Student Name: Zeyan Liang

Student Number: 8392063

I understand the importance of professional integrity in my education and future career in engineering or computer science. I hereby certify that I have done and will do all work on this examination entirely by myself, without outside assistance or the use of unauthorized information sources. Furthermore, I will not provide assistance to others.

Your name (as signature): 

**I agree.**

**Part A: Data preprocessing [10 marks]**

***1. Explain how you would transform the three different data types prior to data mining. (6)***

*namely nominal, binary and numerical data.*

For namely **nominal data**, in this case, the name of country is the nominal data. The number of countries is just three. Therefore, using frequency or response rate is not work for this small dataset. **Label encoder** or **one hot-encoding** are the best choice. The labels are always between 0 and N-classes-1.

For **binary data**, in this case, close to the beach or not close to the beach are equally important.

In symmetric binary, we did the convention: assign 1 to answer “Y” and 0 to answer “N”.

For **Numerical data**, we can change number of stars to a ranked list, the number of restaurants into bins (for example 0-3 3-6…). We also can change the price per individual to bins (for example 0-1000 1000-2000 2000-3000…)

***2. Explain how you would handle noise. (2)***

Binning: Binning methods smooth a sorted data value by consulting the values around it.

Regression: Here data can be smoothed by fitting the data to a function.

***3. Explain how you would handle missing values. (2)***

Use a global constant to fill in for missing values (for example “unknow” or “N/A”).

Use attribute mean or Use attribute mean for all samples belonging to the same class.

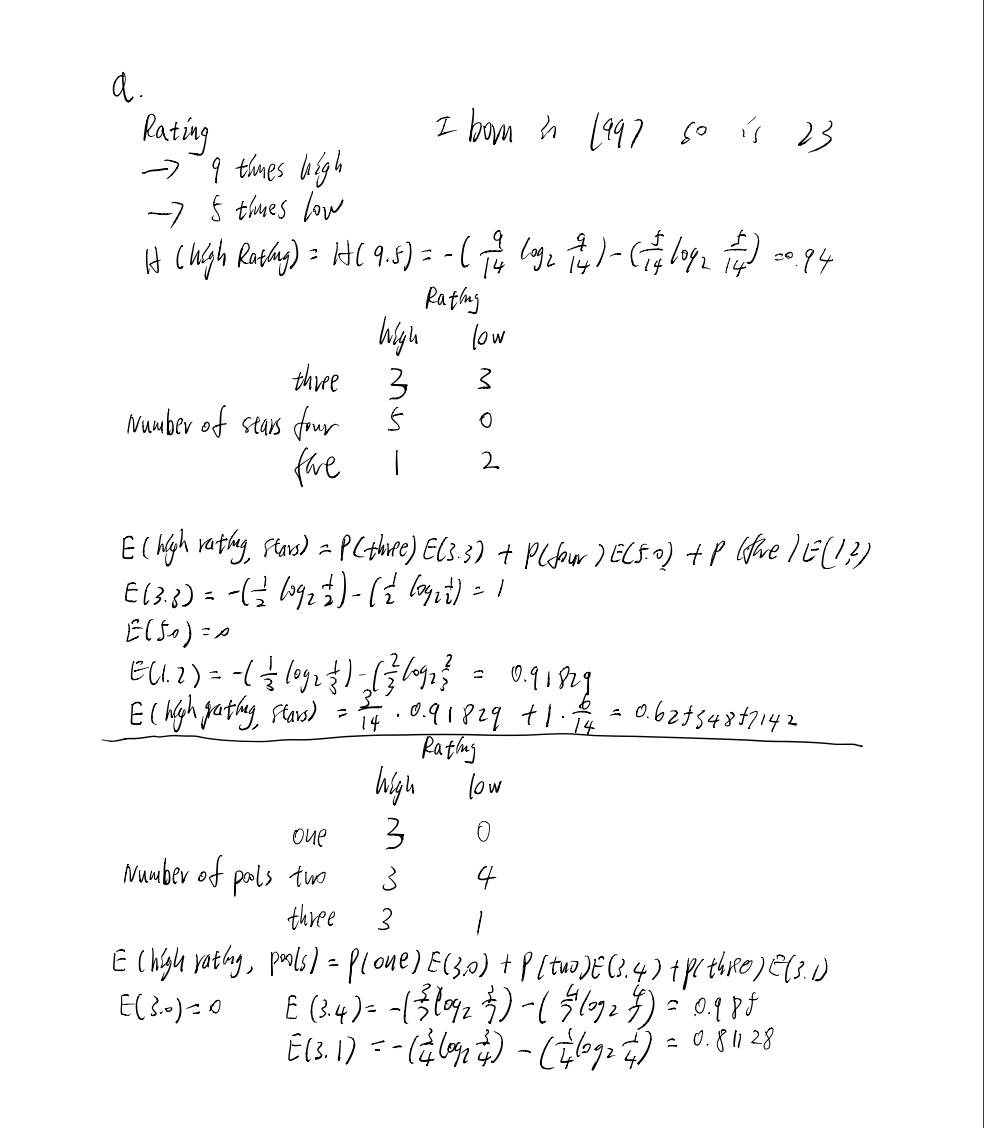
Use a data mining algorithm to predict the most probable value.

