CAB430 Assignment 2

1. Task 1: COVID19 Infection Risk-level Prediction in DMX

1.1 DMX Script with mining structure and two mining models designs to predict Covid-19 Infection Risk Level

```
// Reset line for mining structure
DROP MINING STRUCTURE [Fact Survey]
```

1.1.1 Mining structure

```
//Create mining structure for Fact Survey table
CREATE MINING STRUCTURE [Fact Survey]
(
  [Survey_ID] LONG KEY,
  [Date] TEXT DISCRETE,
  [Risk_infection_level] TEXT DISCRETE,
  [Gender] TEXT DISCRETE,
  [Age] TEXT DISCRETE,
  [Height] DOUBLE DISCRETIZED,
  [Weight] DOUBLE DISCRETIZED,
  [BMI] DOUBLE DISCRETIZED,
  [BloodType] TEXT DISCRETE,
  [Insurance] TEXT DISCRETE,
  [Race] TEXT DISCRETE,
  [Smoking] TEXT DISCRETE,
  [Contact_count] LONG DISCRETIZED,
  [House_count] LONG DISCRETIZED,
  [Working] TEXT DISCRETE,
  [Covid19_symptoms] TEXT DISCRETE,
  [Covid19_contact] TEXT DISCRETE,
  [Asthma] TEXT DISCRETE,
  [Kidney_Diesease] TEXT DISCRETE,
  [Liver_Diesease] TEXT DISCRETE,
  [Diabetes] TEXT DISCRETE,
  [Hiv_positive] TEXT DISCRETE,
  [Hypertension] TEXT DISCRETE,
  [Other_chronic] TEXT DISCRETE,
  [Nursing_Home] TEXT DISCRETE,
  [Health_worker] TEXT DISCRETE
)
WITH HOLDOUT (30 PERCENT or 1000 CASES)

// Process Fact Survey structure
INSERT INTO MINING STRUCTURE [Fact Survey]
(
  [Survey_ID],
  [Date],
  [Risk_infection_level],
```

```
  [Gender],
  [Age],
  [Height],
  [Weight],
  [BMI],
  [BloodType],
  [Insurance],
  [Race],
  [Smoking],
  [Contact_count],
  [House_count],
  [Working],
  [Covid19_symptoms],
  [Covid19_contact],
  [Asthma],
  [Kidney_Diesease],
  [Liver_Diesease],
  [Diabetes],
  [Hiv_positive],
  [Hypertension],
  [Other_chronic],
  [Nursing_Home],
  [Health_worker]
)
OPENQUERY(COVID19Survey,
  'SELECT
     f.Survey_ID,
     f.Date,
       f.Risk_infection_level,
     p.Gender, p.Age, p.Height, p.Weight, p.BMI, p.BloodType, p.Insurance,
p.Race,
     r.Smoking, r.Contact_count, r.House_count, r.Working,
r.Covid19_symptoms, r.Covid19_contact,
     r.Asthma, r.Kidney_disease, r.Liver_disease, r.Diabetes,
r.Hiv_positive, r.Hypertension,
       r.Other_chronic, r.Nursing_home, r.Health_worker
   FROM Fact_survey f
   JOIN Participant p ON f.Participant = p.Participant_ID
   JOIN Response r ON f.Response = r.Response_ID')
```

// Reset line for Participant mining model
DROP MINING MODEL Participant

Mining structure was created and processed by performing an inner join across the 3 fact tables Fact_survey, Participant and Response. This ensure that all the relevant information from the different tables is consolidated into the mining structure, allowing different mining models to be created based on the integrated data.

1.1.2 Model 1: Using the demographical attributes only as input attributes

```
// Add demographical mining model
ALTER MINING STRUCTURE [Fact Survey]
ADD MINING MODEL [Participant]
(
[Survey_ID],
[Gender],
[Age],
[Height],
[Weight],
[BMI],
[BloodType],
[Insurance],
[Race],
[Risk_infection_level] PREDICT
) USING Microsoft_Association_Rules
WITH DRILLTHROUGH
GO
INSERT INTO [Participant]

// Reset line for response mining model
DROP MINING MODEL Attributes
```

1.1.3 Model 2: Using Covid-19 related attributes as input attributes

```
// Add Covid-19 related attributes to mining structure
ALTER MINING STRUCTURE [Fact Survey]
ADD MINING MODEL [Attributes]
(
[Survey_ID],
[Risk_infection_level] PREDICT,
[Smoking],
[Contact_count],
[House_count],
[Working],
[Covid19_symptoms],
[Covid19_contact],
[Asthma],
[Kidney_Diesease],
[Liver_Diesease],
[Diabetes],
[Hiv_positive],
[Hypertension],
[Other_chronic],
[Nursing_Home],
[Health_worker]
) USING Microsoft_Association_Rules
WITH DRILLTHROUGH
GO
INSERT INTO [Attributes]
```

