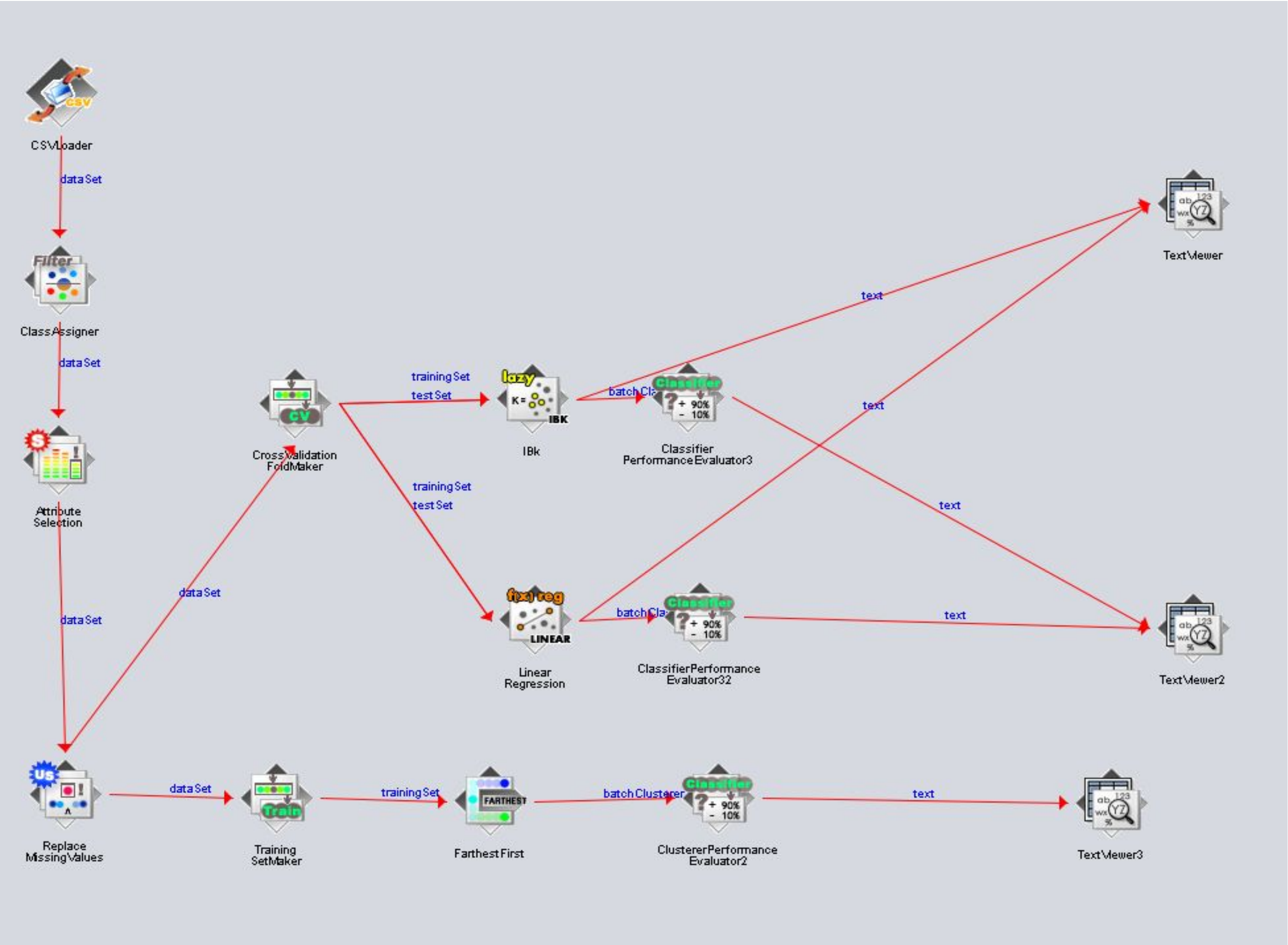
COMP 309 Assignment 2

ID:300432074

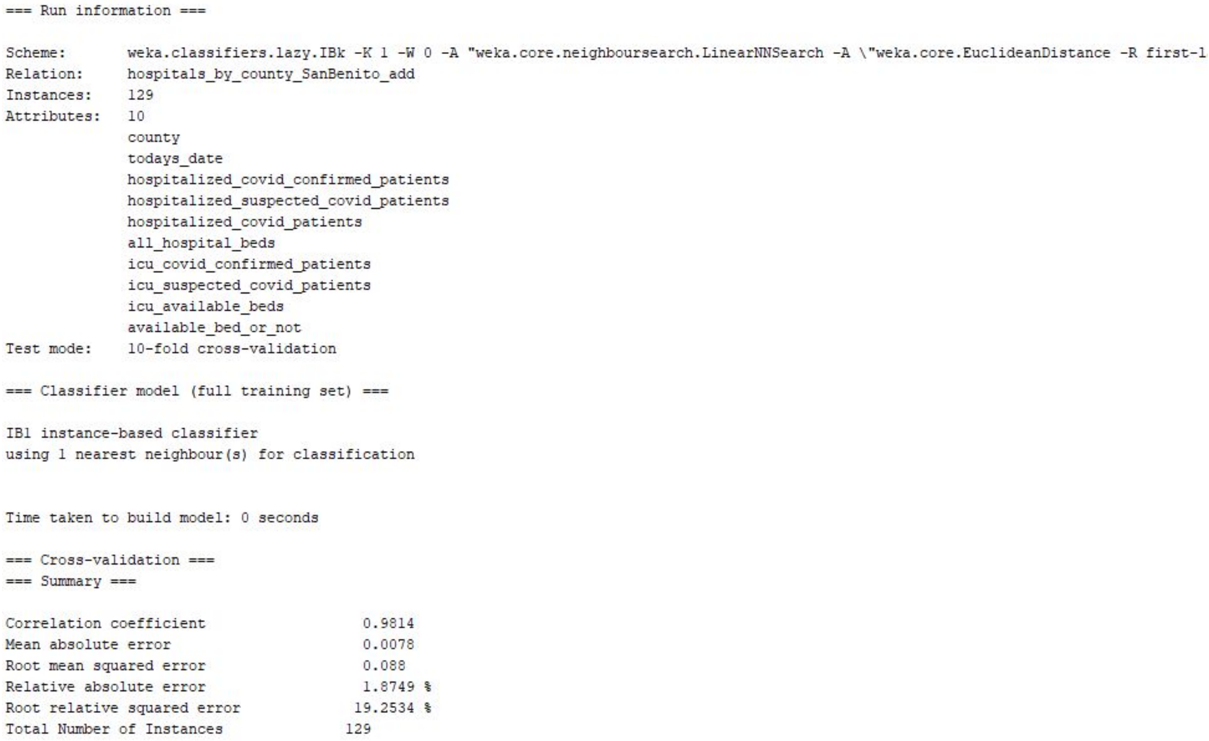
Name:Zhanghao Kong

2.1 Part 1: Core: Pre-processing of COVID-19 cases in a given area [40 marks]

Pipeline:I used IBk , Linear regression and farthest First method.



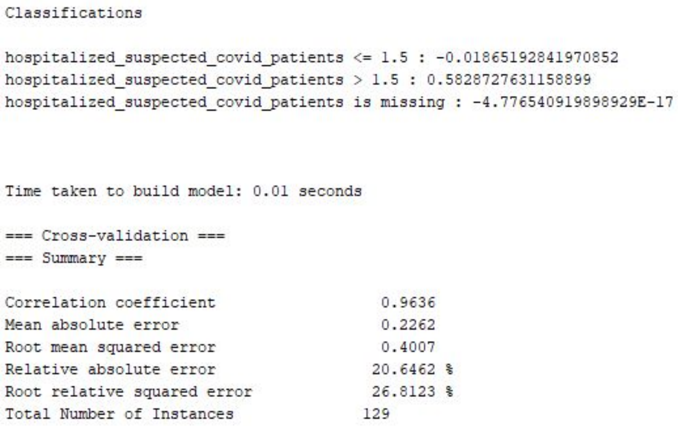
IBk



```
=== Run information ===

Scheme:       weka.classifiers.lazy.IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-l
Relation:     hospitals_by_county_SanBenito_add
Instances:    129
Attributes:   10
              county
              todays_date
              hospitalized_covid_confirmed_patients
              hospitalized_suspected_covid_patients
              hospitalized_covid_patients
              all_hospital_beds
              icu_covid_confirmed_patients
              icu_suspected_covid_patients
              icu_available_beds
              available_bed_or_not
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

IB1 instance-based classifier
using 1 nearest neighbour(s) for classification


Time taken to build model: 0 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.9814
Mean absolute error                      0.0078
Root mean squared error                  0.088 
Relative absolute error                  1.8749 %
Root relative squared error             19.2534 %
Total Number of Instances              129     
```

For doing IBk, because it is a classification method, I made a new attribute called “available_bed_or_not” which records whether the hospital has available beds or not. IBk is very suitable for this purpose.