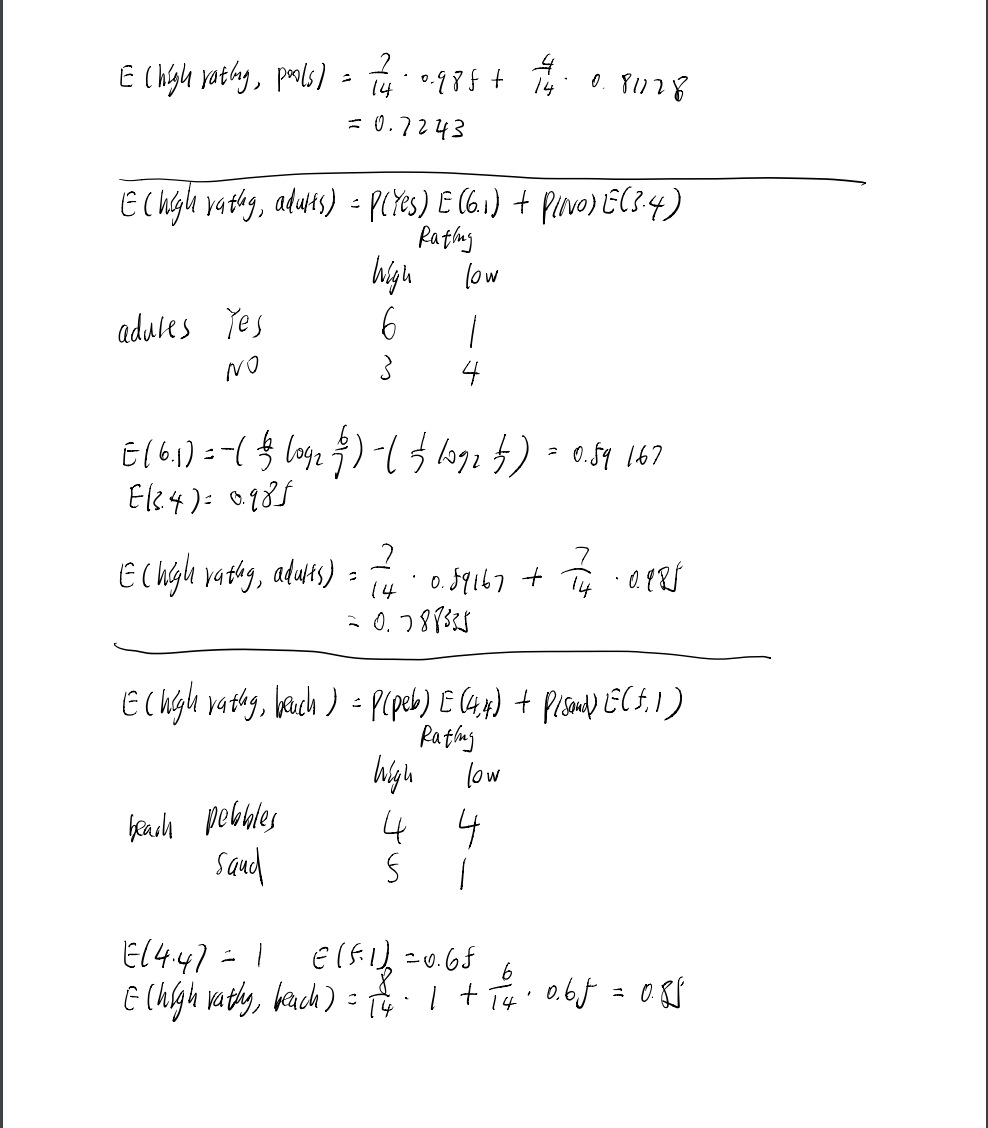
Ask the information from customer or the restaurant, restaurant receipts, consumption records.

