Basic Machine Learning Problem – Decision Trees in Python

Consider the made-up dataset in the attached powerpoint. Each meme card has 3 variables, Broad Audience, Like Rate, and Parent Share Rate. Consider the 15 cards with a confirmed viral yes or viral no as your training set.

The additional 5 meme cards with an unknown viral rate are your test set.

You’ll have to manually input and code the data into python to get started. Then create a **Decision Tree Model** that efficiently and correctly works for all memes in your training set. Please only use the 3 variables associated with the meme and do not create your own variables based on the picture, the pictures content, or the title of the meme. After your model is created, use your model to predict if the 5 test set memes will go viral.

****

**Analysis Questions**

1. How many Meme cards in the test set go viral?
2. What are the titles of the memes that went viral?

Data Analysis and Visualization Problem – License Usage and Optimization

Our team needed a way to determine how many RPA licenses are being used at a particular time.  With only the start and end times for each process and 1000's of processes per month, it can be tricky to sort through it.  Processes are ending and beginning at all hours and minutes of the day.  Some processes take a minute and some take 7 days.

The attached dataset titled "Concurrent Sessions March 2020.xlsx" has all the RPA processes run in production for March.  For each process it only contains the start time, end time, and processId. Each process will occupy 1 and only 1 license between the its start and end time.  Transforming that data sheet into useful insight and a useful graph is difficult.



Here are your questions for the challenge.

**Analysis Questions**

1. How many sessions were active at 3/8/2020 12:01 AM?
2. What is the lowest number of sessions that were used at one time during March?
3. Create a graph or dashboard to show license utilization during March, and explain how you created it.

Text extraction using Python and Regex

In the attached Annual Report for Estrella Resources ltd, we would like to prepare a python code that can extract useful snippets in the running text of the document. There are 5 syntax rules that needed to be coded in python that will help focus our model on the snippets from the document most useful.

**Steps to write a code:**

1. Use an appropriate Python library to read the document text
2. Create search strings for below scenarios (**Document Filer Name = Estrella Resources Ltd**)
   1. (Document Filer Name) NEAR 50 words (purchase) NEAR 50 words (project)
   2. (The Company) NEAR 50 words (sale) NEAR 50 words (project)
   3. (The Company) NEAR 10 words (acqui)
   4. (Document Filer Name) NEAR 15 words (buy) NEAR 15 words (Business)
   5. (Document Filer Name) NEAR 15 words (sold) NEAR 15 words (project)
3. Run all search strings rules on document text
4. Print the matching rules along with its number

**Resources:** Please use attached PDF document for Document Text.



**Expected Output:**

[#]. Excerpt words

For Example, if second string match in the text then output should have string number at the start followed by 15-20 words from matching excerpt. Please refer below sample output.

[b]. Company undertook a successful sales process of the Munda Gold Project with numerous offers being made. Auric Mining Limited

**Analysis Questions**

1. How many rules had matches in the document?
2. Share the output of your code after running through the document.