A picture containing graphical user interface

Description automatically generated

**Semester 1 Alternative Assessment 2021/22 End-of-Module Assignment**

|  |  |
| --- | --- |
| **Module Name and Code** | **MS5104 Decision theory and analytics** |
| **Student Name and ID** | **Nirod kumar das (21240437)** |

1.

Vidflix‘s CEO is interested in exploring the following association rule:

Casa de Papel 🡪 Squid Game ^ Stranger Things

As the Association rules identify the rules that predict how or why certain items are connected.it finds the co-relation and co-occurrence to establish relationship between data items within the data set. So from the above statement we want to see the behavior of users ,If a user is watching “Casa De Papel”,they are also likely to watch “Squid Games” ,”Stranger Things” Combined. We want to find a pattern that helps us to identify the co-relation between “Casa De papel” And “Squid games & Stranger things”.

Support is the proof of no of times a item appears in a data.it is the minimum percentage of frequency of a item in a sample transaction So in this case Support(Casa de papel) indicates how many users watch “casa de papel” from total sample data.

So we can find the support of “Casa de Papel” as

= P(Casa de Papel)/N , Where N=10 (no of users)

Support(Casa de Papel)= 4/10 = 0.4

Support(Squid Games, Stranger Things)

=P(Squid games,Stranger things)= 3/10 =0.3

Support(casa de papel,Squid Games,Stranger Things)

=P(casa de papel,Squid Games,Stranger Things) 3/10 =0.3

Confidence is the frequency of “if-then” statements are true. we can calculate confidence by formula -

Confidence = Support(casa de papel, Squid Games, Stranger Things)/ Support(Casa de Papel)

= 0.3/0.4 = 0.75

The final factor lift can be defined as frequency of an item wit respect to conditional probability. It is the comparison between the actual confidence and expected confidence. Lift can be calculated as

Lift = Confidence/Support(Squid Games, Stranger Things)

= 0.75/0.3

= 2.5

As Lift >1 ,shows the degree to which “casa de papel” and” Squid Games & Stranger Things” are dependent to each other and the persons watch “casa de papel” most likely to watch “stranger things and squid game”. It also indicates there is an opportunity and the rules highly useful for predicting the consequent in future data .

**2.**

**Confirmation bias**

on FilMine’s BA head statement **“we already know which are the right variables to select and build a new and accurate AI-enabled film recommendation system”** seems to have confirmation bias,as he wanted to prove a predetermined assumptions. They may not be showing the full story behind their AI system and showing only skewed data and may be they are briefing in advance to support the conclusion. By not seeing the real facts, only including the information that supports their believe and Interpreting data in a manner to support their existing believes and hiding some most important information .For example, consider a pharma company “austrazenica” claims their vaccine against covid is the best in the world and showed the data of recovery rate without showing is there any side effects or how many people after getting covid after few days of vaccinated or the death rate.

**Selection Bias**

Secondly their model may contain some selection bias, as they have a client base in Europe but we should also consider ,for the US market they have a proper dataset or information and not using the same data or algorithm to predict the US market, as user tendency can be changed or completely deferent from expectation with different region. Considering results from inadequate size of data can also be misleading and can be completely different from population data. For example, In a survey dataset for an OTT platform in Ireland which was mostly answered by males from age-group 20-50 and got a result most popular genres are action and Si-fi but they failed to consider about other gender and age groups which may result in different genre.

**Emergent Bias :**

Filmline’s AI model may have based on historical data to predict the trend or recommend which comes under emergent bias , but we can ignore in our case as all the user data are based upon historical data only.

