Your name: Course code (e.g. COMP309, COMP440):

**COMP 309 | Machine Learning Tools and Techniques**

Assignment 2:

**2.1 Core:**

A screenshot of a map

Description automatically generated

**Linear Regression:**

A screenshot of a cell phone

Description automatically generatedA picture containing indoor, bird

Description automatically generated

The first technique that I used on the dataset was Linear regression, for this I used a variety of filters including remove attributes, replace missing values, Class assigner to ensure that the regression was using hospitalized\_covid\_confirmed\_patients as its predictor and attribute selector to get a better subset of predictors. I found that 10 fold cross validation produced the best results which are shown above with a Correlation Coefficient of 0.97 showing a strong positive linear relationship.

**Naïve Bayes:**

A picture containing receipt, text

Description automatically generated

The second technique that I used on the dataset was Naïve Bayes a technique from the Bayesians tribe. For this classifier I again used a variety of filters including remove attributes, replace missing values, numeric to nominal to change the available beds feature from numeric to nominal and class assigner to assign the available beds feature as the class. I again used 10 fold cross validation for training and testing. The results of this produced a correctly classified instances percentage of 86.05% where the classifier only incorrectly classified 18 instances. The kappa statistic is not the best at 0.61 indicating that predictions are not completely random but somewhat. The precision of the classifier was pretty good sitting at an average of 0.87 where it incorrectly classified more instances of b than a. the ROC area has a value of 0.91 which is really good and indicates that the weighted accuracy of the model is good.

There is an initial obvious difference between the aspects two techniques which is that regression uses a numerical output variable which is continuous in its prediction whereas classification uses a categorical variable as output which is discrete. This is shown in the predictions above where the linear regression model is predicting the number of hospitalized covid confirmed patients which is a continuous variable as it can be any range of numbers from 0 upwards, in comparison the Naïve Bayes model is only classifying its output form 2 possible outcomes which means that the output is discrete.

Through the business understanding of this dataset an initial question noted in the assignment was “is there any evidence of population density affecting the number of cases in a given area?. This question is interesting because the initial dataset does not have enough information to be able to accurately answer that question and therefore more information is needed.

Other questions that could be formed to be asked of the data could be;

does the rate of people with other serious illnesses/ injuries effect the availability of beds in a given hospital?

Or

Is there evidence that the availability of accommodation to homeless people is impacting the number of cases in a given hospital?

**2.2:**

**Feature importance to COVID-19 cases:**

The question I have selected out of the three business understanding questions is;

Is there evidence that the availability of accommodation to homeless people is impacting the number of cases in a given hospital?

I believe this is an interesting question as one would assume that someone who is homeless or living on the street would be more susceptible to the virus that another individual who was not. You would assume that someone in that position would likely have more chances of coming into contact with the virus and therefore if they do not have access to accommodation then they would have more time to be exposed and therefore more chance.

Another publicly available dataset the could be helpful to this question could be the San Benito homeless persons dataset in the same area as the hospital dataset which has already been used. This data was gathered for obvious reasons to keep track of the number of homeless people who both with and without accommodation as well as the availability of accommodation for homeless people, this data is also being collected in order to not only keep track but to monitor the spread of the virus and try to identify different factors which could be causing increases or decreases in new cases every day.

A close up of a map

Description automatically generated

I used a series of different techniques in WEKA knowledgeflow in order to merge the hospitals and homeless datasets into one complete dataset. I used remove attributes to get rid of unnecessary attributes and replace missing values. I had to use remove range for the datasets in order to get the dates all lined up evenly so the join could be made smoothly between all the datasets, the homeless dataset had a lot of missing dates which meant that I had to remove those dates from the other data sets in order for it to be able to join correctly. After the datasets were joined together I once more removed the additional date attributes from each of the new datasets that were not needed anymore which gave me my final dataset. This dataset contains all the information from both datasets combined into one larger set with more information and a larger scope.

A close up of a map

Description automatically generated

Shown above is my feature selection pipeline executed on my merged dataset, specifically this attribute selector was executed with hospitalized\_covid\_confirmed\_patients being the class that the attribute selection based its selection on, I used CfsSubsetEval. This resulted in the following attributes included in the model:

* icu\_covid\_confirmed\_patients (numeric)
* beds\_available (binary)
* hospitalized\_covid\_patients (numeric)

Unsure how I feel about this due to the fact that it removes all of the attributes I merged onto the original dataset.

After running linear regression I got the following results:

A picture containing bird

Description automatically generatedA picture containing indoor, bird

Description automatically generated

As seen in the regression model it appears that hospitalized\_covid\_patients has the highest weight in the regression model.

I didn’t really think that this model really considered the question I chose above related to accommodation available to homeless people and cases in a given hospital. So I decided to do my own attribute selection selecting attributes I thought were most suitable to the question.

The resulting attributes left over were:

* rooms, rooms occupied, hospitalized covid patients, icu suspected covid patients, icu beds available and beds available.

The results were:

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated