Analysis of COVID-19 survival rate in Toronto

Abstract

Coronaviruses are a large group of viruses known to cause the common cold and even more serious illnesses, such as Middle East Respiratory syndrome (MERS) and Severe acute Respiratory syndrome (SARS).mNovel Coronavirus (CoVID-19) is a kind of novel coronavirus that has not been found in humans before 2019.To date, the total number of confirmed cases worldwide has exceeded 18 million. Seven hundred thousand people die of the disease, and the number is rising. How to control COVID-19 and reduce mortality has become a top priority for countries around the world. We now have data sets from Toronto on COVID-19 cases, and we want to analyze these data sets through the KDD process to develop models that actually mitigate the impact of COVID-19 globally and protect specific populations.

Key words: COVID-19 cases, Toronto, KDD process

目录

1.1 Identify the objectives of the business

1.1.1 Background

As a known large group of viruses, coronavirus can cause the common cold and even more serious diseases, such as the Middle East respiratory syndrome (MERS) and severe acute respiratory syndrome (SARS). The novel coronavirus (COVID-19) is a new coronavirus which has not been found in humans before 2019.

The World Health Organization (who) defines coronavirus disease as a pandemic. So far, the total number of confirmed cases in the world has exceeded 18 million. 700000 people have died of the disease, and the number is still on the rise. It not only has a profound impact on human health, but also has an indelible impact on the global economic and social environment. How to curb the novel coronavirus pneumonia and reduce mortality has become the primary problem to be solved.

1.1.2 Business objectives

Analyze the impact of age, gender, source of infection, outbreak associated, hospitalization status. Identify key factors and characteristics that contribute to survive in patients with COVID-19.

Control these characteristics and factors that lead to death in patients with COVID-19, and guide the country, institution or family to locate groups that are more likely to survive from COVID-19.

Protect those at risk for COVID-19 by pre-positioning and taking preventive measures.

Reduce national or regional mortality from COVID-19 through prevention.

1.1.3 Business success criteria

(1) Significant increasing in survival rate of novel coronavirus pneumonia. Obtain characteristics of COVID-19 cases, including sex, age, source of infection, hospital status, etc. Implementing targeted prevention, and successfully improve COVID-19 survival rate.

(2) The project can be completed on time without exceeding the budget.

1.2 Assess the situation

1.2.1 Assumption

(1) Survival rate for COVID-19 patients was associated with age. Generally speaking, young children are at the greatest risk of infection. For example, approximately 57% of malaria occurs in children under 5 years of age. However, in the face of the new coronavirus, the elderly are the most at risk. This may be because older people have potential health problems, especially cardiovascular diseases and respiratory diseases. The elderly are more likely to have these health problems than the young, which may be one of the important reasons why the elderly are not likely to survive the risk of COVID-19.

(2) The gendered impact on health outcomes. In many cases, since most of the world's health workers are women, women seem to be more likely to be diagnosed with covid-19. At the same time, compared with women, the male mortality rate in each country has maintained a higher growth

trend. This may be due to the fact that men have a higher smoking rate and are more likely to suffer from cardiopulmonary diseases.

(3) The probability that someone dies from a disease doesn't just depend on the disease itself, but also on the treatment they receive, and on the patient's own ability to recover from it.

1.2.2 Constraints

(1) The source of Data is single. The data of COVID-19 cases

for this study are from Toronto Public Health in January 2020. The loss of some data will still cause certain errors in the overall statistics of outcome data.

(2) In order to protect personal privacy, some more detailed data cannot be obtained, and the lack of critical information may skew the overall objectivity of the results.

1.3 Determine data mining objectives

1.3.1 Data mining goals

(1) Get a set of decision rules that determine the survival rate of COVID-19.

(2) Use decision rules to predict or classify the COVID- 19 survival rate of a person or group of people in the future

1.3.2 Data mining success criteria

(1) Model quality

More than 85% accuracy; Faster response speed; The output is easy to understand

(2) Engineering dimension

Flexible model; easy to use; tight layout, embeddable and extensible.

(3) Resource quality:

High data noise tolerance; Less sparse data.

(4) Logistical constraints

Simple calculation; less development time.

1.4 Produce a project plan

The overview plan for the study is as shown in the table below.