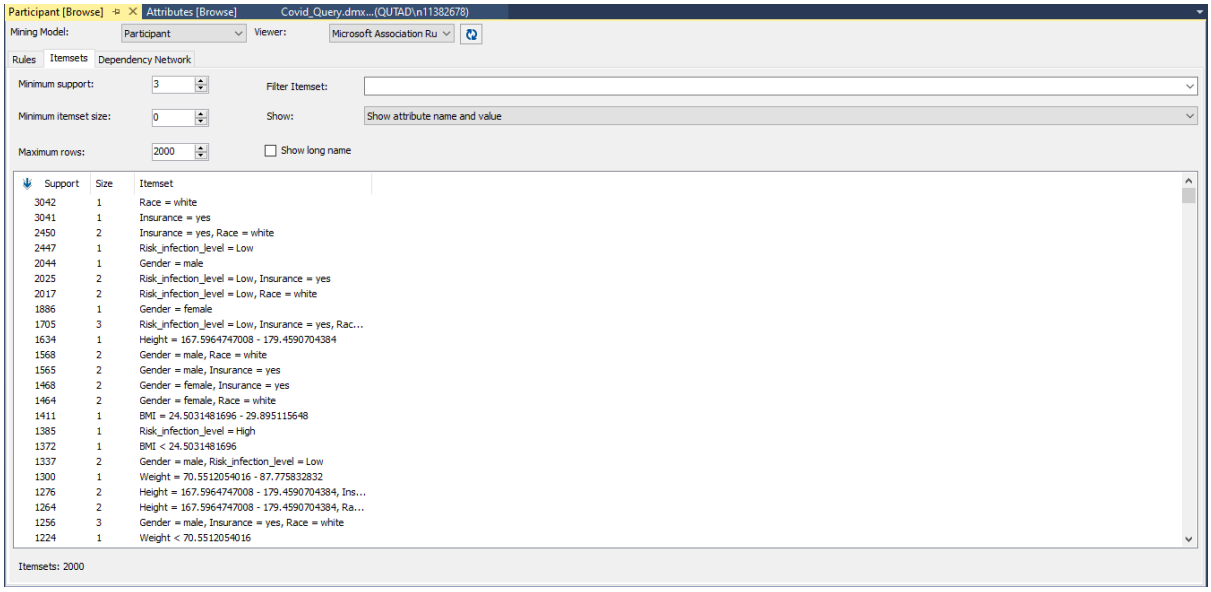
// Browse training cases for both models

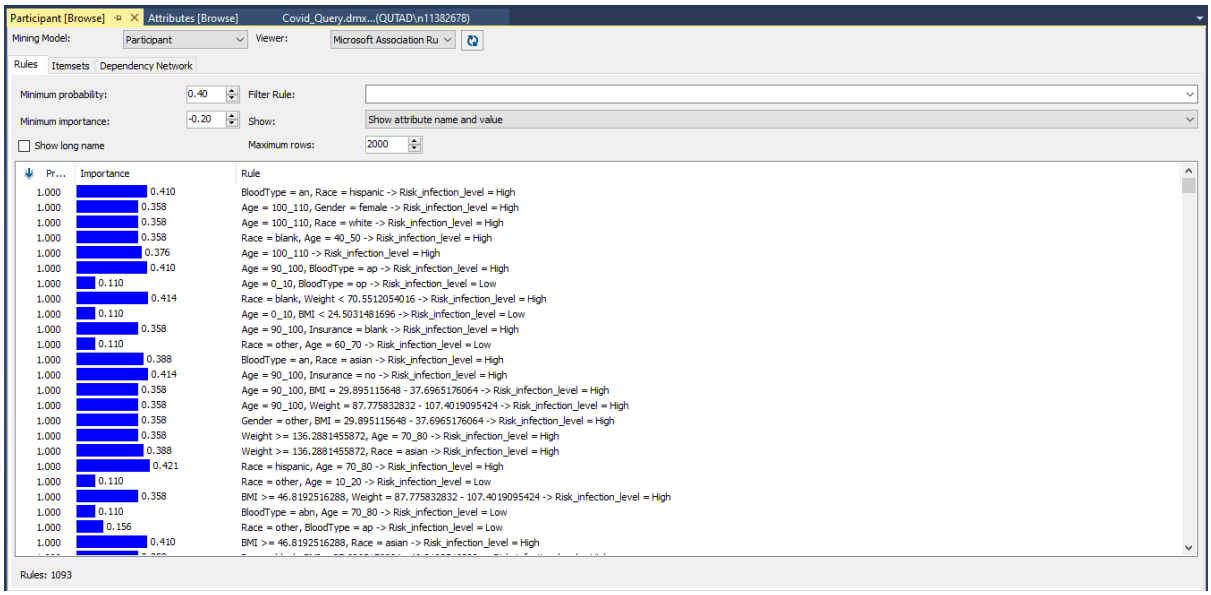
```
SELECT * FROM [Fact Survey].CASES WHERE IsTrainingCase()
```

