**Data Mining Assignment 2 Report**

* **Vishal Deo**

**Under Guidance of :- Prof. Dr. Weimao Ke, Prof Dr. Heejun Kim**

**Index**

|  |  |  |
| --- | --- | --- |
| **Sr.no** | **Topic** | **Page no** |
| **1** | **Data Cleaning Questions** | **3** |
| **2** | **Model Peformance Question** | **6** |
| **3** | **Association Rule Mining and Intrestingess** | **8** |
| **4** | **Yelp Association Rules Insights** | **15** |
|  |  |  |

**Data Cleaning Questions**

**1.1. The first problem**

**1)What is the first problem you found? How did you find it?**

As from the codes below the value [eigenvector] has missing "NA" values we check if there are any "NA" values in any of the coloumns and find that eigenvector has a few.

**2)How will you handle this problem? Please code your approach.**

step 1:- Use is.NA function to check any attribute has any NA values

step 2:- Use np.median to check the median of the attribute eigenvector

step 3:- Use the fill.na to replace the median values with the na values

**3)What are pros/cons of your approach?**

Pros:

1) Easy and fast.

2) Works well with small numerical datasets.

Cons:

1) Doesn’t factor the correlations between features. It only works on the column level.

2) Will give poor results on encoded categorical features (do NOT use it on categorical features).

3) Not very accurate.

4) Doesn’t account for the uncertainty in the imputations.

**1.2. The second problem**

**1) What is the second problem you found? How did you find it?**

To check if there are any attributes which have redundancies by using correlation coefficients

**2) How will you handle this problem?. Please code your approach?**

-------

Visualize the correlation coefficients by using heatmap function of seaborn library where it can be observed that the [degree and betweeness = 0.95] & [degree and eigenvector = 0.95 ] has very high co-related values

Dropping the highly corelated attributes as they might have redudant values by using the drop() function

**3) What are the pros and cons of using this approach?**

----

Pros:-

Reduces Preparation Time: Fewer data ensures the applications operate better.

Decreases overfitting: less repetitive data allows less ability to make noise-based decisions.

Increases the quality: Fewer inaccurate data implies greater simulation performance.

-------------------

Cons :-

Few values which do not have redudant values are also removed resulting in affecting in the model

**1.3 The third problem**

**1) What is the third problem your found? How did you find it?**

Answer:-

found 2 problems :-

a) While building the model i was encountering error that the values in Y are not of integer type so the model cannot be built, after analysis by using the value\_counts() function it was observed the Y variable which is the [Class] attribute had 3 and not 2 values

a) Useful

b) Not Useful

c) Not\_Useful

so values Useful = 1 and Not Useful = 0 were converted to numeric but Not\_Useful were present not allowing the model to run

b)As there were few dimensions which had mostly 0 values such as [joy,love,affection,liking and enthusiasm] as positive and [horror,despair,dislike,sadness] as negative

**2) How will you handle the problems. Please code your approach**.

step 1:- first we will standarize a)not useful b) not\_useful = 0

step 2:- mergeing the dimmensions postive and negative and dropping the dimmensions [joy,love,affection,liking and enthusiasm] and [horror,despair,dislike,sadness]

**3) What are pros/cons of your approach?**

pros:-

1)Faster complilation as less dimmensions

2)model execute sucessful

Cons:-

1)Few imporant values also might be removed

**Model Peformance Question**

### 2.1. Performance

1. **What was the best performance you got? Please provide the code for your best classifier. The student that achieves the best accuracy will be given ONE bonus point towards his/her final grade and, of course, fame and glory.**

**Answer :-**

**Best Performance :- Z\_Scale**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Forward** | **Floating** | **Scoring** | **CV** | **Accuracy** |
| False | True | accuracy | 100 | 0.7601990049751244 |

**2.2. Insight and Explanation**

**Provide a couple of paragraphs description of what you tried, what worked, and what did not work. Describe the lesson you got by this exercise. Please be comprehensive to deliver what you have done. Perhaps, using graphs or tables will be helpful to find a meaningful pattern from your experiments.**

**Answer:-**

**Evaluated the model with the Min**-**max** normalization: Guarantees all features will have the exact same scale but does not handle outliers well.