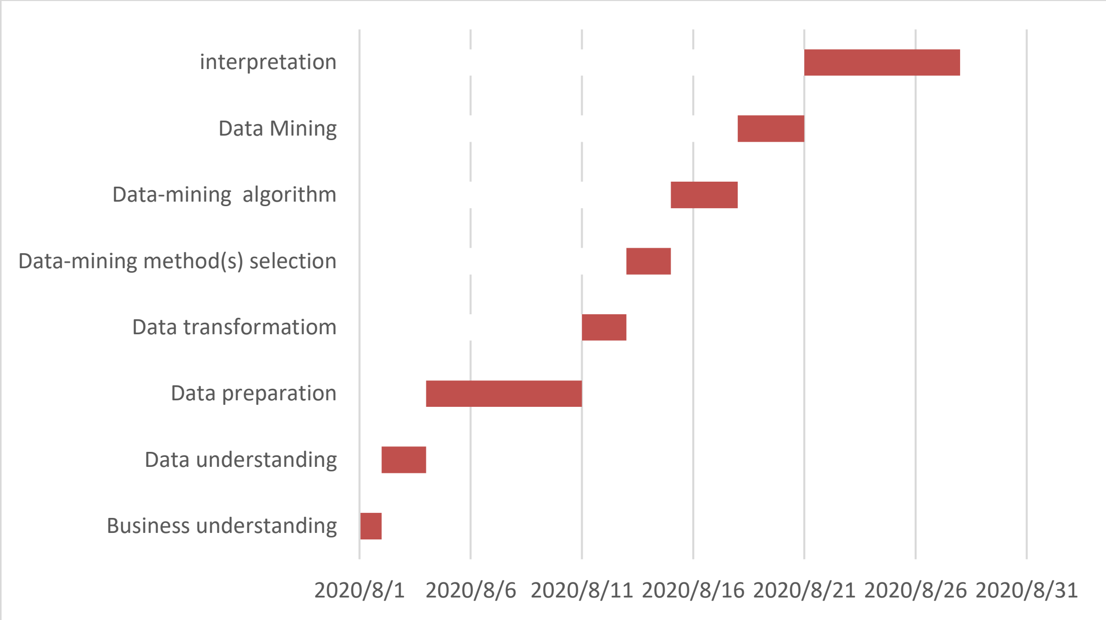


Table 1. project plan overview

2.1 Collect initial data

2.1.1 Source: Kaggle website

https://www.kaggle.com/divyansh22/toronto-covid19-cases

2.1.2 Original source:

Provincial communicable disease reporting system (iPHIS) and Toronto's custom COVID-19 case management system (CORES)

2.2 Describe the Data

2.2.1 Data quantity

(1) Size of dataset: 14912 records and 15 attributes, the data used in this analysis is to record the COVID- 19 survival rate of different populations

(2) Attributes: _id, Outbreak Associated, Age Group, Client Gender, Classification, Source of Infection, Episode Date, Reported Date, Outcome, Currently Hospitalized, Currently in ICU, Currently Intubated, Ever Hospitalized, Ever in ICU, Ever Intubated

(3) Data type

1. Basic data of the people:
   _id: numeric
   Age Group: categorical (string)
   Client Gender: categorical (string)

2. Basic data on virus infection

   Outbreak Associated: categorical (string)

   Classification: categorical (string)

   Source of Infection: categorical (string)

   Episode Date: numeric

   Reported Date: numeric

3. Basic data of severity information

   Outcome: categorical (string)

   Currently Hospitalized: categorical (string)

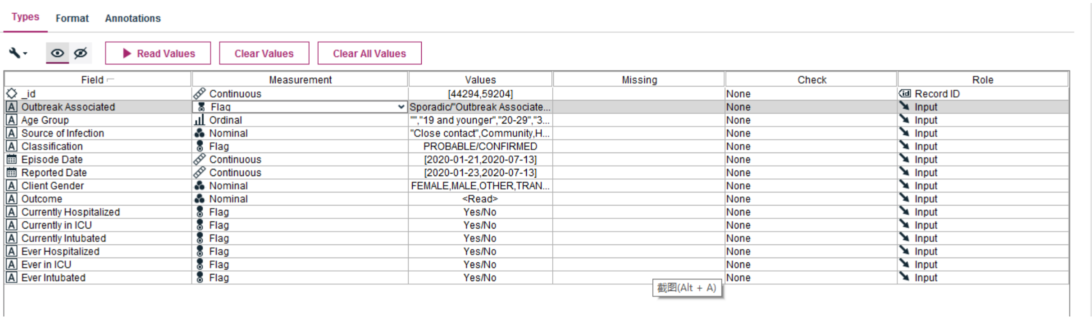
   Currently in ICU: categorical (string)