The 2 mining models are created to predict Risk infection levels based on data from the Participant and the Response table respectively.

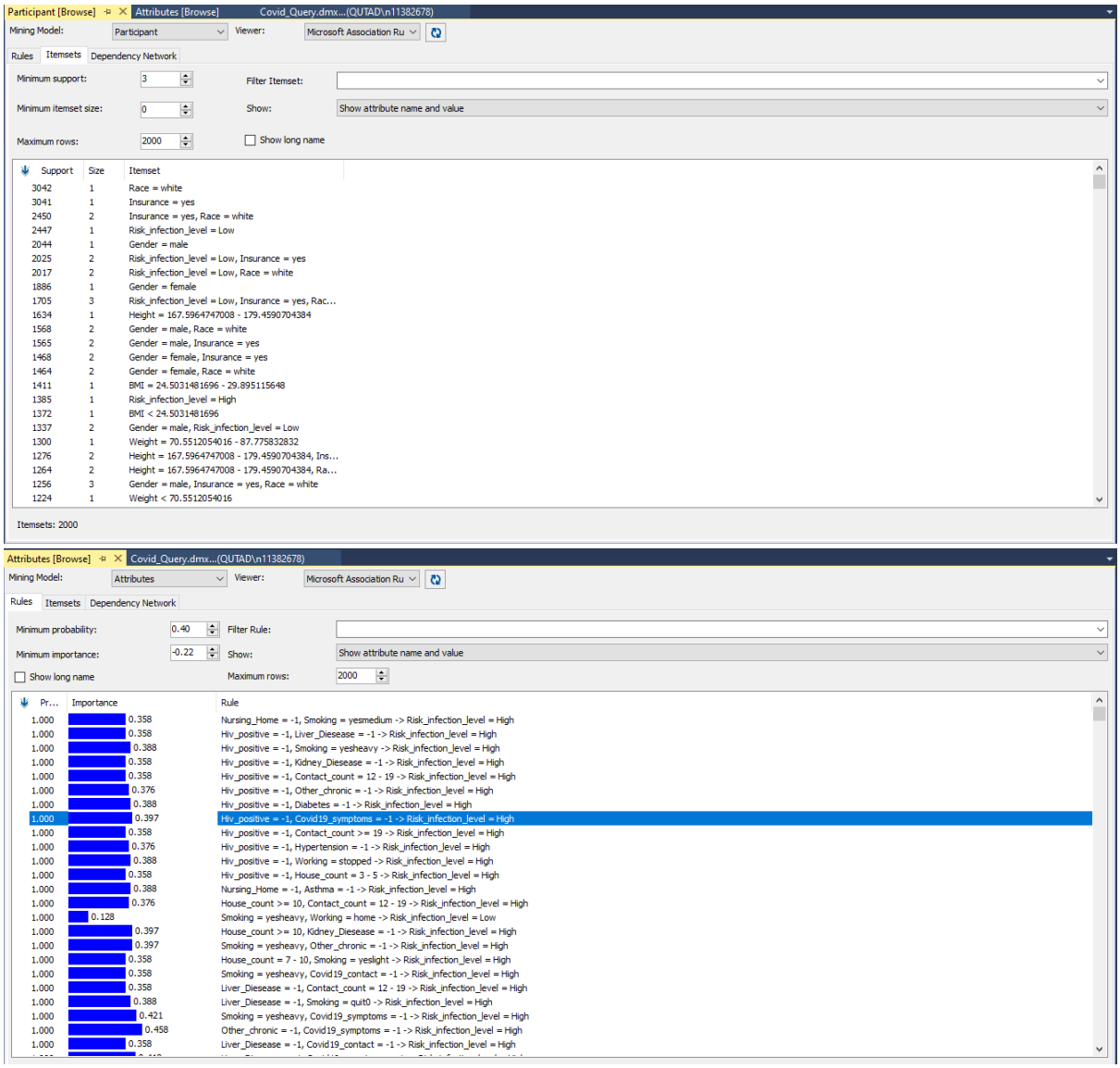
1.2 Processing mining structure and mining models