**Stability Bias :**

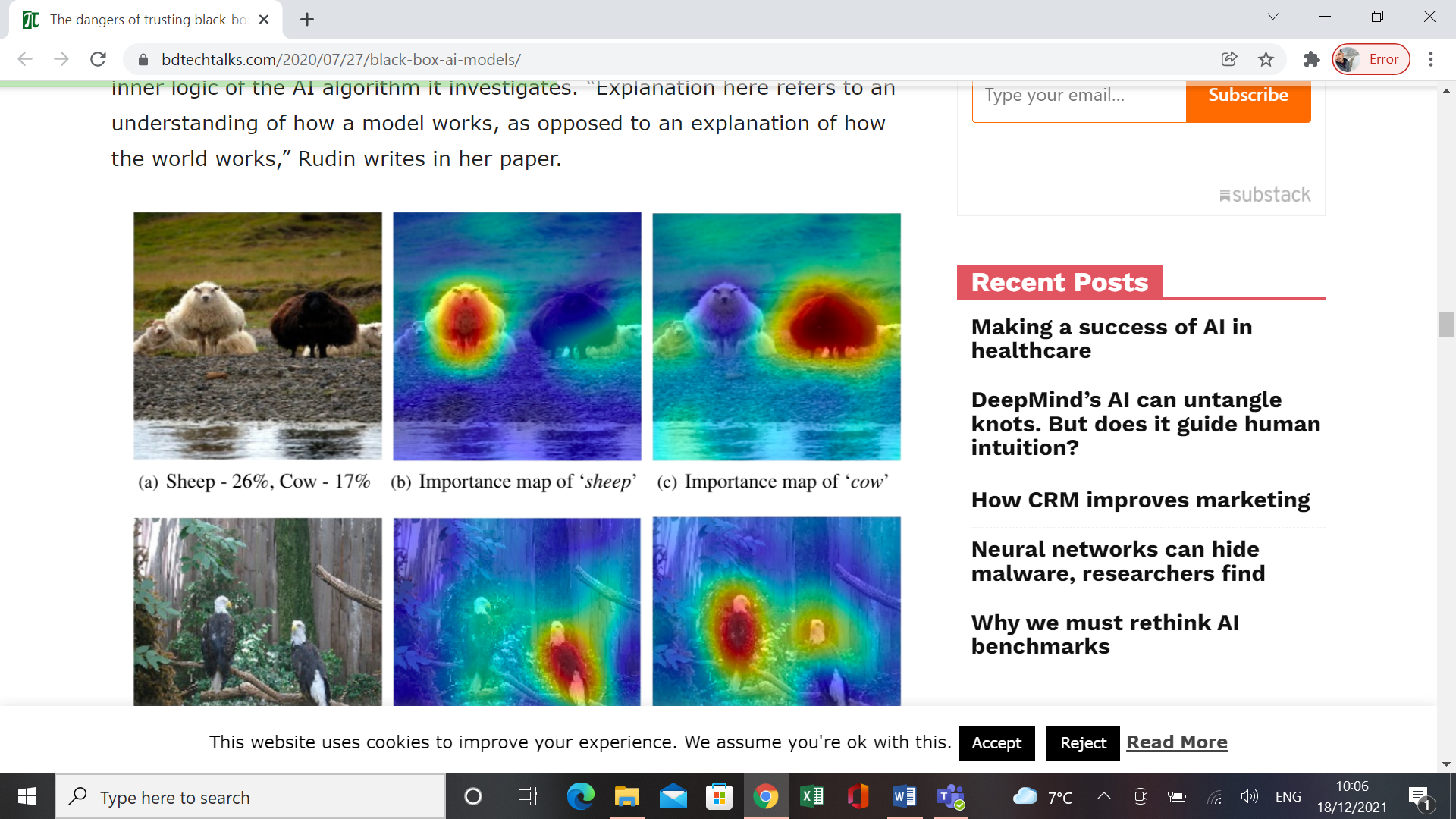
From the line “We already know which are the right variables to select” we may consider this as stability bias ,as its possible that they are considering the same old data or same old methods that they are considering to use this time also without enhancing the model or any enhanced variables which may results in less competitive model or less optimized model.

So irrespective of Filmine’s BA positive comment ,we should consider that their model is well optimized and is a competitive model that we can use the same AI model for US market. And As Filmine already developed a recommendation system for vidflix competator in Europe ,we should consider if there is any significant improvement in their performance and more customer acquisition or improvement in services in Europe market.

**Black Box :**

On his second statement **“even after Vidflix selects to purchase our solution, we will not reveal how our algorithm works as this is a company secret”** which clearly shows appearance of the concept of black box. Only knowing about the input and Output without the knowledge about its internal working ,may raise lot of questions and doubts. As they are not sharing their algorithm, it’s impossible to predict the efficiency of the model and trust the result which might be a result of some bias data. Also it raises a question of Vidflix security and data integrity and it may lead in to data piracy. We should also consider if their algorithm is having any access or is using any sensitive information of Vidflix which can be shared with competitors.