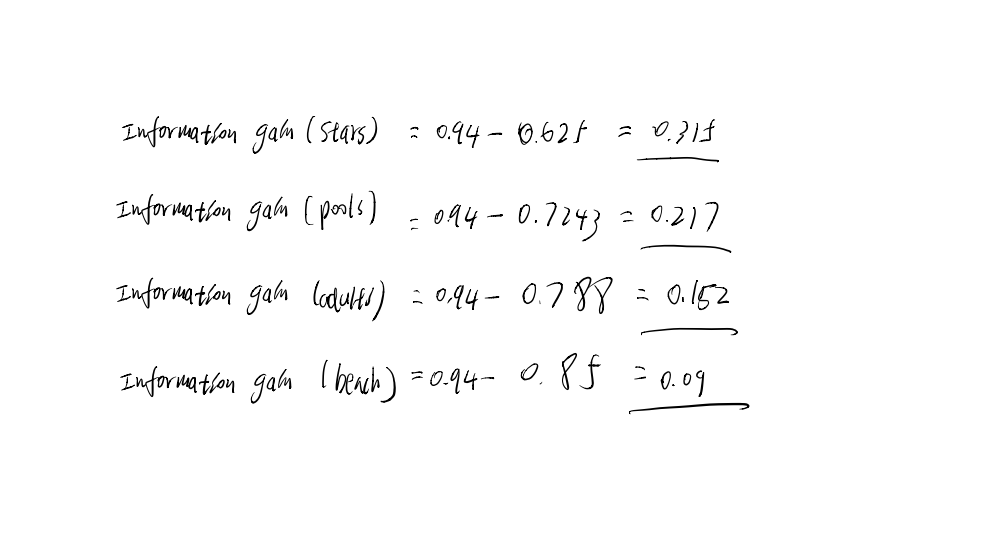
**Part B: Classification [25 marks]**

1.a

***Show how the ID3 decision tree algorithm would construct a model against this data, using Rows 1 to 14 as training set. That is, your answer should detail the steps followed by a decision tree using Information Gain.***







1.b

***Illustrate the steps that a Boosting ensemble, consisting of k-nearest neighbor classifiers, will follow to build a model against this data.***

Firstly, we can choose classiﬁers based upon their individual performance using the out-of-sample accuracy. The selected classiﬁers are then combined sequentially starting from the best model and assessed for collective performance on a validation data set.

We can use benchmark data sets with their original and some added non-informative features for the evaluation of our method.

In this sample data, the comparisons on benchmark classiﬁcation problems and simulated data sets reveal that the proposed ensemble gives better classiﬁcation performance than the usual k-nearest neighbor classifiers and its ensembles and performs comparable to random forest and support vector machines.

2.a

***Illustrate how, and why, class imbalance may lead to low predictive accuracy against the minority class.***

The minority class is harder to predict because there are few examples of this class. This means it is more challenging for a model to learn the characteristics of examples from this class, and to differentiate examples from this class from the majority class.