1.2.1 Itemsets and association rules generated by the model 1





1.2.2 Itemsets and association rules generated by the model 2



1.3 Predictions

1.3.1 Design one batch query in the cases in the test dataset

This batch query uses the testing dataset to predict the infection risk level for each participant case based on their demographic attributes such as gender, age, height, weight, BMI, blood type, insurance status, and race. It applies a trained prediction model using a `NATURAL PREDICTION JOIN` to generate a predicted risk level alongside the actual values for evaluation purposes.

```
// batch query with testing as input, predict infection risk-level for each case in the
// testing dataset based on demographical attributes

SELECT t.[Survey_ID], t.[Gender], t.[Age],t.[Height], t.[Weight], t.[BMI],
t.[BloodType], t.[Insurance], t.[Race],t.[Risk_infection_level],
PREDICT([Risk_infection_level]) AS [Prediction on infection risk level]
From
```

```
[Participant]
NATURAL PREDICTION JOIN
(SELECT * FROM [Participant].CASES WHERE IsTestCase()
) AS t
```

1.3.2 Batch Query Predictions from test case

Survey_ID	Gender	Age	Height	Weight	BMI	BloodType	Insurance	Race	Risk_infection_l...	Prediction on inf...
1169	male	50_60	163.8447108352	97.5888711872	33.7958166272	ap	yes	white	High	Low
557	female	50_60	173.5277725696	57.2756027008	20.9015740848	op	yes	white	High	Low
2667	female	50_60	173.5277725696	79.1635191168	27.1991319088	unknown	yes	white	Low	Low
3698	male	70_80	212.1835050752	97.5888711872	20.9015740848	an	yes	white	Low	Low
2537	female	70_80	163.8447108352	57.2756027008	20.9015740848	ap	yes	white	Low	Low
3952	female	30_40	163.8447108352	79.1635191168	27.1991319088	op	yes	white	Low	Low
179	female	30_40	173.5277725696	57.2756027008	20.9015740848	op	yes	white	High	Low
1183	male	50_60	173.5277725696	79.1635191168	27.1991319088	unknown	yes	white	High	Low
4424	female	30_40	173.5277725696	57.2756027008	20.9015740848	unknown	yes	white	Low	Low
3964	female	10_20	173.5277725696	57.2756027008	20.9015740848	bp	yes	white	Low	Low
1896	female	30_40	173.5277725696	97.5888711872	33.7958166272	unknown	blank	white	Low	High
1659	male	50_60	173.5277725696	79.1635191168	27.1991319088	unknown	yes	white	High	Low
4508	female	30_40	163.8447108352	121.8450275648	57.6596258144	op	yes	mixed	Low	High
2849	male	70_80	173.5277725696	97.5888711872	27.1991319088	op	yes	white	Low	Low
2003	male	80_90	173.5277725696	121.8450275648	42.2578846176	unknown	no	black	Low	High
2544	male	40_50	212.1835050752	158.1440727936	42.2578846176	unknown	yes	white	High	High
4123	male	10_20	212.1835050752	97.5888711872	20.9015740848	unknown	yes	white	Low	Low