Currently Intubated: categorical (string)

Ever Hospitalized: categorical (string)

Ever in ICU: categorical (string)

Ever Intubated: categorical (string)



Field	Measurement	Values	Missing	Check	Role
_id	Continuous	[44294,59204]		None	Record ID
Outbreak Associated	Flag	Sporadic/"Outbreak Associate...		None	Input
Age Group	Ordinal	"","19 and younger","20-29","3...		None	Input
Source of Infection	Nominal	"Close contact",Community,H...		None	Input
Classification	Flag	PROBABLE/CONFIRMED		None	Input
Episode Date	Continuous	[2020-01-21,2020-07-13]		None	Input
Reported Date	Continuous	[2020-01-23,2020-07-13]		None	Input
Client Gender	Nominal	FEMALE,MALE,OTHER,TRAN...		None	Input
Outcome	Nominal	<Read>		None	Input
Currently Hospitalized	Flag	Yes/No		None	Input
Currently in ICU	Flag	Yes/No		None	Input
Currently Intubated	Flag	Yes/No		None	Input
Ever Hospitalized	Flag	Yes/No		None	Input
Ever in ICU	Flag	Yes/No		None	Input
Ever Intubated	Flag	Yes/No		None	Input

figure 1. type of data

2.3 Explore data

2.3.1 data overview

This data set contains demographic, geographic, and severity information for all confirmed and probable cases reported to and managed by Toronto Public Health since the first case was reported in January 2020.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	_id	Outbreak	Age Grou	Source of	Classificat	Episode Date	Reported Date	Client Ger	Outcome	Currently	Currently	Currently	Ever Hosp	Ever in ICU	Ever Intubated	
2	44294	Sporadic	50-59	Institutior	CONFIRMI	2020/3/25	2020/3/27	MALE	RESOLVED	No	No	No	No	No	No	
3	44295	Sporadic	20-29	Communi	CONFIRMI	2020/3/20	2020/3/28	MALE	RESOLVED	No	No	No	Yes	No	No	
4	44296	Sporadic	60-69	Travel	CONFIRMI	2020/3/4	2020/3/8	FEMALE	RESOLVED	No	No	No	Yes	Yes	Yes	
5	44297	Outbreak	50-59	N/A - Outl	CONFIRMI	2020/5/2	2020/5/4	FEMALE	RESOLVED	No	No	No	No	No	No	
6	44298	Sporadic	30-39	Close con	CONFIRMI	2020/5/31	2020/6/6	FEMALE	RESOLVED	No	No	No	No	No	No	
7	44299	Sporadic	20-29	Close con	CONFIRMI	2020/6/1	2020/6/6	MALE	RESOLVED	No	No	No	No	No	No	
8	44300	Sporadic	60-69	Communi	CONFIRMI	2020/5/22	2020/6/1	MALE	RESOLVED	No	No	No	No	No	No	
9	44301	Sporadic	30-39	Close con	PROBABLE	2020/5/26	2020/6/2	MALE	RESOLVED	No	No	No	No	No	No	
10	44302	Sporadic	30-39	Close con	CONFIRMI	2020/5/11	2020/5/16	MALE	RESOLVED	No	No	No	No	No	No	
11	44303	Sporadic	19 and you	Close con	PROBABLE	2020/6/6	2020/6/9	MALE	RESOLVED	No	No	No	No	No	No	
12	44304	Sporadic	30-39	Close con	CONFIRMI	2020/5/17	2020/5/21	MALE	RESOLVED	No	No	No	No	No	No	
13	44305	Outbreak	20-29	N/A - Outl	CONFIRMI	2020/4/23	2020/4/25	MALE	RESOLVED	No	No	No	No	No	No	
14	44306	Sporadic	30-39	Pending	CONFIRMI	2020/4/22	2020/4/22	MALE	RESOLVED	No	No	No	No	No	No	
15	44307	Sporadic	30-39	Close con	CONFIRMI	2020/5/28	2020/6/6	MALE	RESOLVED	No	No	No	No	No	No	
16	44308	Sporadic	80-89	Communi	CONFIRMI	2020/4/11	2020/4/12	MALE	FATAL	No	No	No	Yes	No	No	
17	44309	Sporadic	80-89	Healthcar	CONFIRMI	2020/5/9	2020/6/3	FEMALE	RESOLVED	No	No	No	Yes	No	No	
18	44310	Sporadic	50-59	Institutior	CONFIRMI	2020/4/13	2020/4/15	FEMALE	RESOLVED	No	No	No	No	No	No	
19	44311	Outbreak	30-39	N/A - Outl	CONFIRMI	2020/5/13	2020/5/17	MALE	RESOLVED	No	No	No	No	No	No	
20	44312	Sporadic	20-29	Close con	CONFIRMI	2020/5/19	2020/5/23	MALE	RESOLVED	No	No	No	No	No	No	
21	44313	Outbreak	20-29	N/A - Outl	CONFIRMI	2020/4/18	2020/5/4	FEMALE	RESOLVED	No	No	No	No	No	No	
22	44314	Sporadic	70-79	Close con	CONFIRMI	2020/4/25	2020/4/30	FEMALE	RESOLVED	No	No	No	No	No	No	