**Z**-**score** normalization: Handles outliers, but does not produce normalized data with the exact same scale. Tried using different parameters and parameters and tried setting the "forward" argument of the Sequential Feature Selector as "False" to use backward feature selection. Also tried to set the "floating" argument as "True", then to will use bi-directional stepwise feature selection and worked with CV options models produced the same result. In the backward feature selection, a function for which the model has the lowest ranking. Forward selection goes the other way: it begins from an empty collection of options and introduces a function that better increases the current ranking.

The Min Max Result was below the baseline which is of 0.7502487562189055 and Min-Max produced accuracy of 0.73, Z-score produces the accuracy of 0.760

Output With Different Parameters :-

**Min-Max :-** Ensures that all features have the exact same scale, but do not handle outliers well.

|  |  |  |  |
| --- | --- | --- | --- |
| **Forward** | **Floating** | **CV** | **Accuracy** |
| True | True | 10 | 0.7343283582089553 |
| True | False | 20 | 0.7343283582089553 |
| False | True | 20 | 0.7343283582089553 |

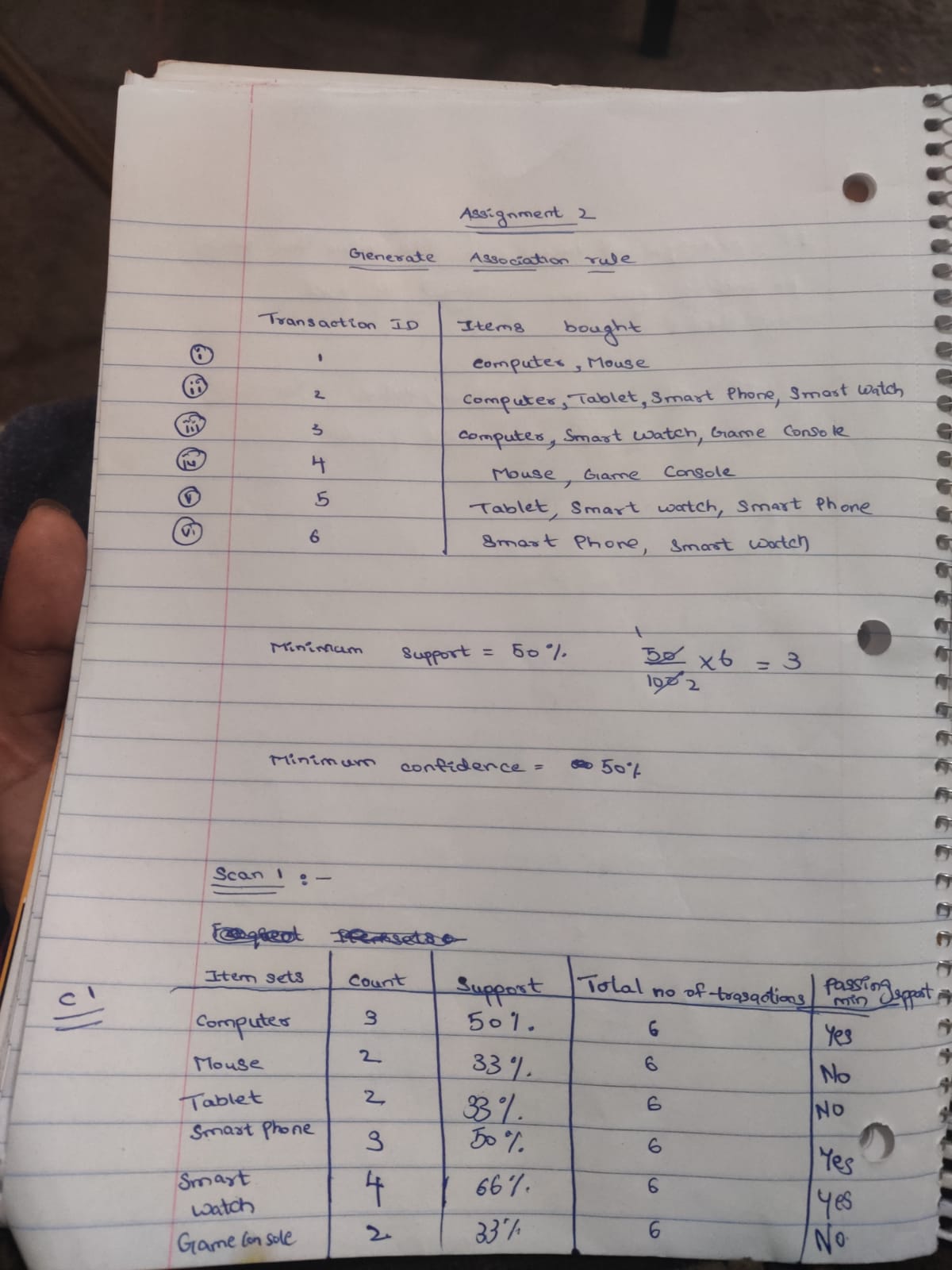
**Z scale:-** Handles outliers, but does not produce normalized data with the exact same scale.