1.3.3 Design one batch query using the data in database COVID19_Survey

This batch query predicts the infection risk level for each participant using health-related attributes such as pre-existing conditions, COVID-19 symptoms, contact history, and work/living situation. It uses a PREDICTION JOIN to apply a trained model on new response data from the COVID19Survey database, generating infection risk predictions based on matched attribute values.

```
// batch querry from database
SELECT
    t.Smoking, t.Contact_count, t.House_count, t.Working,
    t.Covid19_symptoms, t.Covid19_contact,
    t.Asthma, t.Kidney_Disease, t.Liver_Disease,
    t.Diabetes, t.Hiv_positive, t.Hypertension,
    t.Other_chronic, t.Nursing_Home, t.Health_worker,
    PREDICT([Risk_infection_level]) AS [Prediction on infection risk
level]
FROM
    [Attributes]
PREDICTION JOIN
    OPENQUERY([COVID19Survey], '
        SELECT
            Smoking, Contact_count, House_count, Working,
            Covid19_symptoms, Covid19_contact,
            Asthma, Kidney_Disease, Liver_Disease,
            Diabetes, Hiv_positive, Hypertension,
            Other_chronic, Nursing_Home, Health_worker
        FROM dbo.Response
    ') AS t
ON
```

```
[Attributes].Smoking = t.Smoking AND
[Attributes].Contact_count = t.Contact_count AND
[Attributes].House_count = t.House_count AND
[Attributes].Working = t.Working AND
[Attributes].Covid19_symptoms = t.Covid19_symptoms AND
[Attributes].Covid19_contact = t.Covid19_contact AND
[Attributes].Asthma = t.Asthma AND
[Attributes].Kidney_Diesease = t.Kidney_Disease AND
[Attributes].Liver_Diesease = t.Liver_Disease AND
[Attributes].Diabetes = t.Diabetes AND
[Attributes].Hiv_positive = t.Hiv_positive AND
[Attributes].Hypertension = t.Hypertension AND
[Attributes].Other_chronic = t.Other_chronic AND
[Attributes].Nursing_Home = t.Nursing_Home AND
[Attributes].Health_worker = t.Health_worker
```

1.3.4 Batch Query Predictions from database COVID19_Survey

Messages Results

Contact_count	House_count	Working	Covid19_sympto...	Covid19_contact	Asthma	Kidney_Disease	Liver_Disease	Diabetes	Hiv_positive	Hypertension	Other_chronic	Nursing_Home	Health_worker	Prediction on inf...
2	3	home	False	False	False	False	False	True	False	False	False	False	False	Low
4	2	stopped	False	False	False	False	False	False	False	False	True	False	False	Low
5	1	never	True	True	False	False	False	True	False	False	False	False	False	High
1	2	never	False	False	False	False	False	False	False	False	False	False	False	Low
8	3	home	False	False	False	False	False	False	False	False	False	False	False	Low
12	3	travel non critical	False	False	True	False	False	False	False	False	False	False	False	High
4	4	stopped	True	False	True	False	False	False	False	False	False	False	False	High
10	5	travel non critical	False	False	True	False	False	False	False	False	False	False	False	Low
5	1	never	False	False	False	False	False	False	False	True	False	False	False	Low
1	2	never	False	False	True	False	False	False	False	True	False	False	False	Low
3	3	never	False	False	True	False	False	False	False	True	False	False	False	Low
1	1	never	False	False	False	False	False	False	False	False	False	False	False	Low
13	2	travel critical	True	True	False	False	False	False	False	False	False	False	False	Low
2	2	stopped	False	False	False	False	False	False	False	False	False	False	False	Low
21	4	travel critical	False	False	False	False	False	False	False	False	False	False	True	Low
4	2	never	False	False	False	False	False	False	False	False	False	False	False	Low

2. Task 2: COVID19 Infection Risk Prediction in Python

2.1 Attribute Correlations

These ‘Correlation with Risk_Infection’ and ‘Correlation with Covid19_positive’ tables, describes the magnitude and direction of the relationship between the possible predictor variables and the response. At first glance, there seems to be 3 distinct predictor variables that have the strongest

relationship with the outcome each. This may imply that these variables are the most important at predicting the outcome, with a potential to have the most predictive power.

```
 Correlation with Risk_infection:
 Risk_infection            1.000000
Covid19_positive           0.872925
Covid19_symptoms           0.408807
Covid19_contact            0.374394
Contact_count              0.189440
Health_worker              0.181035
Bmi                        0.169186
Diabetes                   0.162780
Public_transport_count     0.161291
Heart_disease              0.139480
Kidney_disease             0.116639
House_count                0.115837
Weight                     0.115758
Nursing_home               0.101290
Compromised_immune         0.092417
Lung_disease               0.076963
Other_chronic              0.069739
Asthma                     0.069005
Liver_disease              0.062092
Hiv_positive               0.048625
Hypertension               0.032634
Height                    -0.082907
Risk_mortality            -0.087730
Name: Risk_infection, dtype: float64
```

```
 Correlation with Covid19_positive:
 Covid19_positive           1.000000
Risk_infection             0.872925
Covid19_symptoms           0.431816
Covid19_contact            0.331895
Bmi                        0.158030
House_count                0.104146
Weight                     0.100846
Kidney_disease             0.098595
Compromised_immune         0.095849
Nursing_home               0.090988
Diabetes                   0.086498
Other_chronic              0.082307
Heart_disease              0.073997
Asthma                     0.067128
Contact_count              0.055458
Hiv_positive               0.054567
Liver_disease              0.033469
Lung_disease               0.029953
Public_transport_count     0.027326
Health_worker              0.017112
Hypertension               0.016920
Risk_mortality            -0.076786
Height                    -0.096941
Name: Covid19_positive, dtype: float64
```