figure 2. data set of COVID-19 cases

2.3.2 Data audit

Most of the data is of type String, representing categories or judgment results.

Field	Sample Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
_id		Continuous	44294	59204	51749	4304.579	0	--	14911
Outbreak Associated		Flag	--	--	--	--	--	2	14911
Age Group		Ordinal	--	--	--	--	--	10	14879
Source of Infection		Nominal	--	--	--	--	--	8	14911
Classification		Flag	--	--	--	--	--	2	14911
Episode Date		Continuous	2020-01-21	2020-07-13	--	--	--	--	14911
Reported Date		Continuous	2020-01-23	2020-07-13	--	--	--	--	14911
Client Gender		Nominal	--	--	--	--	--	5	14911

figure 3. data audit of dataset

2.3.3 Data analyses

This project uses SPSS modeler to analyze the basic data. Through the drawing function of histogram, histogram, scatter plot and broken line chart in the software, the preliminary statistical analysis of the data set can be carried out.

(1) Taking age as an example, it can be seen from the figure that among the confirmed cases, there are more confirmed cases between 20 and 60 years old, almost all of which are more than 1,500 people, while there are fewer cases in other age groups.

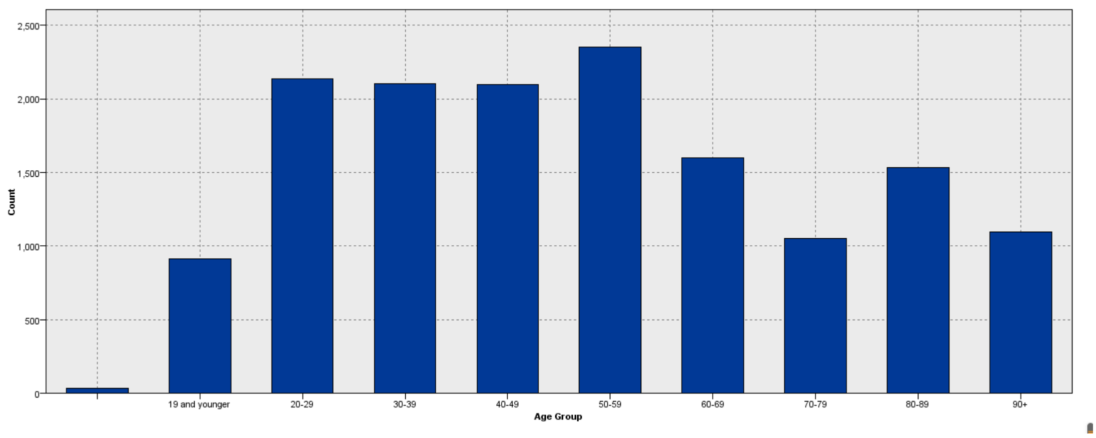
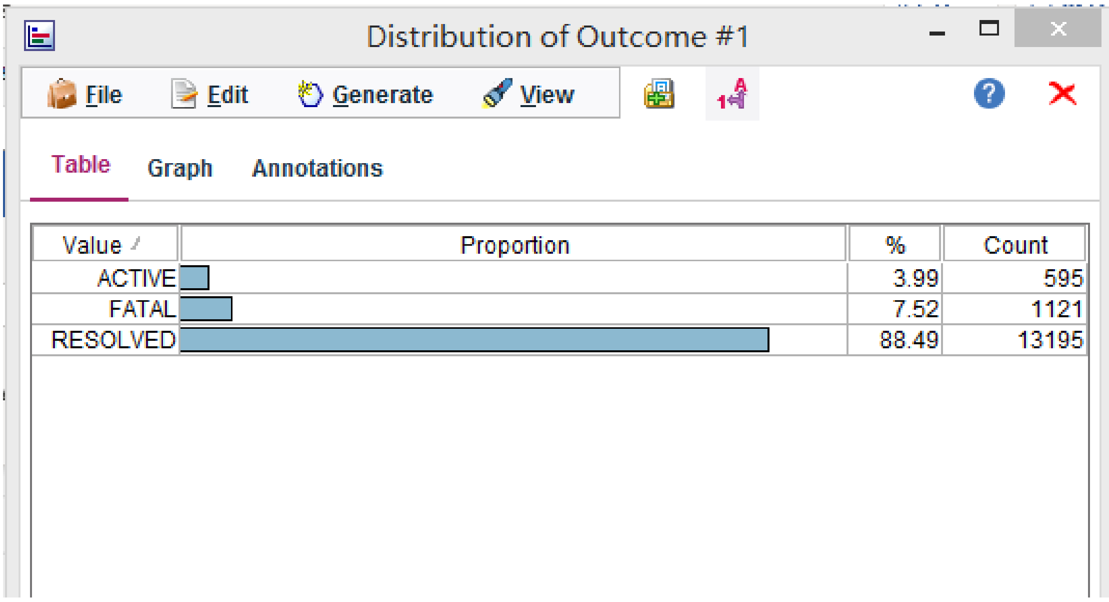