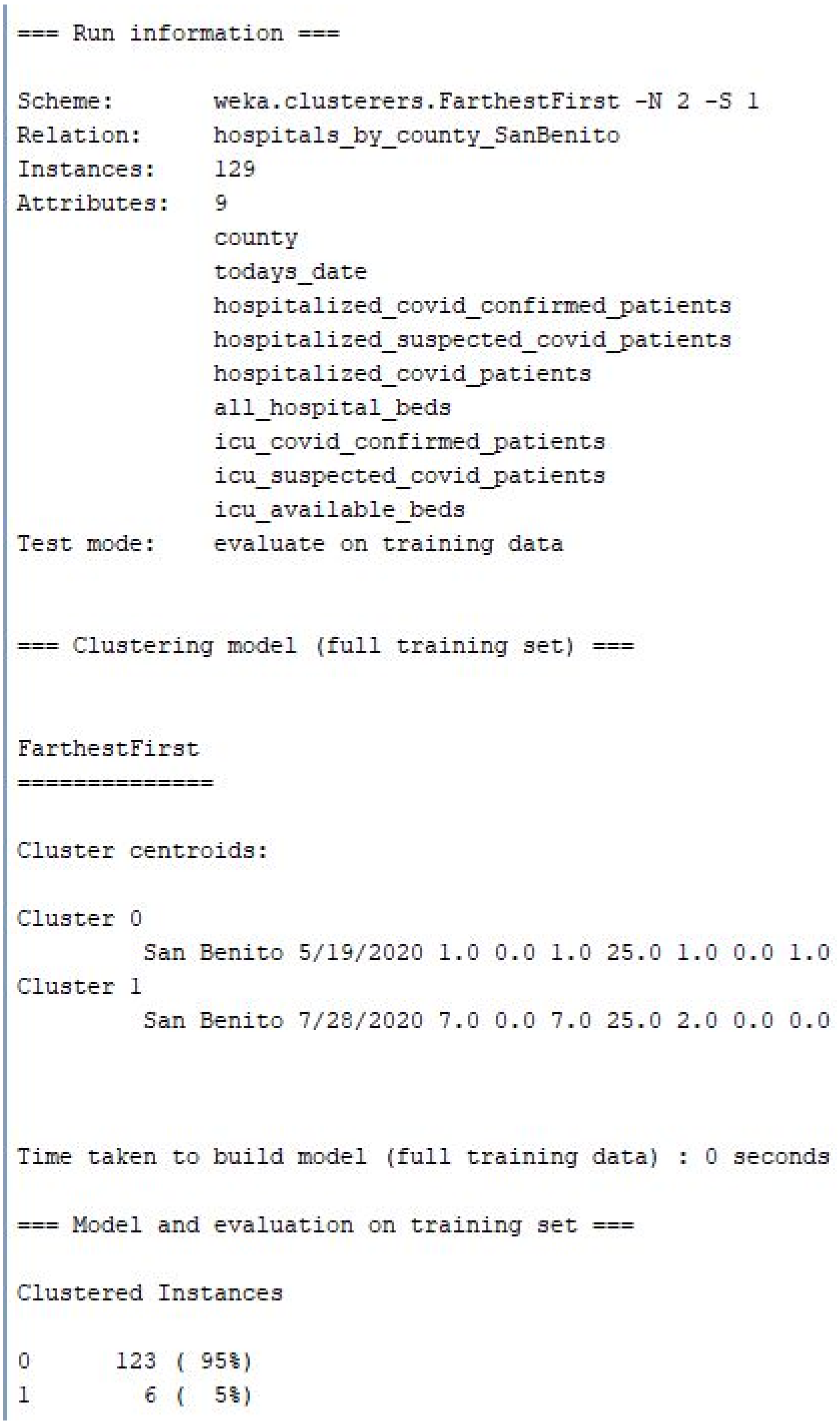
Additive regression



```
Classifications

hospitalized_suspected_covid_patients <= 1.5 : -0.01865192841970852
hospitalized_suspected_covid_patients > 1.5 : 0.5828727631158899
hospitalized_suspected_covid_patients is missing : -4.776540919898929E-17


Time taken to build model: 0.01 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.9636
Mean absolute error                      0.2262
Root mean squared error                  0.4007
Relative absolute error                 20.6462 %
Root relative squared error             26.8123 %
Total Number of Instances              129
```

"hospitalized covid confirmed patients" attribute is the target variable for regression.Because the data value is continuous,so Additive regression can be used. The purpose of our experiment is to predict future trends from known data. This satisfies the purpose of Additive regression.For my data set, it can predict confirmed patients number in future and it also get a results which is very satisfactory.

Farthest First



```
=== Run information ===

Scheme:       weka.clusterers.FarthestFirst -N 2 -S 1
Relation:     hospitals_by_county_SanBenito
Instances:    129
Attributes:   9
              county
              todays_date
              hospitalized_covid_confirmed_patients
              hospitalized_suspected_covid_patients
              hospitalized_covid_patients
              all_hospital_beds
              icu_covid_confirmed_patients
              icu_suspected_covid_patients
              icu_available_beds
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


FarthestFirst
==============

Cluster centroids:

Cluster 0
         San Benito 5/19/2020 1.0 0.0 1.0 25.0 1.0 0.0 1.0
Cluster 1
         San Benito 7/28/2020 7.0 0.0 7.0 25.0 2.0 0.0 0.0



Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      123 ( 95%)
1        6 (  5%)
```

I think this experimental cluster is not applicable. The main reason is that the data have labels. We don't want to divide data into different groups.This deviates from our experimental purpose.

Differences:

clustering differ from classification techniques:

Although both techniques have certain similarities, the difference lies in the fact that classification uses predefined classes in which objects are assigned, while clustering identifies similarities between objects, which it groups according to those characteristics in common and which differentiate them from other.Both result are not suitable to predict increasing patients in the future.
Regression: Because the data value is continuous,so Additive regression can be used. Besides, this method is suitable to predict increasing patients in the future which satisfied the purpose.

business understanding questions

1.is there any evidence of population density affecting the number of cases in a given area?
2.is there any evidence of choosing area position affecting the number of cases ?
3.Does homeless people affect the number of cases?

2.2 Part 2: Completion: Feature importance to COVID-19 cases [40 marks]

The question I chose is number 3:
Does homeless people affect the number of cases?

