)

se share it with your friends and comment below with your queries

Related Articles

Unlock 2020

$$Gini\ Impurity = 1 - \sum_{i=1}^{n} p_i^2$$

neity of the node. **The Gini Impurity of a pure node is zero.** Now, you rmation Gain then, why do we need Gini Impurity?

> *o Information Gain because it does not contain utationally intensive.*

Here are the steps to split a decision tree using Gini Impurity:

1. For each split, individually calculate the Gini Impurity of each
2. as the weighted average Gini Impurity of child nodes
3. Gini Impurity
4. repeat steps 1-3

. How to Select the B...

Decision Tree Splitting Method #4: Chi-Square

Unlock 2020

Chi-square is another method of splitting nodes in a decision tree for datasets having categorical target values. It can make two or more than two splits. It works on the statistical significance of differences between the parent node and child nodes.

$$re = \sqrt{\frac{(Actual - Expected)^2}{Expected}}$$

class in a child node based on the distribution of classes in the
a class in a child node.

quare for a class. Take the sum of Chi-Square values for all the
for that node. Higher the value, higher will be the differences
between parent and child nodes, i.e., higher will be the homogeneity.

ng Chi-Square:

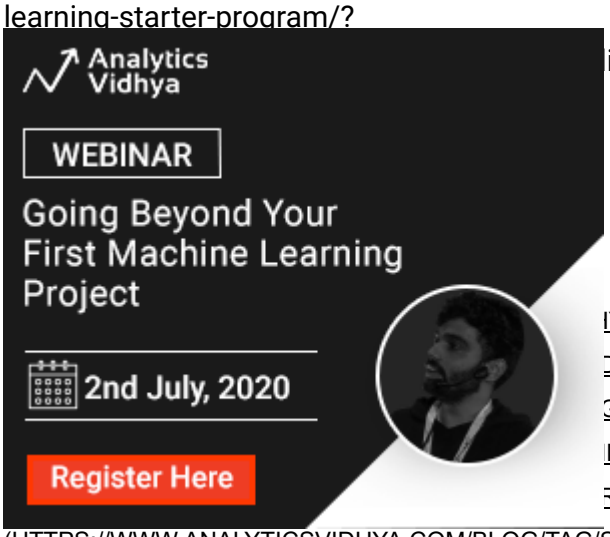
Chi-Square value of each child node by taking the sum of Chi-

split as the sum of Chi-Square values for all the child nodes
alue
repeat steps 1-3

re in the context of a decision tree:

Decision Tree Full Course | #5. Chi-Square to Select...

(https://courses.analyticsvidhya.com/bundles/machine-learning-starter-program/?utm_source=Sticky_banner2&utm_medium=display&utm_campaign=unlock_2020)



(https://datahack.analyticsvidhya.com/contest/webinar-going-beyond-your-first-machine-learning-p/?utm_source=Sticky_banner2&utm_medium=display&utm_campaign=webinar)

(https://www.analyticsvidhya.com/blog/2016/04/tree-based-algorithms-complete-tutorial-scratch-in-python/)
sed Algorithms: A Complete
from Scratch (in R & Python)

(https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/)
Decision Tree vs. Random Forest - Which Algorithm Should you Use?

IYA.COM/BLOG/TAG/CHI-SQUARE/), DECISION TREES
DECISION-TREES/), GINI IMPURITY
GINI-IMPURITY/), INFORMATION GAIN
NFORMATION-GAIN/), REDUCTION IN VARIANCE
REDUCTION-IN-VARIANCE/), SPLIT DECISION TREE
(HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/SPLIT-DECISION-TREE/)

NEXT ARTICLE

8 SQL Techniques to Perform Data Analysis for Analytics and Data Science

(https://www.analyticsvidhya.com/blog/2020/07/8-sql-techniques-data-analysis-analytics-data-science/)

•••

PREVIOUS ARTICLE

Hands-on NLP Project: A Comprehensive Guide to Information Extraction using Python

(https://www.analyticsvidhya.com/blog/2020/06/nlp-project-information-extraction/)

Unlock 2020

(https://www.analyticsvidhya.com/blog/author/abhishek-shrm/)

Vww.Analyticsvidhya.Com/Blog/Author/Abhishek-Shrm/)

s diving into data and generating insights from it. He is inspired self every day. He is always ready for making machines to learn ode and writing technical blogs.