For example, in this sample data, low rating is the minority class, high rating is the majority class. The abundance of examples from the majority class can swamp the minority class. Most machine learning algorithms for classification predictive models are designed and demonstrated on problems that assume an equal distribution of classes. This means that an imbalance of a model may focus on learning the characteristics of the abundant observations only, neglecting the examples from the minority class that is, in fact, of more interest and whose predictions are more valuable.

2.b

***Show how the SMOTE algorithm may be used to address the class imbalance problem, with reference to the sample data.***

1. Randomly select n samples of minority class (low rating class)

Number of restaurants: 2 Year opend” 1995 rating: low

2.Then find m number nearest samples from the random selected sample.

Number of restaurants: 1 Year opend” 1990 rating: low

Number of restaurants: 3 Year opend” 1998 rating: low (student number end with 3)

3. Then select any one of the m nearest samples

Number of restaurants: 3 Year opend” 1998 rating: low

4. Choose one of these two points, this is the new data sample

We can repeat step 4 to produce more data points in order to make this data set more balance than before.

**Part C: Cluster Analysis [19 marks]**

***1. Explain how, and why, you would decide to use a density-based algorithm such as DBSCAN to cluster this data. Your answer should discuss to the pros and cons of density-based algorithms, with reference to the data as listed in the table.***

Firstly, DBSCAN define clusters as the largest set of points connected by density can divide regions with sufficiently high density into clusters, and can find clusters of arbitrary shapes in the spatial database of noise.

1.Import the data sample.

2. First determine the radius r and minPoints. Starting from an arbitrary data point that has not been visited, centering on this point, r is the radius of the circle contains the number of points greater than or equal to minPoints, if greater than or equal to minPoints then change the point It is marked as a core point, otherwise it will be marked as a noise point.

3. Repeat step 2. If a noise point exists in a circle with a core point of radius, this point is marked as a boundary point, otherwise it is still a noise point. Repeat step 1 until all points have been visited.

Pros: 1. DBSCAN does not need to know the number of clusters to be formed in advance.

2. DBSCAN can find clusters of any shape.

3. DBSCAN can identify noise points.

4. DBSCAN is not sensitive to the order of samples in the database, that is, the input order of Pattern has little effect on the results. However, for the boundary samples between clusters, the clusters may be detected according to which clusters are detected first, and their attribution may fluctuate.

Cons: 1. DBSCAN does not reflect high-dimensional data well.

2. DBSCAN does not reflect the changing density of the data set well.

3. If the density of the sample set is not uniform and the cluster interval is very different, the cluster quality is poor.

***Suppose that you decide to apply a partitioning-based algorithm, such as k-means, to this data.***

***2. Explain how you would use Silhouette plots to determine the best value of k.***

***Silhouette plots is an evaluation method of clustering effect. In this sample data, assume we use***

***k-means to Cluster data.***

1. Randomly select k points and use them as the center point of the first iteration.

2. Calculate the distance from each set of data to k points and assign each point to k clusters.

3. Update the center point: add the value of each attribute of each point in the k clusters and calculate the average value to obtain a new center point.

4. Iterate many times until the value of the center point no longer changes.

Use Silhouette plots to evaluate the Clustering is good or not.

1. Calculate the average distance ai from sample i to other samples in the same cluster. The smaller ai indicates that the sample i should be clustered into this cluster. Let ai be the intra-cluster dissimilarity of sample i. The mean a i of all samples in cluster C is called the cluster dissimilarity of cluster C.

2. Calculate the average distance bij of all samples from sample i to some other cluster Cj, which is called the dissimilarity between sample i and cluster Cj. Defined as the dissimilarity between clusters of sample i: bi = min {bi1, bi2, ..., bik} The larger the bi, the less the sample i belongs to other clusters.

3. According to the intra-cluster dissimilarity a i of sample i and the inter-cluster dissimilarity b i, define the contour coefficient of sample i.

