Assignment 1

Name: Shounak Rangwala

NetID: snr85

Q2)

1. Descriptive analytics: this uses the past recorded data to get patterns and make observations on it that could potentially be used in future predictions. Most businesses generally use this form of analytics because they have the data ready to perform such analysis on and they need to figure out the trends in their business over the last few years

Example : A company trying to see how their new strategies are improving their revenue would hold a descriptive analysis of its income and expenditure over the last few quarters and understand the patterns that affect the revenue. Clustering or feature-extraction can be used here.

Modelling technique: Unsupervised learning can be used because we are not sure what concept we are going to be needing. This modelling technique helps to define an attribute of the data that can be used to provide insight into the patterns in the data.

1. Predictive analytics: After determining the patterns in the historical data obtained from descriptive analysis, we can predict the outcomes of these same in the future. This is used for setting future goals in businesses and decide on how successful a business strategy can be in the future.

Example: Car manufacturers would do a predictive analysis before launching a new model so they have a benchmark of many sales can be anticipated to happen in the first 2 months.

Modelling technique: We can use supervised learning to focus on some parameters to monitor in the future. This helps in forecasting.

1. Prescriptive analytics: After predictive analytics, we have an idea of what the future holds and so we can come up with solutions that have a high probability of ensuring that result. Prescriptive analytics analyses different solutions that the business can use to get the results desired from the predictive analytics.

Example: Car manufacturers may look to find which sales offer can be most desirable to different types of customers so that their new model can meet the desired/predicted result.

Modelling technique: We can use a combination of unsupervised and supervised learning.We can identify different solutions using unsupervised learning and then from supervised learning we can learn how these solutions can be made better by identifying important aspects of it.

Q3)

Concept : Self-Driving car prediction system. Based on the image captured by the front camera of the car, the system has to predict a steer value (in the range of -1 to 1). Negative steer value will steer the car to the left and positive value will steer it to right. Magnitude will decide how sharp the turn is.

Instance: The data from the database which is used to train the model is the instance.

Attributes: The data instance fed into the model to train it is consists of a image and the steer float value. During testing the predicted attributed will be the steer value.

Concept:

This is the thing we want the model to learn and give the output/prediction of. In the case above, we want the self-driving car to predict a steer value based on the input.

Instance:

Each data example that we give the model to learn is called the instance. It has several attributes in it. In the case it is the data containing the image and the steer value.

Attribute:

Each aspect of the instance that defines the instance. All the instances in the database have the same number of attributes and the collective values of these attributes makes each instance unique.

Q4)

Missing values: Sometimes in the database there will be some attributes that do not have a value entered into them.

Why they occur: These usually occur because the values of the attributes may not have been known during the recording of the data. Sometimes the information may have been purposefully left blank and while conducting machine learning on these datasets, its, important to have the domain expert find out the significance of missing values.

Example: While filling out the survey many participants prefer anonymity and may hold back from disclosing their age, name , gender etc. In that case many instances (survey responses) will have missing values in the attributes of name, age and gender.

Incorrect values: These values are outliers or rogue values that were entered wrong by mistake or by purpose. These values can potentially skew the models ability to learn.

Why they occur: There are many reasons that one finds incorrect values in the database. Sometimes the data becomes outdated and things such as the address and phone numbers have changed and may be incorrect in the database. Sometimes typographical errors can cause this too

Example:some participants in the survey write a wrong answer either purposefully or by mistake.

Q5)

False positive: a result that shows a positive outcome of a test when the outcome is negative or not positive is called false positive

False negative: a result of a test that shows a negative outcome when the actual is positive or not negative.

Q6)

Training set: part of the dataset that is used to train the model in supervised learning so that it could learn to predict a certain attribute. The training set can be passed through the model as many times as you want till you are satisfied with the learning.

Test set: part of the same dataset that the training set is part of. This is used to test how well the model is trained by passing the data and comparing the predicted value of the attribute to the actual value of the same.

Q7) A set of instances is complete when all the different cases including borderline cases are present in the dataset. All the rules formulated by learning from the dataset summarize all the instances present in that dataset. There may be similarities in the instances

A set of instances is representative because it represents all the qualities of a complete set but it is smaller in size than a complete set. Representative sample is a subset of the complete sample. Machine learning models often use this sample because its smaller size makes processing faster and it correctly represents all of the data.

Q8)

Data quality refers to how good the data is in the sense that can it be reliably used for the modelling application we have in mind. This data should be able to give businesses actionable results.

Poor data quality happens because of incorrect data, missing data or sometimes accumulation of data from different sources having different formats so that the whole dataset cannot be used in the same model.