(https://courses.analyticsvidhya.com/bundles/machine-learning-starter-program/?

(https://twitter.com/Abhi_dev2196)

667569132/)

isplay&utm_campaign=unlock_2020)

Reply

lyticsvidhya.com/blog/2020/06/4-ways-split-decision-tree/#comment-162096)

Method #3: Gini Impurity. You mension there Gini Impurity is a t variable is continuous.

Is it correct?

(https://datahack.analyticsvidhya.com/contest/webinar-going-beyond-your-first-machine-learning-p/?utm_source=Sticky_banner2&utm_medium=display&utm_campaign=webinar)

I have read below blog so I am confuse with it.

https://www.analyticsvidhya.com/blog/2016/04/tree-based-algorithms-complete-tutorial-scratch-in-python/

(https://www.analyticsvidhya.com/blog/2016/04/tree-based-algorithms-complete-tutorial-scratch-in-python/)

ABHISHEK SHARMA **Reply**

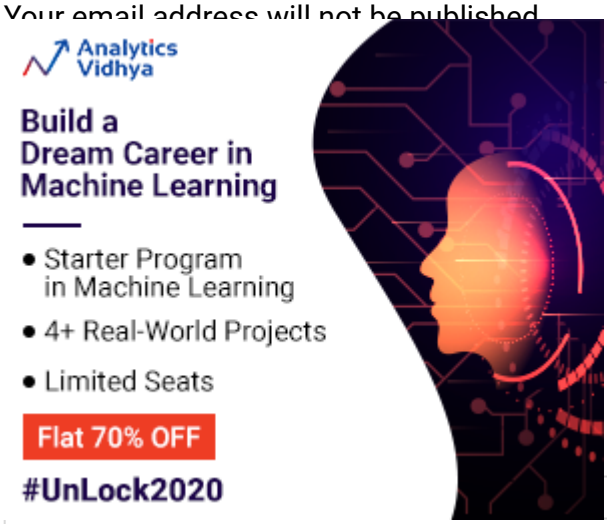
June 30, 2020 at 2:08 pm (https://www.analyticsvidhya.com/blog/2020/06/4-ways-split-decision-tree/#comment-162097)

Hi Maneesh, Thank you for pointing it out. I have made the necessary improvements.

Unlock 2020

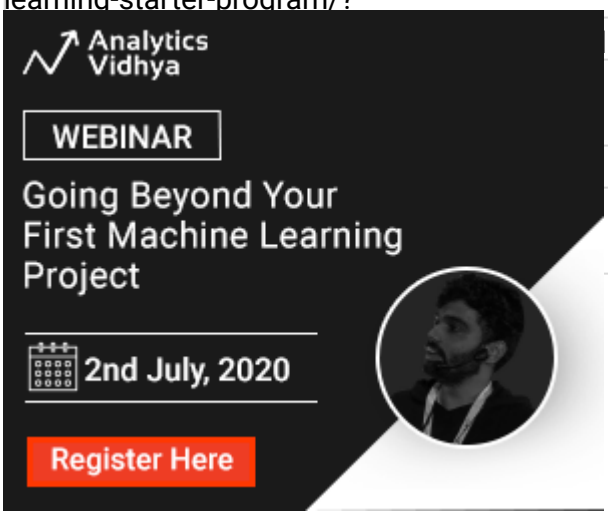
LEAVE A REPLY

Your email address will not be published.