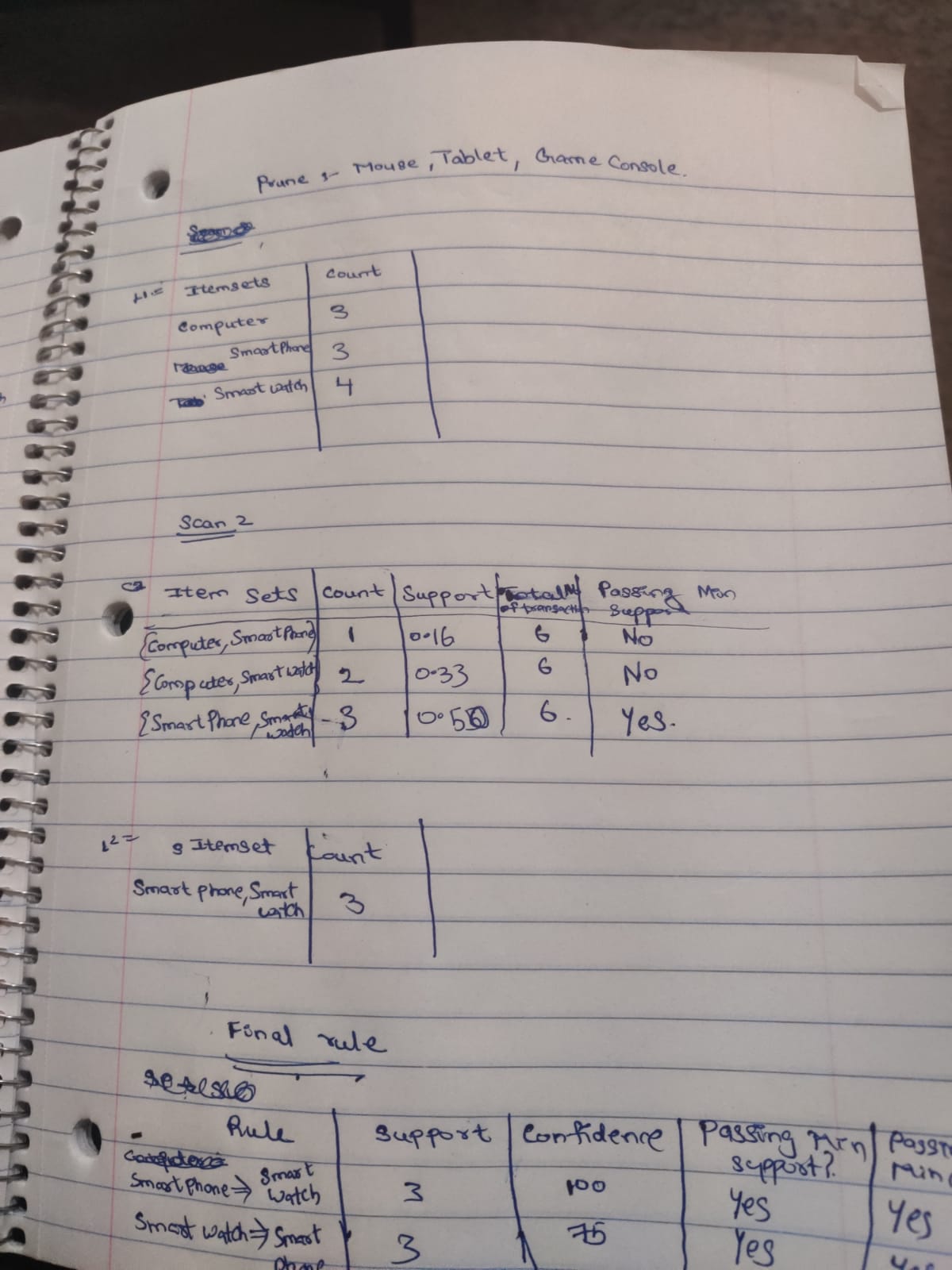
|  |  |  |  |
| --- | --- | --- | --- |
| **Forward** | **Floating** | **CV** | **Accuracy** |
| False | True | 100 | 0.7601990049751244 |
| True | True | 10 | 0.7601990049751244 |