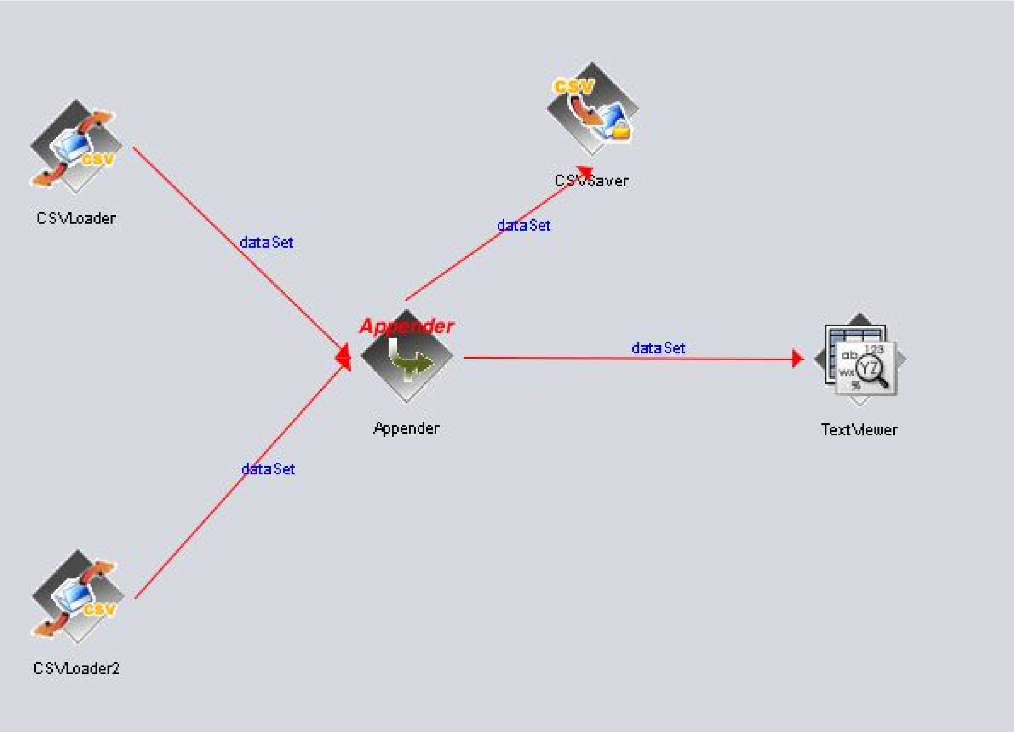
Why it is interesting?
Covid-19 is a highly infectious virus. The arrival of homeless people may become a source of infection. These people may have carried the virus, causing covid-19 to infect San Benito.By studying the relationship between the two, we can know whether there is a link between the increase of patients and the arrival of homeless.

Selected dataset： homeless impact SanBenito.csv

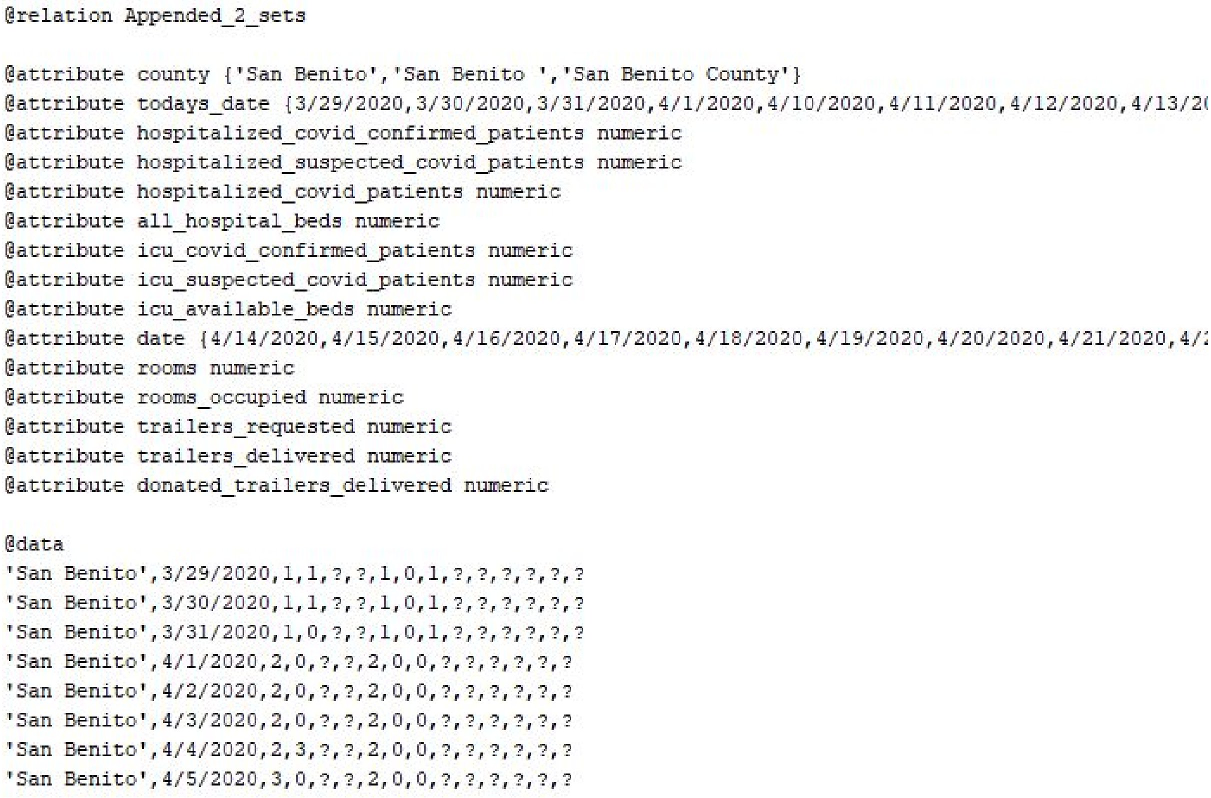
This data set records the time and number of homeless people arriving in the country, which can produce a good comparison with previous data sets. We can compare two datasets to see if there is an impact between the changes in each other.