figure 4 confirmed or suspected cases over different age group

(2) In all confirmed and suspected cases, the mortality rate is about 7.52%, and the survival rate of current cases is higher than 88%. Explore the influence of age, sex, source of infection, and severity of disease on the survival rate of infected patients.



Value	Proportion	%	Count
ACTIVE		3.99	595
FATAL		7.52	1121
RESOLVED		88.49	13195

figure 5. Current survival and mortality rates

2.4 Verify the data quality

2.4.1 Data missing

First of all, all magnetic field measurements are correct. As can be seen from Figure 6, there are basically extreme values and outliers in the data. This is because most of the data is of Type String and cannot be evaluated. The Data is relatively complete, only the age group has a small number of empty strings and white space. This may be because it is difficult to collect complete information during the COVID-19 outbreak.

2.4.2 Data errors and metric errors

No data errors were found, and the deviation values that emerged are phenomena that warrant further analysis.

Complete fields (%): 93.33% Complete records (%): 99.79%

Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value	Empty String	White Space
_id	Continuous	0	0	None	Never	Fixed	100	14911	0	0	0
Outbreak As...	Flag	--	--	--	Never	Fixed	100	14911	0	0	0
Age Group	Ordinal	--	--	--	Never	Fixed	99.785	14879	0	32	32
Source of Inf...	Nominal	--	--	--	Never	Fixed	100	14911	0	0	0
Classification	Flag	--	--	--	Never	Fixed	100	14911	0	0	0
Episode Date	Continuous	0	0	None	Never	Fixed	100	14911	0	0	0
Reported Date	Continuous	0	0	None	Never	Fixed	100	14911	0	0	0
Client Gender	Nominal	--	--	--	Never	Fixed	100	14911	0	0	0
Outcome	Nominal	--	--	--	Never	Fixed	100	14911	0	0	0
Currently Ho...	Flag	--	--	--	Never	Fixed	100	14911	0	0	0
Currently in I...	Flag	--	--	--	Never	Fixed	100	14911	0	0	0
Currently Intu...	Flag	--	--	--	Never	Fixed	100	14911	0	0	0
Ever Hospital...	Flag	--	--	--	Never	Fixed	100	14911	0	0	0
Ever in ICU	Flag	--	--	--	Never	Fixed	100	14911	0	0	0
Ever Intubated	Flag	--	--	--	Never	Fixed	100	14911	0	0	0

figure 6. the data quality of data set

3.1 Select the Data

By observing complete data and the distribution of missing data, I decided to select ten attributes in the dataset.