In this excerise I learned about Logistic Regression Classification and Scaling methods (Min-Max and Z-Scale) the importance of the parameters of these functions .

**3. Association Rule Mining and Intrestingess**

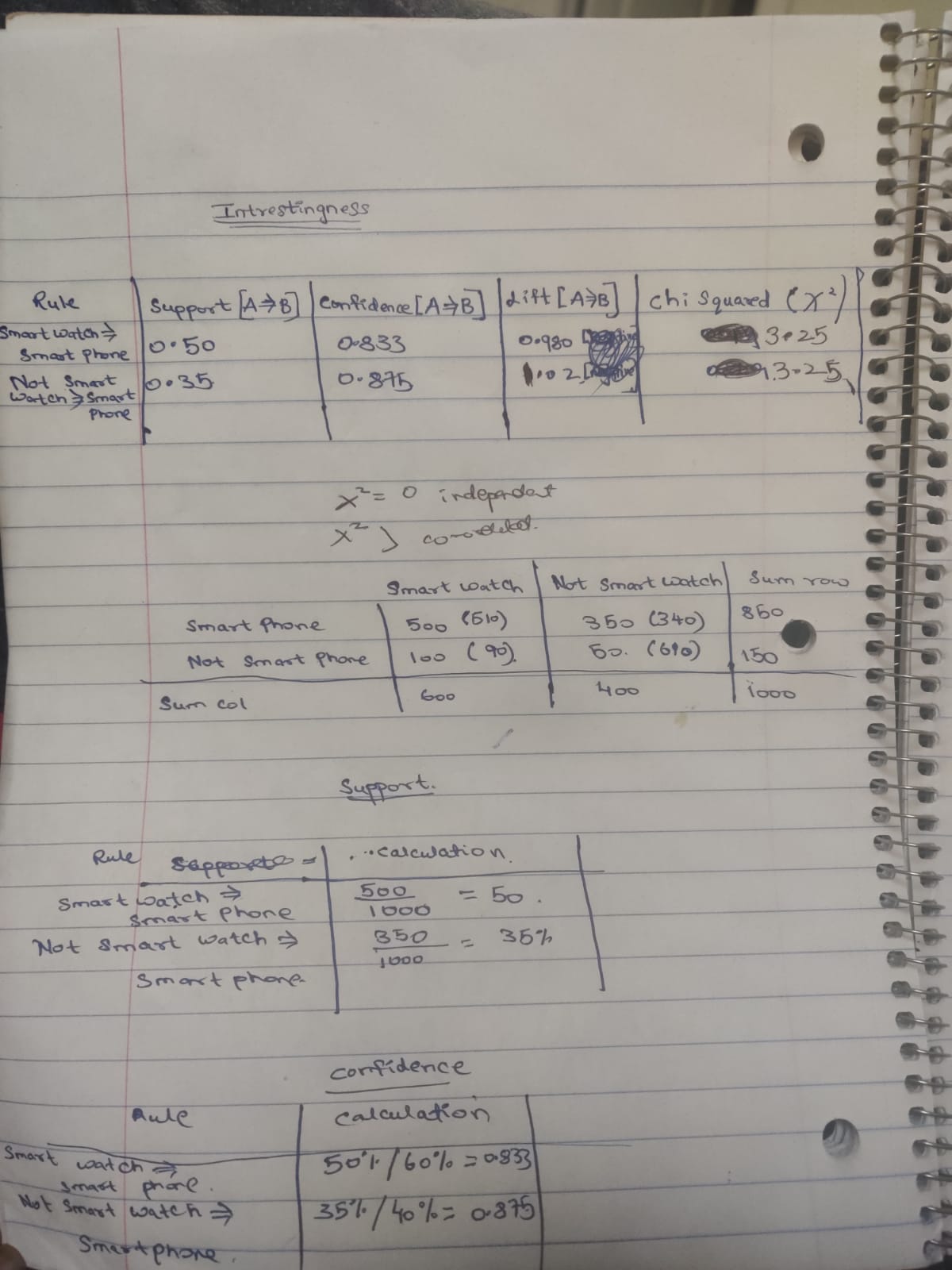
**3.1. Rule Calculation**

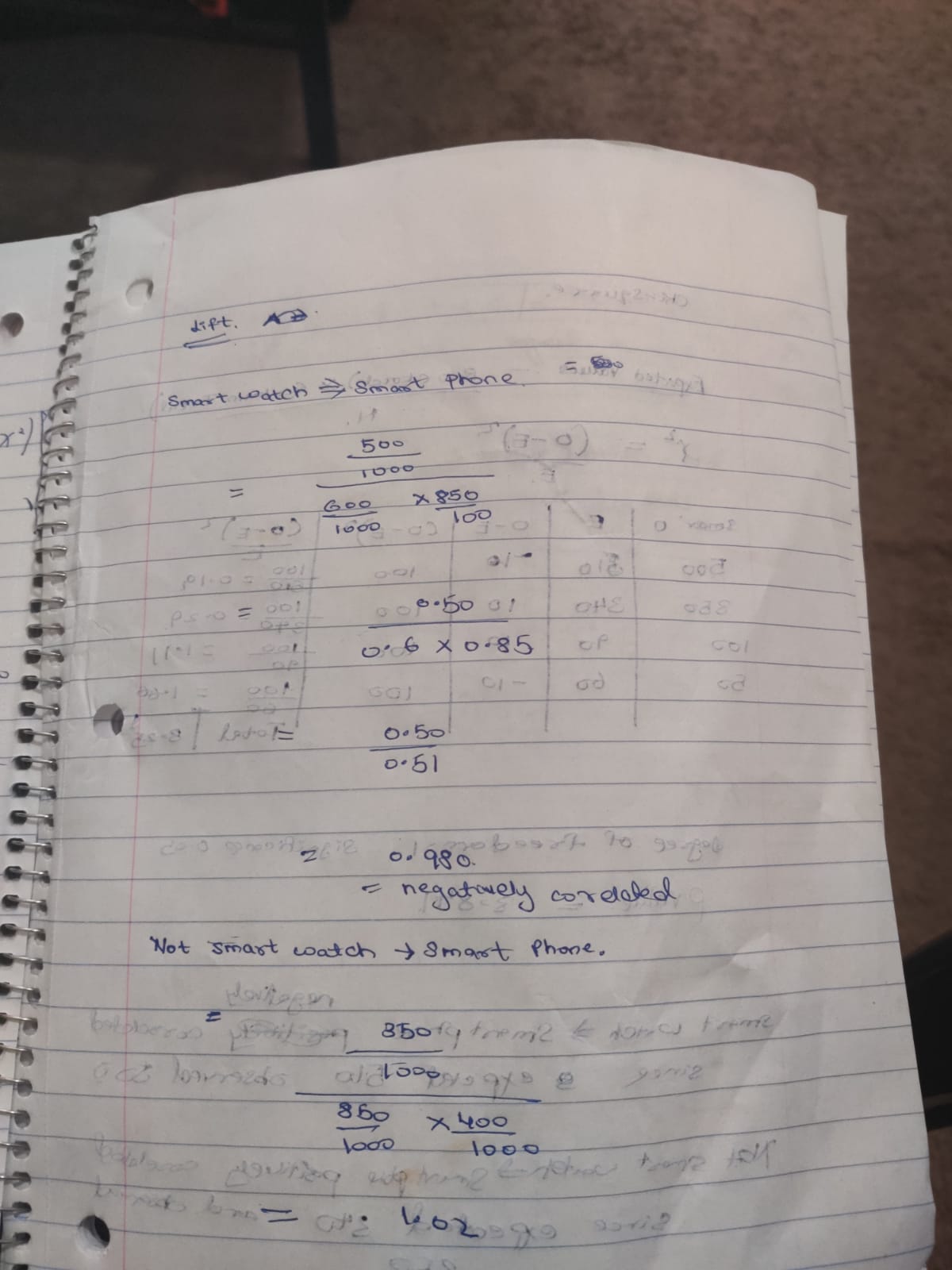
3.1.1 Calculation of support 

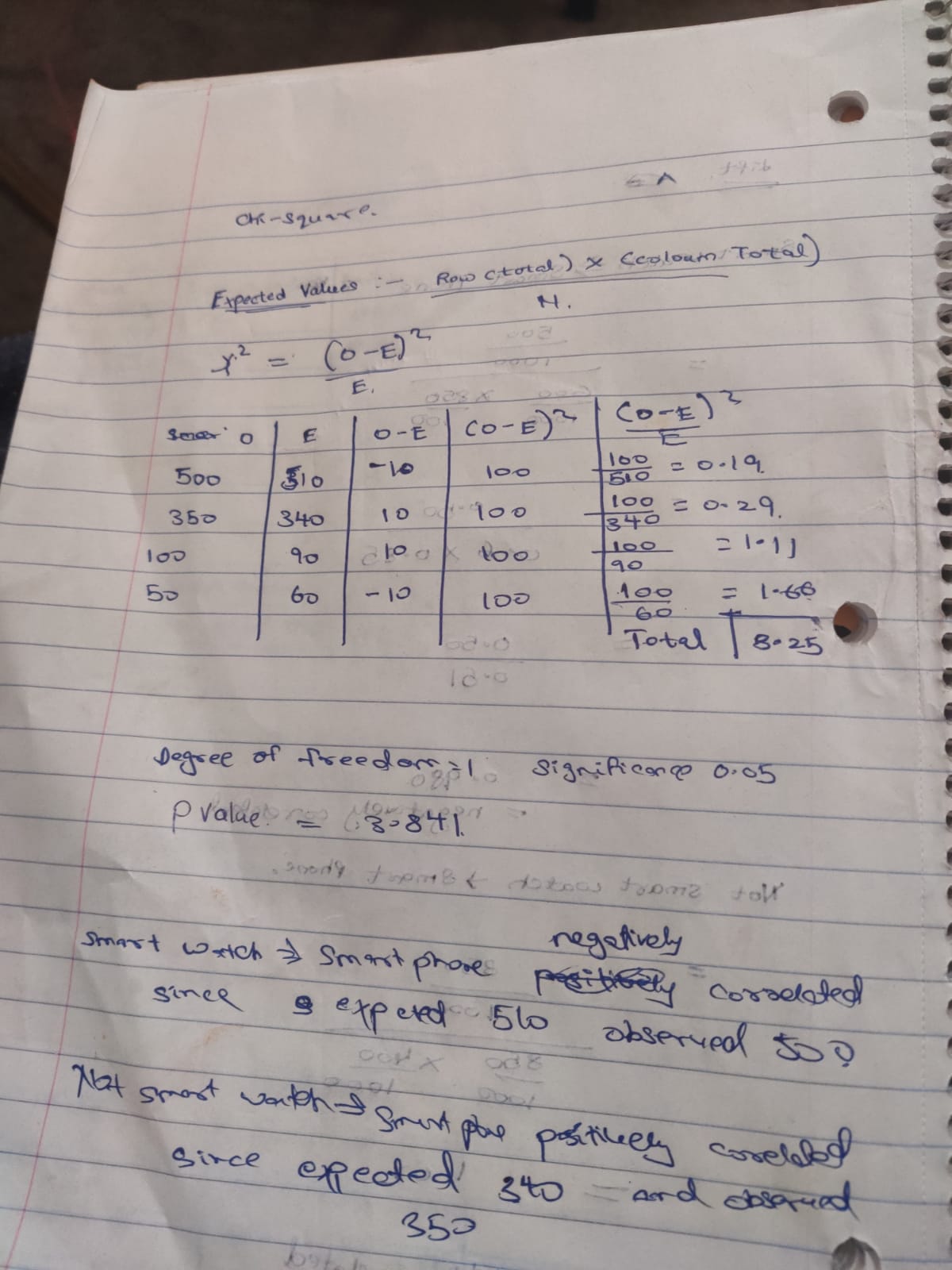


**Final Rule :- Smart Phone 🡪 Smart Watch**

**Smart Watch 🡪 Smart Phone**

**Intrestingness**





**3.2.2 Insight and Explanation**

**If you calculated correctly, each measure might tell you different stories. Let's assume that both minimum support and minimum confidence are 35% (i.e., .35). What do these interestingness measures tell you? Please explain what each measure means. In other words, can you tell whether there is a strong association or not? What are