2.2 Feature Selection

When looking at feature selection using ANOVA and Chi-Square scores for the predictors for risk infection and covid 19 positive outcomes, there are more variables identified as having explanatory power. Compared to the correlation results, ANOVA and Chi-squared filtered out variables that may not contribute much to the model's predictive power. Overall, streamlining the model training process and improves performance by removing irrelevant or redundant features

```
---- Feature selection for anova ----

 input Column names: ['Risk_mortality', 'Height', 'Weight', 'Bmi', 'Contact_count', 'House_count', 'Public_transport_count'
, 'Covid19_symptoms', 'Covid19_contact', 'Asthma', 'Kidney_disease', 'Liver_disease', 'Compromised_immune', 'Heart_disease'
, 'Lung_disease', 'Diabetes', 'Hiv_positive', 'Hypertension', 'Other_chronic', 'Nursing_home', 'Health_worker']

 target column names: ['Risk_infection', 'Covid19_positive']

Scores for Risk_infection: [ 20.23670943   2.40881121   5.43055119   8.32045252  26.0151959
   9.23488296  50.29142932  47.60350065  67.86227526   2.5921231
   5.73786535  80.38753305   2.89046427  16.64879846  24.77265857
  94.20178931   0.47505714   3.83059337   1.62643231  18.04867313
 192.47180768]
Top 10 features for Risk_infection from ANOVA: ['Risk_mortality' 'Contact_count' 'Public_transport_count'
 'Covid19_symptoms' 'Covid19_contact' 'Liver_disease' 'Lung_disease'
 'Diabetes' 'Nursing_home' 'Health_worker']

Scores for Covid19_positive: [  18.95906198   45.86967915   79.2757932   168.68642764   53.24644755
   94.66912292    1.39309552 1131.806123    611.27709904   22.35201988
   48.47333148    5.53761993   45.78636207   27.18730867    4.4344069
   37.22440453   14.74692029    1.4141403    33.6807463    41.22229605
    1.44629255]
Top 10 features for Covid19_positive from ANOVA: ['Height' 'Weight' 'Bmi' 'Contact_count' 'House_count' 'Covid19_symptoms'
 'Covid19_contact' 'Kidney_disease' 'Compromised_immune' 'Nursing_home']

Transformed dataset shape using ANOVA for Risk_infection: (4940, 10)

Transformed dataset shape using ANOVA for Covid19_positive: (4940, 10)
---- Feature selection for chi-sqaure ----

Scores for Risk_infection: [5200.59834214   55.00005781 1101.73929194  526.94509892 4178.21267373
  265.73576546 8635.78337301 1062.32819642 1372.42116784   74.1689708
  180.75231808 1717.21889273   88.28327009  478.97324921  688.3066541
 1757.83432819   15.67791494  101.73234032   49.73670072  529.93476828
 2611.70188867]
Top 10 features for Risk_infection from CHI-SQAURED: ['Risk_mortality' 'Weight' 'Contact_count' 'Public_transport_count'
 'Covid19_symptoms' 'Covid19_contact' 'Liver_disease' 'Lung_disease'
 'Diabetes' 'Health_worker']

Scores for Covid19_positive: [166.01726768  31.74778151 493.97201545 328.41530838 299.29046398
  85.47025153   9.63555384 816.71650103 482.36596748  19.50255579
  47.28277254   5.48100902  42.52691722  26.04188563   4.32366019
  33.90821953  14.65537061   1.15950355  31.1356899   40.52503125
   1.35539245]
Top 10 features for Covid19_positive from CHI-SQAURED: ['Risk_mortality' 'Weight' 'Bmi' 'Contact_count' 'House_count'
 'Covid19_symptoms' 'Covid19_contact' 'Kidney_disease'
 'Compromised_immune' 'Nursing_home']

Transformed dataset shape using chi-squared for Risk_infection: (4940, 10)

Transformed dataset shape using chi-squared for Covid19_positive: (4940, 10)
```

2.3 Prediction and evaluation

Two models, GaussianNB and Decision Trees, were developed with feature selections from either ANOVA or Chi-squared with k = 5,8,10,12,15. These models with minor k adjustments and feature selection methods were used to test which combination yielded the best results. That is, the highest accuracy score after performing predictions on the same and separate test dataset. The strategy is to follow a filter feature selection approach by a univariate method. Each attribute has an ANOVA or Chi-squared score individually, and the top features (k) is selected based on the score.