4. When si is **close to 1**, it means that the clustering of sample i is reasonable;

     si is **close to -1**, indicating that sample i should be classified into another cluster;

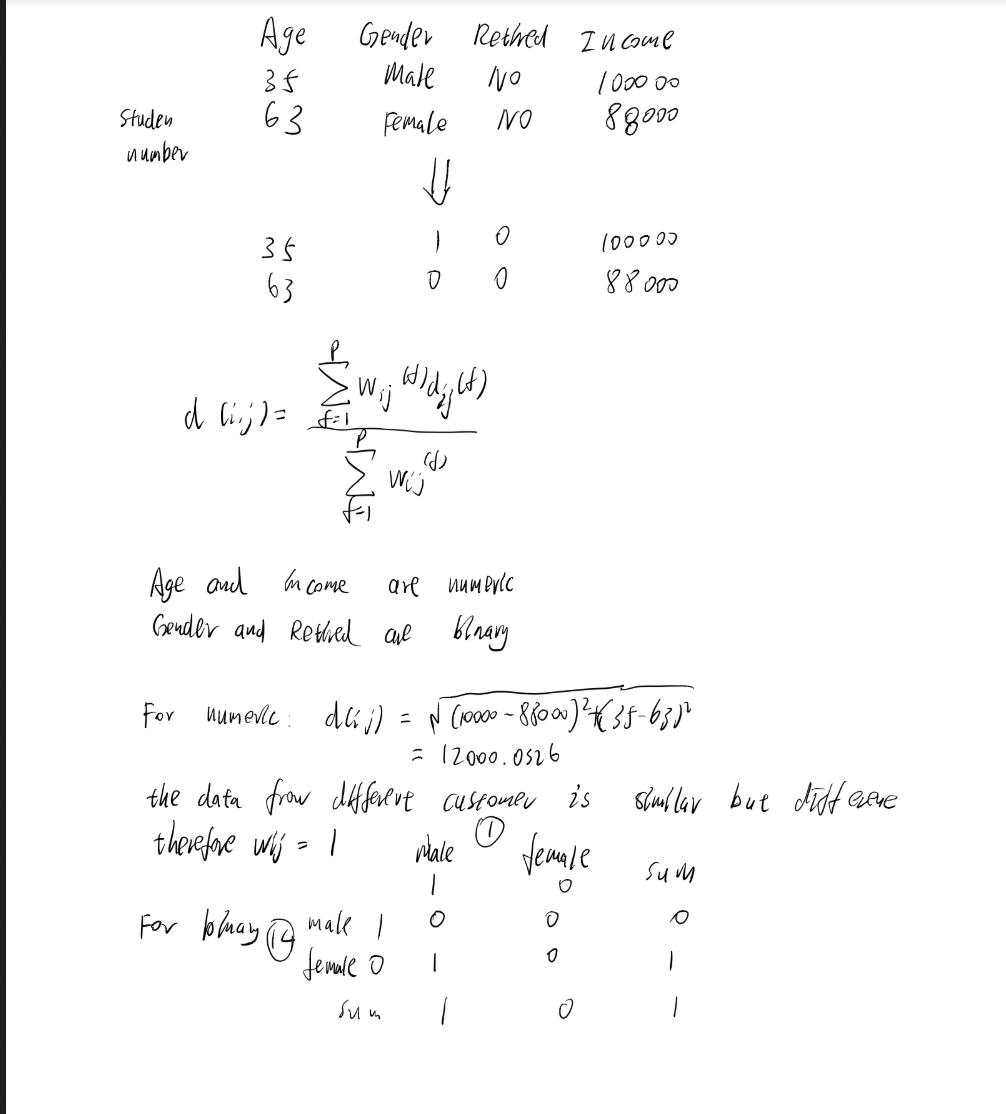
    If si is **approximately 0**, it means that sample i is on the boundary of two clusters.

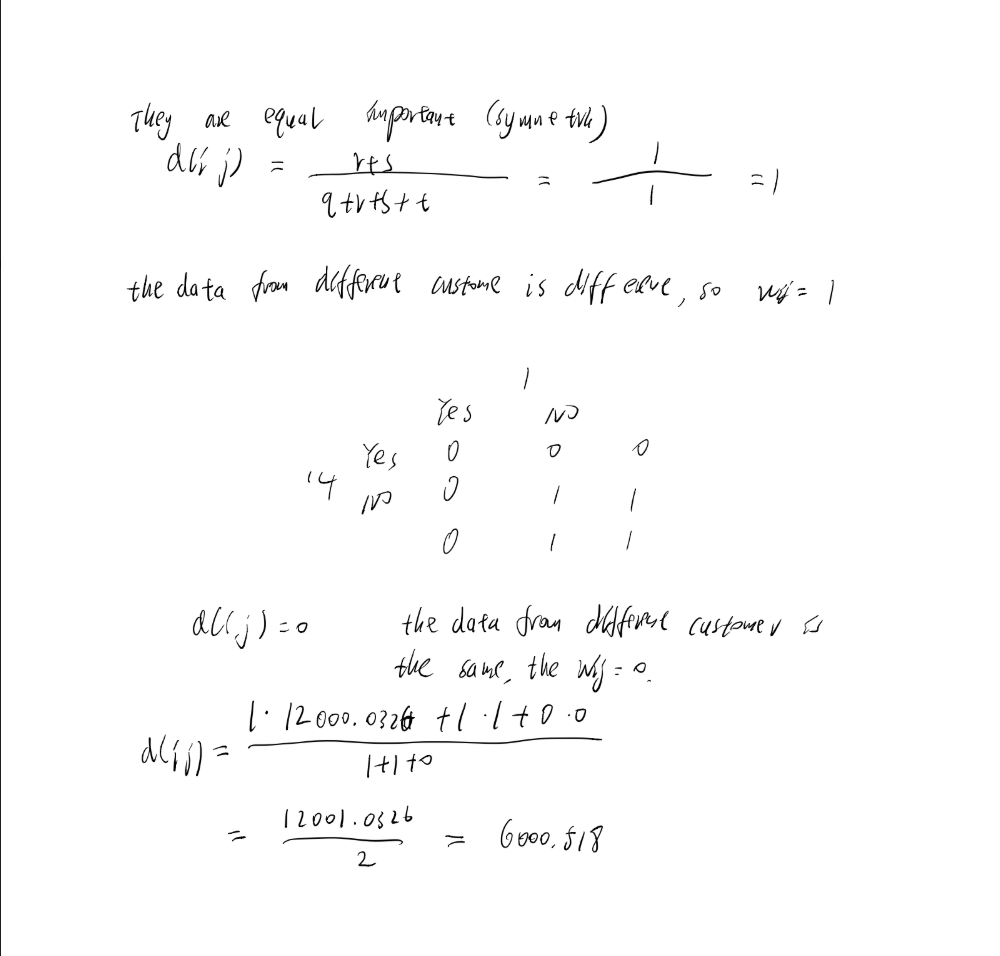
***3. First, insert your own value for the Age attribute, that corresponds to the last two digits of your student number. That is, if you student number is 30009834, you should enter an Age of 34.***

Age Gender Retired Income

35 Male No 100000

63 Female No 88,000

Next, calculate the distance between Customer 1 and Customer 12.



***Part D: Outlier Detection [6 marks]***

***1. Explain what a global outlier is and give your own example of a global outlier that may be presence in the Opération Soleil data. (3)***

A data point is considered a global outlier if its value is far outside the entirety of the data set in which it is found.

This is an example

Number of Stars Number of restaurants Beach Price per individual Country Rating

* 3 Y 800000 Mexico High

In this case,

***2. Explain what a collective outlier is and give your own example of a collective outlier that may be presence in the Opération Soleil data.***

A subset of data points within a data set is considered anomalous if those values as a collection deviate significantly from the entire data set, but the values of the individual data points are not themselves anomalous in either a contextual or global sense. In time series data, one way this can manifest is as a normal peaks and valleys occurring outside of a time frame when that seasonal sequence is normal or as a combination of time series that is in an outlier state as a group.

This is example

Age Gender Retired Income

32 Male Yes 99960

29 Male Yes 99920

The average age of retired person is greater than 55 years old.

32 and 29 are not reasonable in this sample data. But some special occupations do retire early and have high pensions, such as athletes.