the pros/cons of each measure, particularly related to this example? How are you going to use these measures in the future according to what you learned?**

**Answer:-**

**Insight 1 :-**

**Support :-** Support is an example of how frequently the item set occurs in the data collection.

For example:- In the Smart Phone and Smart Watch table the frequency of Smart phones bought when customer purchases Smart Watch is 500 out of the total 1000 transactions then the Support is 50% where as if a customer does not purchase smart watch but purchases smartphone has a frequency of 350 out of 1000 transactions so if we assume that the support is 35% both the rules are above the threshold

**Pros :-** Shows Frequency of the transactions and help to determine association rule

**Cons :-** Can be misleading

**Insight 2:-**

**Confidence:-** Confidence is the what we set where Y is the item that is purchased when X is purchased. So out of total no of times X is purchased , no of times Y is purchased when X is purchased. We set confidence for considering the meaningful associations between the items.

**For example :-** In the Smart Phone and Smart Watch table the confidence for customer purchasing smart phone if someone purchases smart watch is 0.833 while the confidence of curstomer not purchasing smart watch when they buy smartphone is 0.875 so when we take the support and confidence together it seems confusing and a requirement for additional analysis as even though support for Smart Phone and Smart Watch Purchased together is higher than the second rule the confidence shows contrary to what we observe

**Insight 3 :-**

**Lift :-** Lift can be understood as a ratio of two percentages,interestingness measure lift captures such as correlation in the sense that it tells us whether the LHS influences the RHS positively or negatively. Therefore, using Lift instead of confidence as a criteria for discovering association rules can be more effective. If the Lift is 1 it is idependent, if lift<1 negatively correlated and if lift>1 there is postive correlation

**For Example :-** In the Smart Phone and Smart Watch Table the Lift for customer purchasing is 0.980 there appears to be a negative co-relation and not smart watch and smart phone are postively corelated

**Cons :-** influenced by null transactions but in the table of Smart Watch and Smart Phone there not as many null transactions

**Insight 4:- Chi Squared Measure :-** We use this test when we have two categorical variables and want to determine whether there is a significant association between the two variables.

Forexample:- In the Smart Phone and Smart Watch table since the chi-squared values are not significant the chi square value is 3.25 which is not significant with significance 0.05 and degree of freedom 1, we reject the null hypothesis that there's no difference between the means and conclude that a significant difference does exist

**Pros:-** Identify strong association rule

**Cons :-** influenced by null transactions

**Strong Association Rule** :- Both of the rules above the support and confidence threshold but as confidence value shows contrary to support we cannot determine

**How are you going to use Intresginess measure ?**

I look ahead to analyze the covid situation using the intresginess measure by anlazying the pattern between a people in an area and wether they wear mask or not to anaylze the increasing awarness among people with the passage of the time since the virus started for each month for a specific region can be intresting insights

Example :- [hypothetical values]

Total 1000 people in Jan

Wears Mask || Does not Wear Mask

Covid 200 400

Not Covid 100 300

Total 1000 people in March

Wears Mask || Does not Wear Mask

Covid 600 50

Not Covid 300 50

Intresingness measure will help analyze change /awarness in these months and changing patterns in this unprecidented times

**Yelp Association Rules Insights**

### 4.3. Insight and Explanation

**Provide a couple of paragraphs' description of what rules you have found. Did you find any interesting rules? Pick one interesting rule and explain how it is interesting and how you used evaluation measures. Pick one not-interesting rule and explain how it is not interesting and how evaluation measures overate the rule. Describe the lesson you got by this exercise. Please be comprehensive to deliver what you learned. Perhaps, using graphs or tables will be helpful in finding a meaningful pattern from your experiments.**

* **The following Intresting rules were found after applying the threshold of:-**

**antecedent**  **consequent**

(FleschReadingEase) (correct\_spell\_ratio)

**Metrics :-**

**antecedent support :-** 0.541

**consequent support** 0.539

**confidence :-**0.604436

**Support :-** 0.327

**Confidence:-** 0.604436

**Lift :-** 1.121403

**Leverage :-** 0.035401

**Conviction:-** 1.165425

As Flesch reading ease test measures the readability of a text and correct spell ratio is associated with readablity thus the rule higher the fleschReadingEase the higher the Correct Spell Ratio is affirmed by the rule and though the support for the rule is less the confidence is high

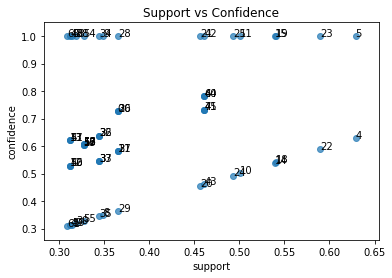
and conviction is high which means the consequent is dependent on antecedent

**Less Intresting Rule :-**

**antecedent**  **consequent**

review stars correct\_Spell\_ratio

The Review stars and correct\_spell ratio have Conviction = Infinity and Leverage = 0 and the Lift = 1 when Lift is 1 it shows the values are independnt when leverage is zero then it means independt thus we can see that the rule is not significant and not a intresting association rule



**Graph of the association rules with Scatter plot of Confidence and Support**

Got a broader understanding of leverage and conviction and how these values are useful in determining the association rules like Leverage computes the difference between the observed frequency of A and C appearing together and the frequency that would be expected if A and C were independent. An leverage value of 0 indicates independence and conviction, A high conviction value means that the consequent is highly depending on the antecedent. The understanding of these metrics to see which rules are useful has been very insightful in the excerise.