Filmline’s comment on trade secrecy can be understandable as it helps them to protect their design , information or knowledge from their competitors. for example “ [Google's](https://www.investopedia.com/articles/investing/020515/business-google.asp) search algorithm exists as intellectual property in code and is regularly updated to improve and protect its operations”. But as a client company it makes Vidflix in dark. Also if there is any problem or bug and algorithm showing any sensitive outcome, without the involvement of filmline team ,we can’t fix the problem or bug which may result in damaging company’s reputation and it also increase the dependency of Vidflix on Filmine .For example ,Due to some bug or error in the AI model, it 18+ category movies in kids section or showing kids and anime movies in adult section .That can create severe damage to brand image and create unpleasant user experience.

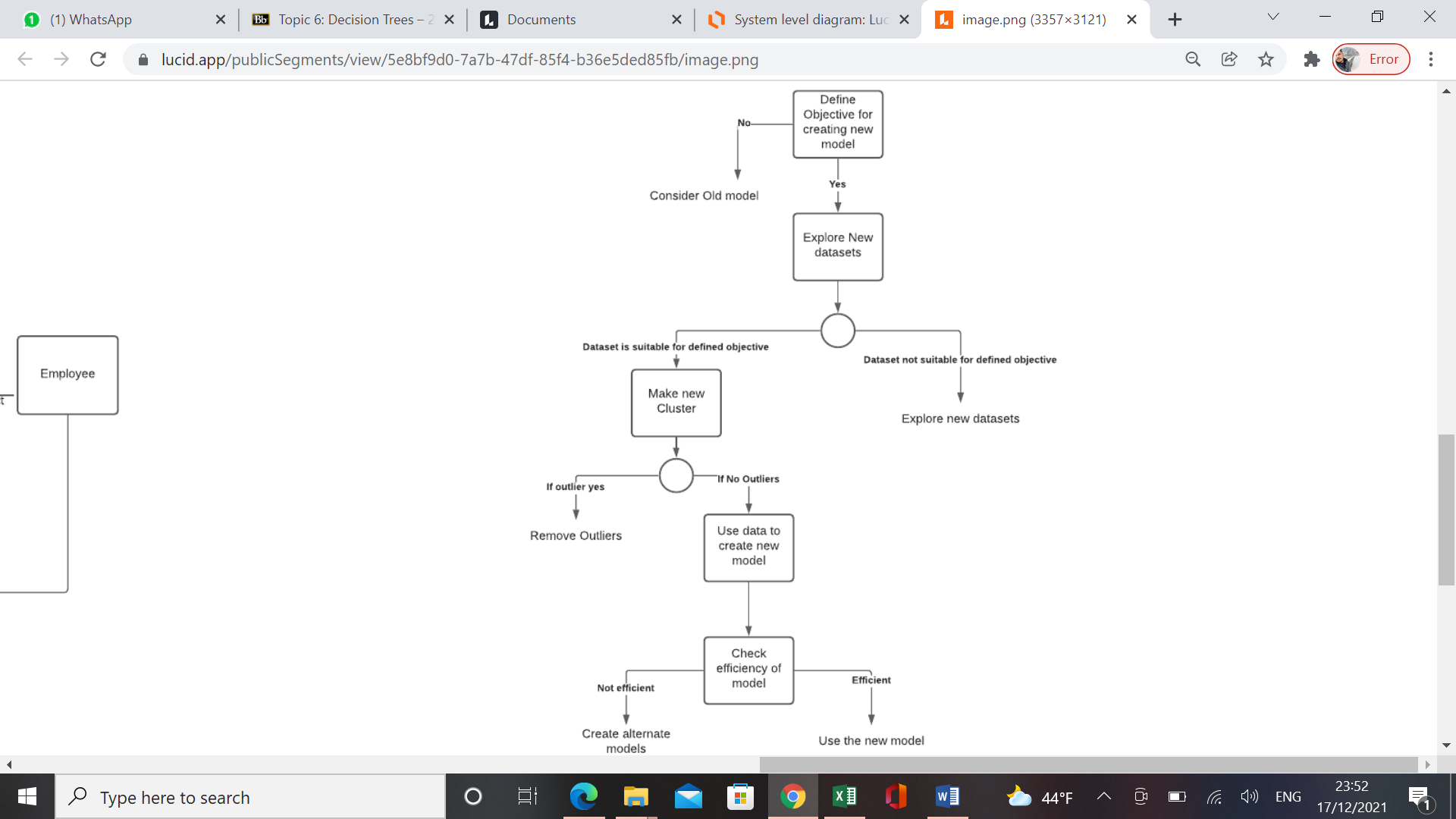


The above picture shows a real life example of failure of AI which is considering a brown sheep as cow. In such cases the company will be completely helpless as no one knows how the process works without the involvement of Filmine.

So as a conclusion to the second statement, I want to comment that ,If we Vidflix is purchasing Filmline’s solution with all trade secret may create a confusion and raise question on efficiency and company’s security policy, which may lead to adverse effect to the company and which is risky. Even though Filmline want to keep their algorithm as trade secret but they should have some transparency about their model and data they have been using. Also we can propose them ,they can patent their model instead of trade secret agreement, so that they can reveal the model and it’s working procedure which will lead us to take a conclusive decision.

**3.**

**Decision tree and clustering -**



As per the above decision tree we can explore to different datasets and decide usability of new clustering.

Clustering is a process of grouping object or items of similar categories. These similarities we can measure in terms of User Demographics, Region, language preference, age group etc. Before defining the clustering method we need to define some major parameters as well as objective .Also we should also consider if for the method is applicable for the particular data or particular issue. After selecting the proper method we should exclude the unnecessary data that’s not going to help in our objective keeping in mind not to exclude important information or columns which can play a vital role. We should consider about Outliers before clustering which can result in skewed data. As clustering compares each segmentations with permutation and combination we should give enough time to the system to predict the best cluster. After getting output we should consider statistical tools to evaluate output depending upon Problem statement or type of data or clustering method.

