My project

Introduction

I tried out some data which comes from:
https://healthdata.gov/dataset/community-health-status-indicators-chsi-combat-obesity-heart-disease-and-cancer
They provide csv files with information about the citizens from many cities around USA regarding their health status in a general context. Data exists from roughly 3200 cities with more than 400 attributes. There were no metrics provided for individual patients but rather some statistical mean values and indicator values.

Previous Work

I have not had time to comlete this yet.

Work in progress

I used a decision tree implementation from assignment 1 to work with the data to try to find out if or how lung cancer deaths relates to other attributes.

Dropping data

Since there are a lot of attributes reported for each city I decided to drop most of them and focus on a few attributes.
There are different reason to why I drop some attributes and here are the most important ones:

- For the scope of this assignment I had to chose fewer data attributes in order be able to finish it.
- Some of the attributes are less interesting since the data is incorrect or missing. For example some values that should be positive are negative for an absolute majority of the cities which leads to fewer data points. Hence those attributes are dropped.
- Other data points all had the same values, for example carbon monoxide air quality indicator, which in all cases were set to 1. So there is no interesting variety of information there.

I ended up using data from 9 attributes and 1 class.

Chosen data attributes and class

The following attributes were used:
Ozone_Ind,Particulate_Matter_Ind,RHI_Brst_Cancer_Ind,RHI_Col_Cancer_Ind,RHI_CHD_Ind,RHI_MVA_Ind,RHI_Stroke_Ind,Toxic_Chem,Population_Density

The following class was used:
Lung_Cancer

Ozone_Ind – an air quality indicator regarding ozone levels. Unfortunately they only had the values ’1’ and ’2’ and not a true metric. But I decided to pick it anyway.
Particulate_Matter_Ind – an air quality indicator regarding particles. These values were also limited to 1 and 2. But I decided to pick it too.
RHI_Brst_Cancer_Ind – an indicaor value for breast cancer.
RHI_Col_Cancer_Ind – an indicator value for colon cancer.
RHI_CHD_Ind – an indicator value for coronary heart disease.
RHI_MVA_Ind – an indicator value for motor vehicle injuries.
RHI_Stroke_Ind – an indicator value for stroke.
Toxic_Chem – a metric for how many tonnes of toxic chemicals that have polluted the nature.
Population_Density – a metric for population density.
Lung_Cancer – a value that indicates the amount of lung patients that died of lung cancer.

Data rows that were dropped in runtime

The amount of rows, about 3200, allowed me to make some choices and still keep significant parts of the data. For example I decided that all rows containing negative values should be dropped since negative values means the data is missing or irrelevant.
The end result of this led to a reduction from approximately 3200 rows to about 2200 rows and since that’s still a lot of cities I decided to keep this data reduction choice.

Data conversion

Since I used my own implementation from assignment 1 it was limited to use attributes with multiple choices rather than continuous values. For this reason I decided to convert the data points from the columns Toxic_Chem, Population_Density and Lung_Cancer.
The conversion was chosen as follows (the values chosen is still work in progress):
Toxic_Chem:

Low	below 7000
Medium	7000 to 40000
High	above 40000

Population_Density:

Low	below 40
Medium	40 to 100
High	above 100

Lung_Cancer:

Low	below 60
High	above 60

Results so far

I trained 80% of the city rows and used the provided shuffle scheme and dropped the last 20% of the list.
The first split choice of the attributes that was chosen in the decision tree with my implementation was RHI_CHD_Ind. This means that, using my current value thresholds at data conversion above and my implementation of the ID3 algorithm, the most information gain is found using this attribute.
Whether or not this is correct or not is a matter of future work. For example other threshold values might yield a different result.

Future work

I have not yet had time to correlate the trained data to predict if the lung cancer death rate is High or Low which should be fairly straight forward. But since I haven’t implemented i.e. Naïve Bayes classifier I decided to wait until later.