2 examples of this are:

1. A survey taken among student about returning back to campus after COVID-19 shows irregularities in the way of half-filled forms, incorrect age values, incorrect majors etc. This data quality is poor because no statistical analysis done on this data will be able to give accurate results/outcomes.

The way to control this is through data quality management and enforcing rules will data is being entered in the survey so that the participant enters a response that can be comprehended by the model and the data is now ready for analysis.

1. In the self-driving car example, the data was of poor quality because of the format in which it was being fed into the model. Values that the model considered were numbers were in fact strings. This makes the data quality poor even though the data is correct.

The way to correct this is by proper data preprocessing so that data is converted to the proper format.

Q9)

1. Since data mining techniques have the potential to identify personal sensitive information from the given data, how are people going to be convinced to give their data while still trying to preserve anonymity?

Ans: Sometimes because they want anonymity, people will give false/incorrect personal data so that after data mining they cannot be identified from the records. Sometimes they may leave the data entry blank. If we try to enforce the no-blank data policy, we may end up losing the participants all together and fall short of required data. One way would be to carry out the data mining, get the results and not disclose the data to the public if you feel that there is a way that the personal information of the people will be disclosed. If you have to disclose some data, maybe disclose a random sample that will provide no idea about the participants.

Sometimes it is necessary to decide beforehand what data is crucial for the study and ask only that from the participants. Domain experts can be consulted about which attributes are needed for the learning process and which are not. Also test runs on the data can be carried out by third parties to find if the personal information is at risk of being revealed.

1. Before beginning data mining, we are supposed to clean the data, but if the denormalized data has missing values, unbalanced values or sparse values, and we manually have to go over all the cases, then why do we need to do data mining?

Ans: Data mining is meant to substitute human labour and browse through reams of data and find patterns that would take lot of time and human resources. However, data purification is another process all together, we are not trying to find any insight in the data, we are just trying to eliminate any and all outliers so that the main process of data mining is fluid. Data experts may be consulted about the potential significance of unbalanced, missing or incorrect data but this process is mostly independent of data mining process. Any observations made by these experts can be added to the findings of the actual analysis. Both these processes happen in parallel and save time and labour.

1. Is unsupervised learning the only way of removing bias from the data mining process?

Ans: It would seem that way because if at any point of the machine learning process, humans determine parameters, concepts, goals etc, then there will be inherent bias. This bias may be in data selection, allowing certain outliers and not allowing others etc. Supervised learning often produces output that the supervisor needs and has use for. This may be counted as bias. However unsupervised learning has a limited number of uses. It will tell us of patterns and trends we do not know but it cannot be used in businesses beyond this point. Businesses need actual expected goals and determined numbers, this can only be guaranteed by supervised learning. Hence, we can only try to minimize the bias but cannot remove it completely.

Q11) Outliers: I looked for null values in the dataset by appending the dataframe and removing duplicates. I found that there are no null values in any column in the dataset.

If there would have been null values, I would have either removed those entries (if the number of null values was small) or I would have removed the entire column (if the number of null values were significantly large). If these values were numerical in type, I may have changed their values to the mean of the column values. It requires a domain expert to have the final say on what to do with null values: should we disregard those instances, or should we add values to those blanks.

In the case of diabetes\_s.csv, we observe that the values of the ‘age’ attribute of several instances are incorrect/absurd. These values range from -102 to 600. We have to find the number of such outliers in the data for which we did the below snippet assuming the age range from 0 years to 100 years.

A screenshot of a cell phone

Description automatically generated

The number of such outliers are 669 in number while the total number of instances are 770. Obviously, this is a significant number of outliers.

However, we do not wish to lose the data present in other fields of this dataset because that data might be of good quality and be useful in further analysis. **Thus we must disregard this column all together from the dataset.**

So the final outcome will be like this:

A screenshot of a cell phone

Description automatically generated

Q12)

The following way:

A screenshot of a cell phone

Description automatically generated

Q13)

A screenshot of a cell phone

Description automatically generated

Q14) The average, standard deviation, minimum and maximum values of the BMI column

A screenshot of a cell phone

Description automatically generated

Q15) 1) Gini tree

A picture containing screenshot

Description automatically generated

2) Entropy tree

A screenshot of a cell phone

Description automatically generated

Q16) OneR stands for one rule. This means you pass through all the predictors in the dataset and counts the frequency by which the value of the predictor produces the desired target. Whichever predictor has the least total error after adding up the frequencies of all the values of the predictor: that predictor will be made the one rule for the dataset.

The oneR method is better for marking the baseline for machine learning models because it is simply stating that the more frequent predictor producing the least error in the output is probably the desired rule. This is generally the case and is simpler for us to understand the classification.