Select attribute: Age Group, Client Gender, Source of Infection, Outcome, Currently Hospitalized, Currently in ICU, Currently Intubated, Ever Hospitalized, Ever in ICU and Ever Intubated. Because these attributes may have different degrees of influence and response in analyzing COVID- 19 survival rate of different populations.

(1) Source of Infection

The condition of confirmed cases may be related to different sources of infection, so different sources of infection may increase or decrease the survival rate of COVID-19.

(2) Client Gender

Gender may lead to different COVID- 19 survival rate because of men and women differences in personality and physical conditions.

(3) Age Group

People in different age groups have large differences in their physical, mental, or economic conditions, so this attribute should be retained, and its effect on COVID- 19 survival rate.

(4) Outcome

Data on the status of all confirmed and probable cases at present, which is the direct data of COVID- 19 survival rate.

(5) Currently Hospitalized and Ever Hospitalized

Whether or not a patient is admitted to hospital represents to some extent, the level of treatment for a confirmed case, which may affect the patient's level of recovery.

(6) Currently in ICU and Ever in ICU

The ICU represents the level of treatment for severe cases and affects the survival rate of COVID-19 cases.

(7) Currently Intubated and Ever Intubated

Intubated as treatment may play a decisive role in the rehabilitation of COVID-19 cases and increase the survival rate of COVID-19.

3.2 Clean the Data

3.2.1 Useless data and processing methods

Useless Data	reason	method
_id	no research value	Use filter
Outbreak Associated	Incidents of various COVID-19 outbreaks Not relevant to aim of this project	Use filter
Classification	It had nothing to do with survival after covid-19.	Use filter
Episode Date	Means the time of coVID-19 diagnosis is independent of survival rate.	Use filter
Reported Date	Means the time of coVID-19 diagnosis is independent of survival rate.	Use filter