Data preparation:

Merge pipeline:

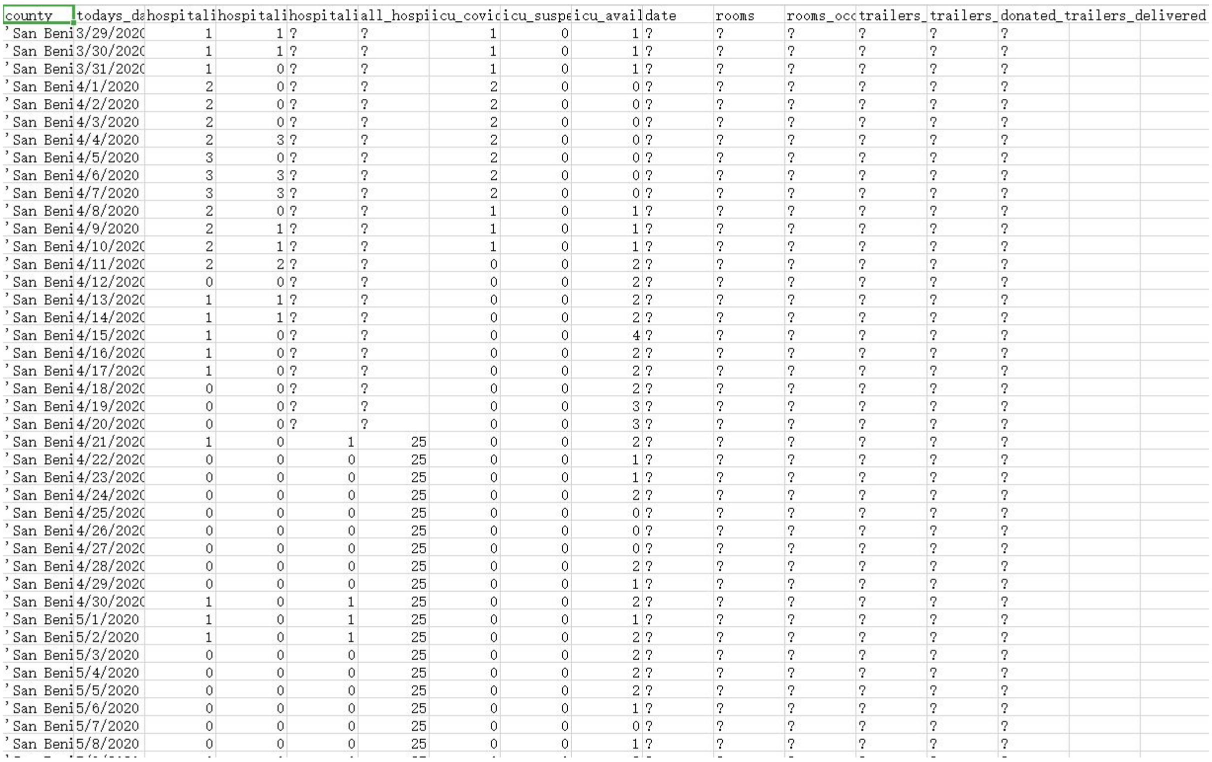


After Merge:



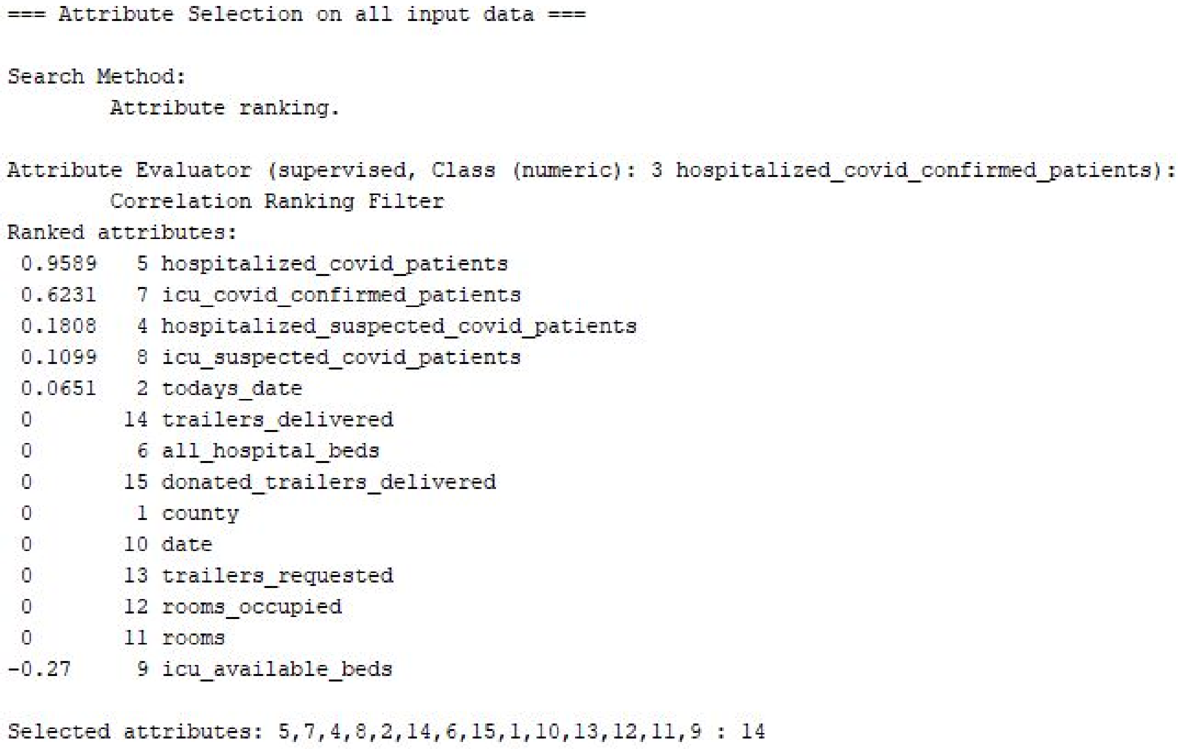
```
@relation Appended_2_sets

@attribute county {'San Benito','San Benito ','San Benito County'}
@attribute todays_date {3/29/2020,3/30/2020,3/31/2020,4/1/2020,4/10/2020,4/11/2020,4/12/2020,4/13/2
@attribute hospitalized_covid_confirmed_patients numeric
@attribute hospitalized_suspected_covid_patients numeric
@attribute hospitalized_covid_patients numeric
@attribute all_hospital_beds numeric
@attribute icu_covid_confirmed_patients numeric
@attribute icu_suspected_covid_patients numeric
@attribute icu_available_beds numeric
@attribute date {4/14/2020,4/15/2020,4/16/2020,4/17/2020,4/18/2020,4/19/2020,4/20/2020,4/21/2020,4/
@attribute rooms numeric
@attribute rooms_occupied numeric
@attribute trailers_requested numeric
@attribute trailers_delivered numeric
@attribute donated_trailers_delivered numeric

@data
'San Benito',3/29/2020,1,1,?,?,1,0,1,?,?,?,?,?,?
'San Benito',3/30/2020,1,1,?,?,1,0,1,?,?,?,?,?,?
'San Benito',3/31/2020,1,0,?,?,1,0,1,?,?,?,?,?,?
'San Benito',4/1/2020,2,0,?,?,2,0,0,?,?,?,?,?,?
'San Benito',4/2/2020,2,0,?,?,2,0,0,?,?,?,?,?,?
'San Benito',4/3/2020,2,0,?,?,2,0,0,?,?,?,?,?,?
'San Benito',4/4/2020,2,3,?,?,2,0,0,?,?,?,?,?,?
'San Benito',4/5/2020,3,0,?,?,2,0,0,?,?,?,?,?,?
```

Merged file:



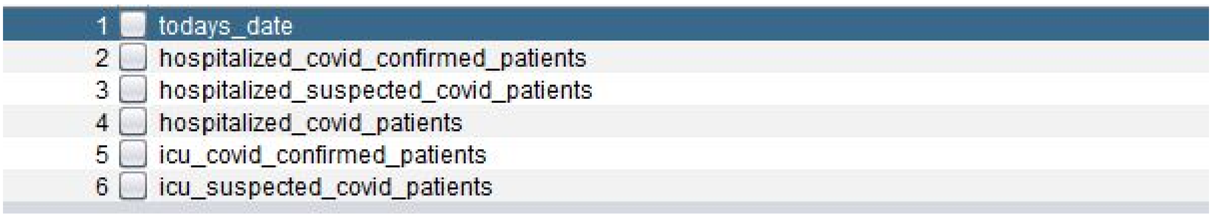
county	todays_da	hospitali	hospitali	hospitali	all_hospi	icu_covi	icu_suspe	icu_avail	date	rooms	rooms_occ	trailers_	trailers_	donated_trailers_delivered
'San Beni	3/29/2020	1	1	?	?	1	0	1	?	?	?	?	?	?
'San Beni	3/30/2020	1	1	?	?	1	0	1	?	?	?	?	?	?
'San Beni	3/31/2020	1	0	?	?	1	0	1	?	?	?	?	?	?
'San Beni	4/1/2020	2	0	?	?	2	0	0	?	?	?	?	?	?
'San Beni	4/2/2020	2	0	?	?	2	0	0	?	?	?	?	?	?
'San Beni	4/3/2020	2	0	?	?	2	0	0	?	?	?	?	?	?
'San Beni	4/4/2020	2	3	?	?	2	0	0	?	?	?	?	?	?
'San Beni	4/5/2020	3	0	?	?	2	0	0	?	?	?	?	?	?
'San Beni	4/6/2020	3	3	?	?	2	0	0	?	?	?	?	?	?
'San Beni	4/7/2020	3	3	?	?	2	0	0	?	?	?	?	?	?
'San Beni	4/8/2020	2	0	?	?	1	0	1	?	?	?	?	?	?
'San Beni	4/9/2020	2	1	?	?	1	0	1	?	?	?	?	?	?
'San Beni	4/10/2020	2	1	?	?	1	0	1	?	?	?	?	?	?
'San Beni	4/11/2020	2	2	?	?	0	0	2	?	?	?	?	?	?
'San Beni	4/12/2020	0	0	?	?	0	0	2	?	?	?	?	?	?
'San Beni	4/13/2020	1	1	?	?	0	0	2	?	?	?	?	?	?
'San Beni	4/14/2020	1	1	?	?	0	0	2	?	?	?	?	?	?
'San Beni	4/15/2020	1	0	?	?	0	0	4	?	?	?	?	?	?
'San Beni	4/16/2020	1	0	?	?	0	0	2	?	?	?	?	?	?
'San Beni	4/17/2020	1	0	?	?	0	0	2	?	?	?	?	?	?
'San Beni	4/18/2020	0	0	?	?	0	0	2	?	?	?	?	?	?
'San Beni	4/19/2020	0	0	?	?	0	0	3	?	?	?	?	?	?
'San Beni	4/20/2020	0	0	?	?	0	0	3	?	?	?	?	?	?
'San Beni	4/21/2020	1	0	1	25	0	0	2	?	?	?	?	?	?
'San Beni	4/22/2020	0	0	0	25	0	0	1	?	?	?	?	?	?
'San Beni	4/23/2020	0	0	0	25	0	0	1	?	?	?	?	?	?
'San Beni	4/24/2020	0	0	0	25	0	0	2	?	?	?	?	?	?
'San Beni	4/25/2020	0	0	0	25	0	0	0	?	?	?	?	?	?
'San Beni	4/26/2020	0	0	0	25	0	0	0	?	?	?	?	?	?
'San Beni	4/27/2020	0	0	0	25	0	0	0	?	?	?	?	?	?
'San Beni	4/28/2020	0	0	0	25	0	0	2	?	?	?	?	?	?
'San Beni	4/29/2020	0	0	0	25	0	0	1	?	?	?	?	?	?
'San Beni	4/30/2020	1	0	1	25	0	0	2	?	?	?	?	?	?
'San Beni	5/1/2020	1	0	1	25	0	0	1	?	?	?	?	?	?
'San Beni	5/2/2020	1	0	1	25	0	0	2	?	?	?	?	?	?
'San Beni	5/3/2020	0	0	0	25	0	0	2	?	?	?	?	?	?
'San Beni	5/4/2020	0	0	0	25	0	0	2	?	?	?	?	?	?
'San Beni	5/5/2020	0	0	0	25	0	0	2	?	?	?	?	?	?
'San Beni	5/6/2020	0	0	0	25	0	0	1	?	?	?	?	?	?
'San Beni	5/7/2020	0	0	0	25	0	0	0	?	?	?	?	?	?
'San Beni	5/8/2020	0	0	0	25	0	0	1	?	?	?	?	?	?

Next, i used attribute rank method in WEKA and result here(choose hospitalized covid confirmed patients as target):



```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (numeric): 3 hospitalized_covid_confirmed_patients):
        Correlation Ranking Filter
Ranked attributes:
 0.9589    5 hospitalized_covid_patients
 0.6231    7 icu_covid_confirmed_patients
 0.1808    4 hospitalized_suspected_covid_patients
 0.1099    8 icu_suspected_covid_patients
 0.0651    2 todays_date
 0        14 trailers_delivered
 0         6 all_hospital_beds
 0        15 donated_trailers_delivered
 0         1 county
 0        10 date
 0        13 trailers_requested
 0        12 rooms_occupied
 0        11 rooms
-0.27      9 icu_available_beds

Selected attributes: 5,7,4,8,2,14,6,15,1,10,13,12,11,9 : 14
```

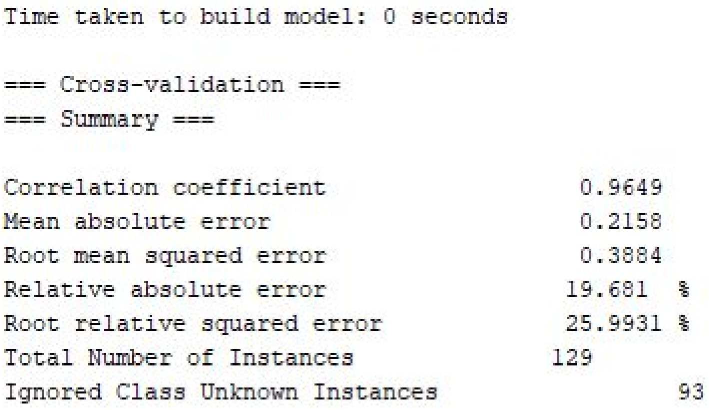
So i decided to delete all attributes which rank is 0 or below.
Rest attributes :



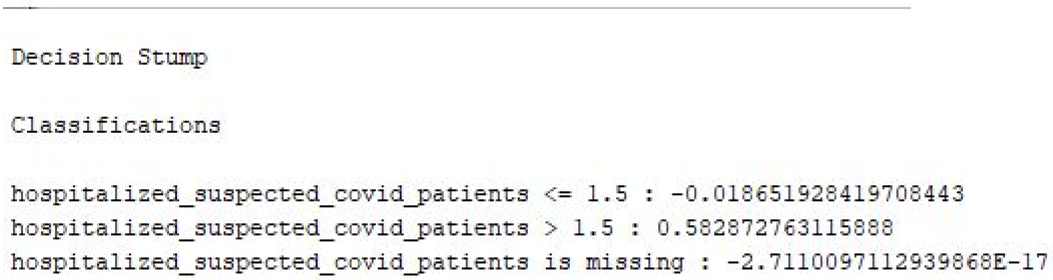
```
1  □ todays_date
2  □ hospitalized_covid_confirmed_patients
3  □ hospitalized_suspected_covid_patients
4  □ hospitalized_covid_patients
5  □ icu_covid_confirmed_patients
6  □ icu_suspected_covid_patients
```

By attribute rank method in WEKA, we can easily find that these six attributes shown above are important for output.