(https://courses.analyticsvidhya.com/bundles/machine-learning-starter-program/?utm_source=Sticky_banner2&utm_medium=display&utm_campaign=unlock_2020)

Name (required)



(https://datahack.analyticsvidhya.com/contest/webinar-going-beyond-your-first-machine-learning-p/?utm_source=Sticky_banner2&utm_medium=display&utm_campaign=webinar)

Unlock 2020

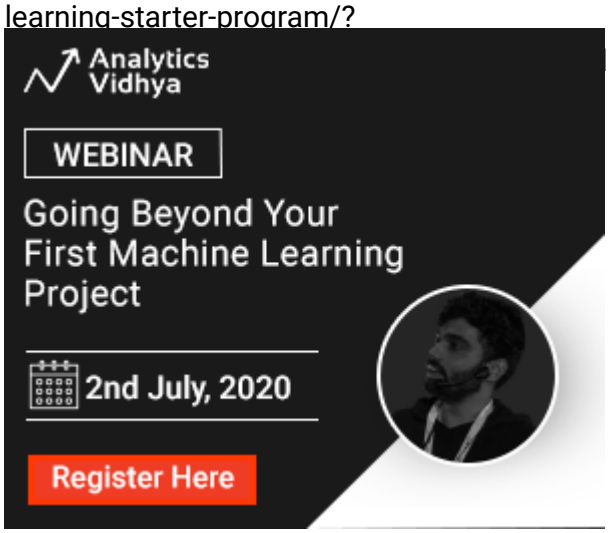
POPULAR POSTS

Unlock 2020

Hands-on NLP Project: A Comprehensive Guide to Information Extraction using Python (https://www.analyticsvidhya.com/blog/2020/06/nlp-project-information-extraction/)

ers' Library Version (v0.8.0)
20/06/hugging-face-tokenizers-nlp-library/)

(https://courses.analyticsvidhya.com/bundles/machine-learning-starter-program/?
isplay&utm_campaign=unlock_2020)



(https://datahack.analyticsvidhya.com/contest/webinar-going-beyond-your-first-machine-learning-p/?
utm_source=Sticky_ba utm_me n=di &utm_cam ai n=w)

(https://www.analyticsvidhya.com/)

Download App (https://play.google.com/store/apps/details? (https://apps.apple.com/us/app/analytics-

Unlock 2020

5 Powerful Excel Dashboards for Analytics Professionals (https://www.analyticsvidhya.com/blog/2020/06/5-excel-dashboards-analytics/)

Professional Should Try
020/06/3-advanced-excel-charts-every-analytics-professional-

achine Learning [Solution: SkillPower – Machine Learning, DataFest
log/2017/04/40-questions-test-data-scientist-machine-learning-
fest-2017/)

hms (with Python and R Codes)
017/09/common-machine-learning-algorithms/)

ustering Techniques (Skill test Solution)
017/02/test-data-scientist-clustering/)

RECENT POSTS

8 SQL Techniques to Perform Data Analysis for Analytics and Data Science (https://www.analyticsvidhya.com/blog/2020/07/8-sql-techniques-data-analysis-analytics-data-science/)

JULY 1, 2020

4 Simple Ways to Split a Decision Tree in Machine Learning (https://www.analyticsvidhya.com/blog/2020/06/4-ways-split-decision-tree/)

JUNE 30, 2020

Unlock 2020

id=com.analyticsvidhya.android) vidhya/id1470025572)

Analytics Vidhya

About Us (https://www.analyticsvidhya.com/about-me/)

Our Team (https://www.analyticsvidhya.com/about-me/team/)

Careers (https://www.analyticsvidhya.com/about-me/career-analytics-vidhya/)

Contact us (https://www.analyticsvidhya.com/contact/)

Data Science

Blog (https://www.analyticsvidhya.com/blog/)

Hackathon (https://datahack.analyticsvidhya.com/)

Discussions (https://discuss.analyticsvidhya.com/)

Apply Jobs (https://www.analyticsvidhya.com/jobs/)

Companies

Post Jobs (https://www.analyticsvidhya.com/corporate/)

Trainings (https://courses.analyticsvidhya.com/)

Hiring Hackathons (https://datahack.analyticsvidhya.com/)

Advertising (https://www.analyticsvidhya.com/contact/)

Visit us

(https://www.linkedin.com/company/analytics-vidhya/) (https://www.facebook.com/analyticsvidhya) (https://twitter.com/antgrasso) (https://www.youtube.com/channel/UCH6gDteHtH4hg3o2343iObA)

Privacy Policy Terms of Use Refund Policy

Unlock 2020