A snapshot of the results found below shows that the best classification algorithm is Decision Trees with ANOVA as the best feature selection method at k = 10. Thus, it was found that the Top 10 features for Risk Infection prediction using ANOVA is: Risk mortality, contact count, house count, covid19 symptoms, covid19 contact, kidney disease, compromise immune and nursing home. This results in an accuracy of 0.5617, which is the best result as prediction accuracy decreases as k continues to increase subsequently. This suggests that as more attributes are added to the model, the classification model is unable to generalize from overfitting by adding more “noise” to the training.

```
Feature selection: ANOVA, Prediction algorithm: Naive Bayes , for K =  8
Accuracy:  0.395748987854251

Feature selection: Chi-sqaure, Prediction algorithm: Naive Bayes, for K =  8
Accuracy:  0.49696356275303644

Feature selection: ANOVA, Prediction algorithm: Decision Trees, for K =  8
Accuracy:  0.6214574898785425

Feature selection: Chi square, Prediction algorithm: Decision Trees, for K =  8
Accuracy:  0.4817813765182186

Feature selection: ANOVA, Prediction algorithm: Naive Bayes , for K =  10
Accuracy:  0.5232793522267206

Feature selection: Chi-sqaure, Prediction algorithm: Naive Bayes, for K =  10
Accuracy:  0.5121457489878543

Feature selection: ANOVA, Prediction algorithm: Decision Trees, for K =  10
Accuracy:  0.5617408906882592

Feature selection: Chi square, Prediction algorithm: Decision Trees, for K =  10
Accuracy:  0.5293522267206477

Feature selection: ANOVA, Prediction algorithm: Naive Bayes , for K =  12
Accuracy:  0.5344129554655871

Feature selection: Chi-sqaure, Prediction algorithm: Naive Bayes, for K =  12
Accuracy:  0.5242914979757085

Feature selection: ANOVA, Prediction algorithm: Decision Trees, for K =  12
Accuracy:  0.5759109311740891

Feature selection: Chi square, Prediction algorithm: Decision Trees, for K =  12
Accuracy:  0.4959514170040486
```

3. Task 3: COVID19 Positive Prediction in Python

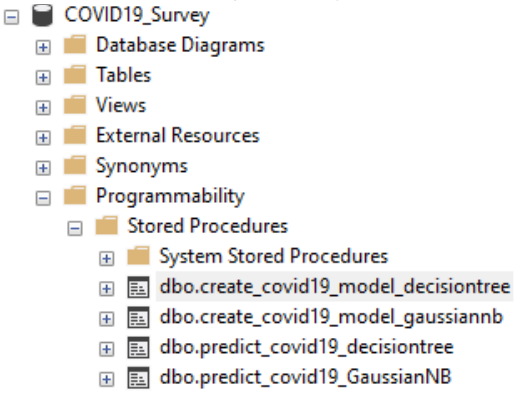
3.1 Stored procedures for creating machine learning models

Part 1 - Screenshot of models in the COVID19_ML_models

	model_name	model
1	DecisionTree_model	0x800363736B6C6561726E2E747265652E747265650A4465...
2	GaussianNB_model	0x800363736B6C6561726E2E6E616976655F62617965730A...

3.2 Stored procedures for generating predictions

3.2.1 Screenshots of stored procedures



3.2.2 Results from decision tree predictions

	Covid19_positive	prediction	correct	accuracy
1	1	1	1	0.420388349514563
2	1	1	1	0.420388349514563
3	1	1	1	0.420388349514563
4	1	0	0	0.420388349514563
5	1	1	1	0.420388349514563
6	1	0	0	0.420388349514563
7	1	0	0	0.420388349514563
8	1	0	0	0.420388349514563
9	1	0	0	0.420388349514563
10	1	0	0	0.420388349514563
11	1	0	0	0.420388349514563
12	1	0	0	0.420388349514563
13	1	0	0	0.420388349514563

3.2.2 Results from GaussianNB prediction

	Covid19_positive	prediction	correct	accuracy
1	1	1	1	0.431067961165049
2	1	1	1	0.431067961165049
3	1	1	1	0.431067961165049
4	1	1	1	0.431067961165049
5	1	1	1	0.431067961165049
6	1	1	1	0.431067961165049
7	1	0	0	0.431067961165049
8	1	0	0	0.431067961165049
9	1	1	1	0.431067961165049
10	1	0	0	0.431067961165049
11	1	1	1	0.431067961165049
12	1	0	0	0.431067961165049
13	1	0	0	0.431067961165049

3.3 Predictions and evaluation

The first set of input attributes drops Nursing_home which is the lowest correlation out of the batch and adds Risk_infection which was not counted in previous prediction as it was the target attribute but has high correlation to Covid19_positive as shown in the correlation section of this report.