By doing Additive Regression again, the result shows below:



```
Time taken to build model: 0 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.9649
Mean absolute error                      0.2158
Root mean squared error                  0.3884
Relative absolute error                 19.681  %
Root relative squared error             25.9931 %
Total Number of Instances              129
Ignored Class Unknown Instances                 93
```

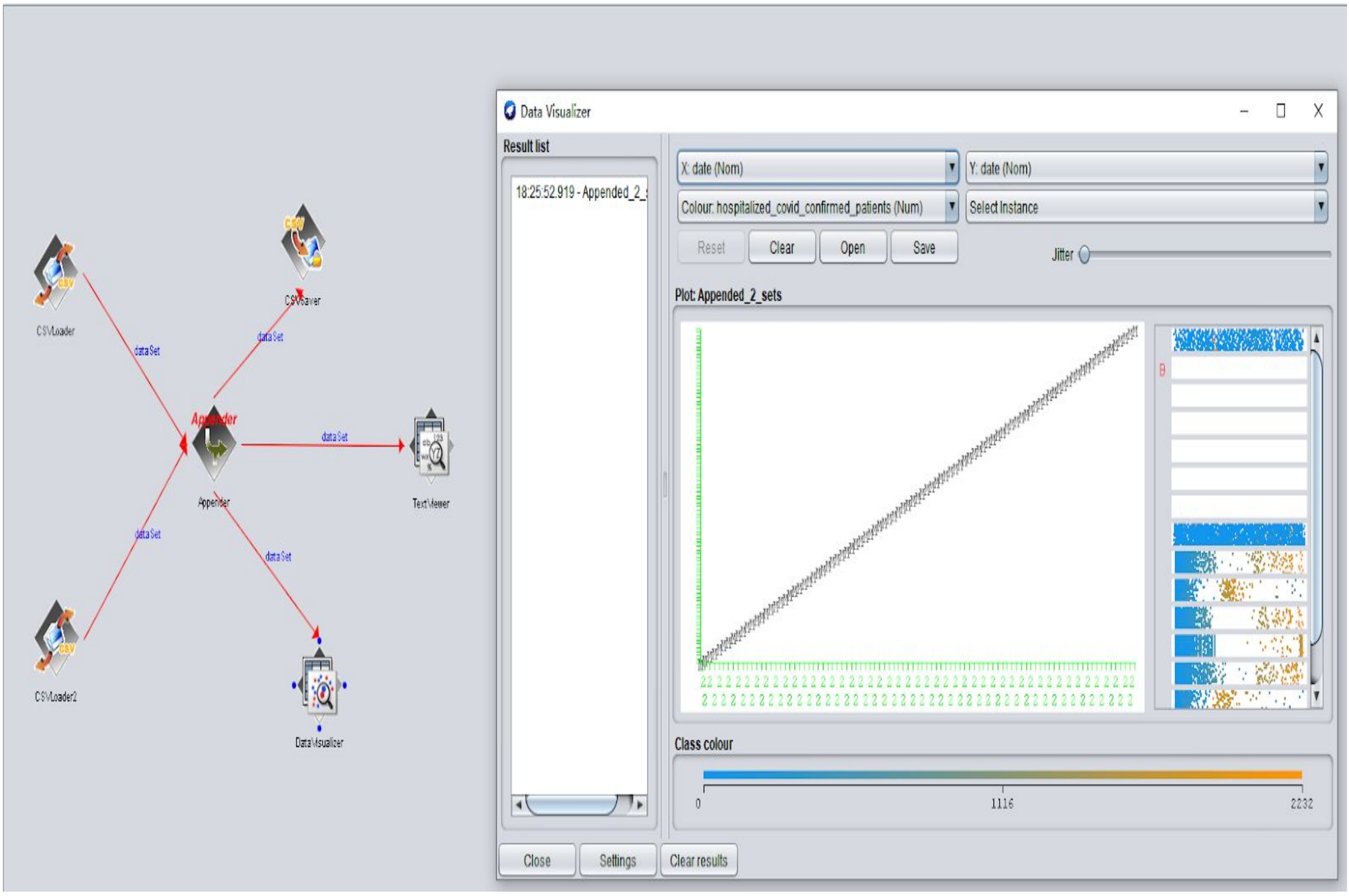


```
Decision Stump

Classifications

hospitalized_suspected_covid_patients <= 1.5 : -0.018651928419708443
hospitalized_suspected_covid_patients > 1.5 : 0.582872763115888
hospitalized_suspected_covid_patients is missing : -2.711009711293986 8E-17
```

And we can find that hospitalized_suspected_covid_patients attribute have the highest weights in regression.

2.3 Part 3: Challenge: Visualisation of results [20 marks]

graph I got from pipeline:



18:25:52.919 - Appended_2_s

X: date (Nom)

Y: date (Nom)

Colour: hospitalized_covid_confirmed_patients (Num)

Select Instance

Plot: Appended_2_sets