Table 2. Useless data and processing methods

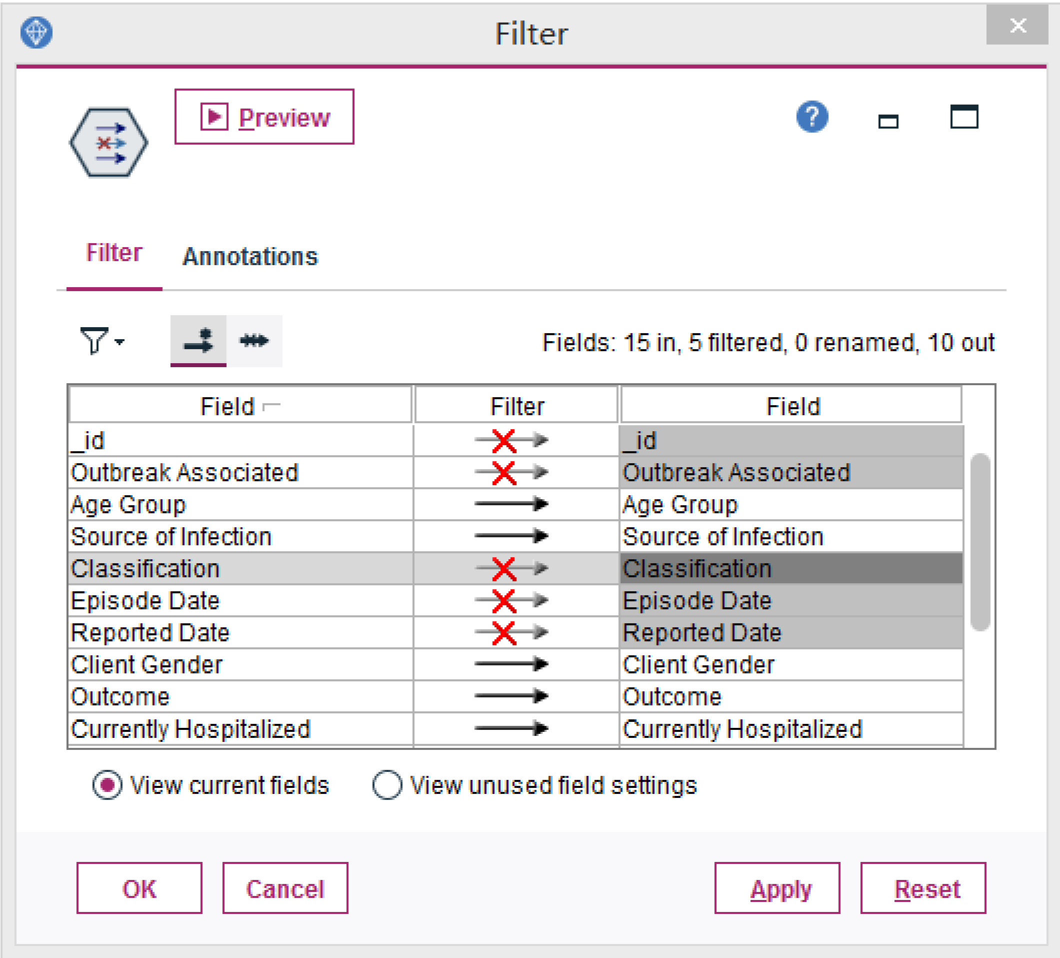


figure 7. filter useless fields

3.2.2. Missing data

(1) According to the review report, there were seven missing values that were represented by "" in the original data set and could not be found. So we first use Type to define and check missing values.

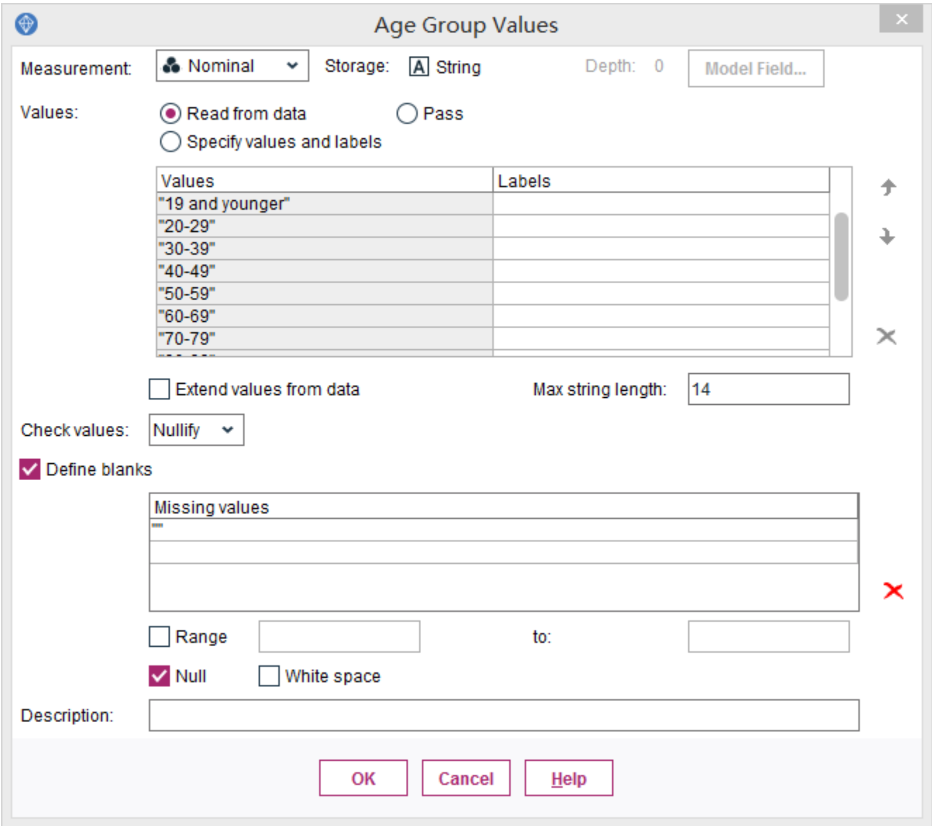


figure 8. define missing values.