The second set of input attributes is similar to the ones in the previous sub-question but the difference is that there is no date constraint.

3.3.1 First set of Input attributes

The input dataset includes demographic and health-related features such as Height, Weight, and BMI, along with exposure-related factors like Contact_count and House_count. It also captures symptom presence (Covid19_symptoms), exposure history (Covid19_contact), and pre-existing conditions (Kidney_disease and Compromised_immune). These features are used to predict the target variables: Covid19_positive (whether a participant tested positive) and Risk_infection (a classification of their infection risk). The data, drawn from a filtered subset of records dated between April 1st and 30th, 2020, is suitable for building classification models like Decision Trees or Naive Bayes to assess infection likelihood based on observable and reported attributes. To improve accuracy from the stored procedures model, the Nursing_home attribute was removed and replaced with the Risk_infection attribute.

3.3.1.1 Decision Tree Model Results

	Covid19_positive	prediction	correct	accuracy
1	1	1	1	0.790291262135922
2	1	1	1	0.790291262135922
3	1	1	1	0.790291262135922
4	1	0	0	0.790291262135922
5	1	1	1	0.790291262135922
6	1	0	0	0.790291262135922
7	1	1	1	0.790291262135922
8	1	1	1	0.790291262135922
9	1	0	0	0.790291262135922
10	1	1	1	0.790291262135922
11	1	0	0	0.790291262135922
12	1	1	1	0.790291262135922
13	1	1	1	0.790291262135922

3.3.1.2 GaussianNB Algorithm Results

	Covid19_positive	prediction	correct	accuracy
1	1	1	1	0.994174757281553
2	1	1	1	0.994174757281553
3	1	1	1	0.994174757281553
4	1	1	1	0.994174757281553
5	1	1	1	0.994174757281553
6	1	1	1	0.994174757281553
7	1	1	1	0.994174757281553
8	1	1	1	0.994174757281553
9	1	1	1	0.994174757281553
10	1	1	1	0.994174757281553
11	1	1	1	0.994174757281553
12	1	1	1	0.994174757281553
13	1	1	1	0.994174757281553

3.3.2 Second set of Input attributes

The second set of input attributes is the same as the first, but the date constraints are removed entirely.

3.3.2.1 Decision Tree Model Results

	Covid19_positive	prediction	correct	accuracy
1	0	1	0	0.837449392712551
2	0	0	1	0.837449392712551
3	1	1	1	0.837449392712551
4	0	0	1	0.837449392712551
5	0	0	1	0.837449392712551
6	0	0	1	0.837449392712551
7	1	1	1	0.837449392712551
8	0	0	1	0.837449392712551
9	0	0	1	0.837449392712551
10	0	0	1	0.837449392712551
11	0	0	1	0.837449392712551
12	0	0	1	0.837449392712551
13	1	1	1	0.837449392712551

3.3.1.2 GaussianNB Algorithm Results

	Covid19_positive	prediction	correct	accuracy
1	0	0	1	0.759919028340081
2	0	0	1	0.759919028340081
3	1	1	1	0.759919028340081
4	0	0	1	0.759919028340081
5	0	0	1	0.759919028340081
6	0	0	1	0.759919028340081
7	1	1	1	0.759919028340081
8	0	0	1	0.759919028340081
9	0	0	1	0.759919028340081
10	0	0	1	0.759919028340081
11	0	0	1	0.759919028340081
12	0	0	1	0.759919028340081

3.4 Final Evaluation on the machine learning models

Overall, the best-performing machine learning model is the Decision Tree algorithm trained on the full dataset using the attributes: Covid19_positive, Height, Weight, Bmi, Contact_count, House_count, Covid19_symptoms, Covid19_contact, Kidney_disease, Compromised_immune, and Risk_infection. This model achieved an accuracy of 0.837. While the GaussianNB model attained a higher accuracy of 0.9942 using the initial feature set, the Decision Tree model is considered more balanced and robust, as it was evaluated on the entire dataset without date constraints, reducing the risk of overfitting and offering more generalizable predictions.

5. Statement of completeness

All tasks have been completed. Colin had attempted tasks 1,2 and 3. Kayathri has attempted task 2, 3 and finalized the report.