**Combining clustering and classification :**

While clustering is an unsupervised model, Classification is based upon labelled data or supervised model. When we only have data labels , clustering doesn’t help in founding class boundaries. So it is advised to blend clustering with classification. For example suppose vidflix is having 20k movies in library from which only 10-20% movies have been assigned with genres. So clustering each genres and then label each cluster using majority label in the particular cluster. Also clustering can be used as a processing tool followed by classification.

After a well processed clustering , lets assume that we have dataset of customer demographics including age group ,marital status, gender etc and as per that we can offer customized promotional plans to particular customer segments and introduce promotional prices to attract new customers and increase no of subscriptions .For example- most married couples or a family account want if they can add more devices (mobiles/laptop/Tv) under a single subscription with single user ID ,and we can understand from our clustering which users are married or a family account and we can offer them promotional schemes for the same. And also as per the age category we can select popular movies streaming by the same age groups and show them in a trending list.

Secondly, let’s consider if we have a customer survey data which includes psychographics data like personality, interest, hobbies etc , and as per that we can cluster groups and we can suggest particular genre or similar type shows and also we can see the top genres or categories of our users and we can include more no of movies of similar categories. which will include values to Vidflix and help in boosting customer satisfaction. For example : people who like to watch action movies may like Sci-fi or drama genre. For penetrating Us market also we can analyze most popular categories or top age groups or top preferred language , and we can include those movies in our platform and we can use our clustering for assortment planning.

**Combining Decision tree and smart Method :**

Finally after evaluating and making new clustering from different datasets, we can include smart method as a deciding factor and Assigning weightage of each cluster. For the same we have to assign an objective and all relevant attributes and proxy attribute. After then the CEO can assign values randomly on his preference from 0-100 to different attributes . On another sheet ,the CEO can give swing weight to each attributes and we can calculate normalized upon that .In the end we can multiply normalized weight with the assigned value for each attributes and finalize the key attribute, which we can use in decision making.

References :

<https://www.investopedia.com/terms/t/trade-secret.asp>

<https://en.wikipedia.org/wiki/Association_rule_learning>

https://bdtechtalks.com/2020/07/27/black-box-ai-models/