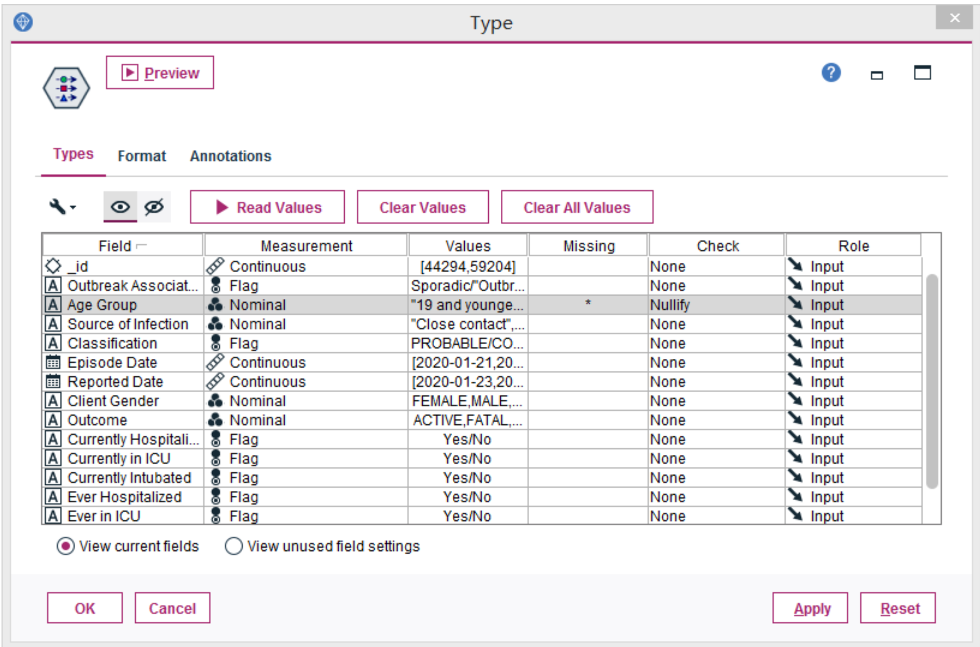


figure 9. check missing values

（2）For the missing value of the data, empty string values and white space are treated as distinct from null values. Empty strings are treated as equivalent to white space for most purposes. So I select the missing value in age group and treat them as blanks. As shown in the following figure:

Complete fields (%): 90% Complete records (%): 99.79%

Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value	Empty String	White Space
Age Group	Ordinal	--	--	--	Blank & Null Val...	Random	99.785	14879	0	32	32
Source of Inf...	Nominal	--	--	--	Never	Fixed	100	14911	0	0	0
Client Gender	Nominal	--	--	--	Never	Fixed	100	14911	0	0	0
Outcome	Nominal	--	--	--	Never	Fixed	100	14911	0	0	0
Currently Ho...	Flag	--	--	--	Never	Fixed	100	14911	0	0	0
Currently in I...	Flag	--	--	--	Never	Fixed	100	14911	0	0	0
Currently Intu...	Flag	--	--	--	Never	Fixed	100	14911	0	0	0
Ever Hospital...	Flag	--	--	--	Never	Fixed	100	14911	0	0	0
Ever in ICU	Flag	--	--	--	Never	Fixed	100	14911	0	0	0
Ever Intubated	Flag	--	--	--	Never	Fixed	100	14911	0	0	0

figure 10. fill blanks with an estimated value.

Complete fields (%): 100% Complete records (%): 100%

Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value	Empty String	White Space
Age Group	Categorical	--	--	--	Never	Fixed	100	14911	0	0	0
Source of Inf...	Nominal	--	--	--	Never	Fixed	100	14911	0	0	0
Client Gender	Nominal	--	--	--	Never	Fixed	100	14911	0	0	0
Outcome	Nominal	--	--	--	Never	Fixed	100	14911	0	0	0
Currently Ho...	Flag	--	--	--	Never	Fixed	100	14911	0	0	0
Currently in I...	Flag	--	--	--	Never	Fixed	100	14911	0	0	0
Currently Intu...	Flag	--	--	--	Never	Fixed	100	14911	0	0	0
Ever Hospital...	Flag	--	--	--	Never	Fixed	100	14911	0	0	0
Ever in ICU	Flag	--	--	--	Never	Fixed	100	14911	0	0	0
Ever Intubated	Flag	--	--	--	Never	Fixed	100	14911	0	0	0

Figure11. the data quality after replacing the missing value.

3.3 Construct the data

Create three new attributes. Use the derived node to generate the field Outcome, Intubated and Hospitalized. Both the former hospitalization and the current hospitalization belong to the hospitalization treatment, which can be combined into one field. Similarly, Intubated experiences can also compose a field. There are three values in Outcome ACTIVE, RESOLVED and FATAL. ACTIVE and RESOLVED both represent patients still alive and can be combined into a single value.

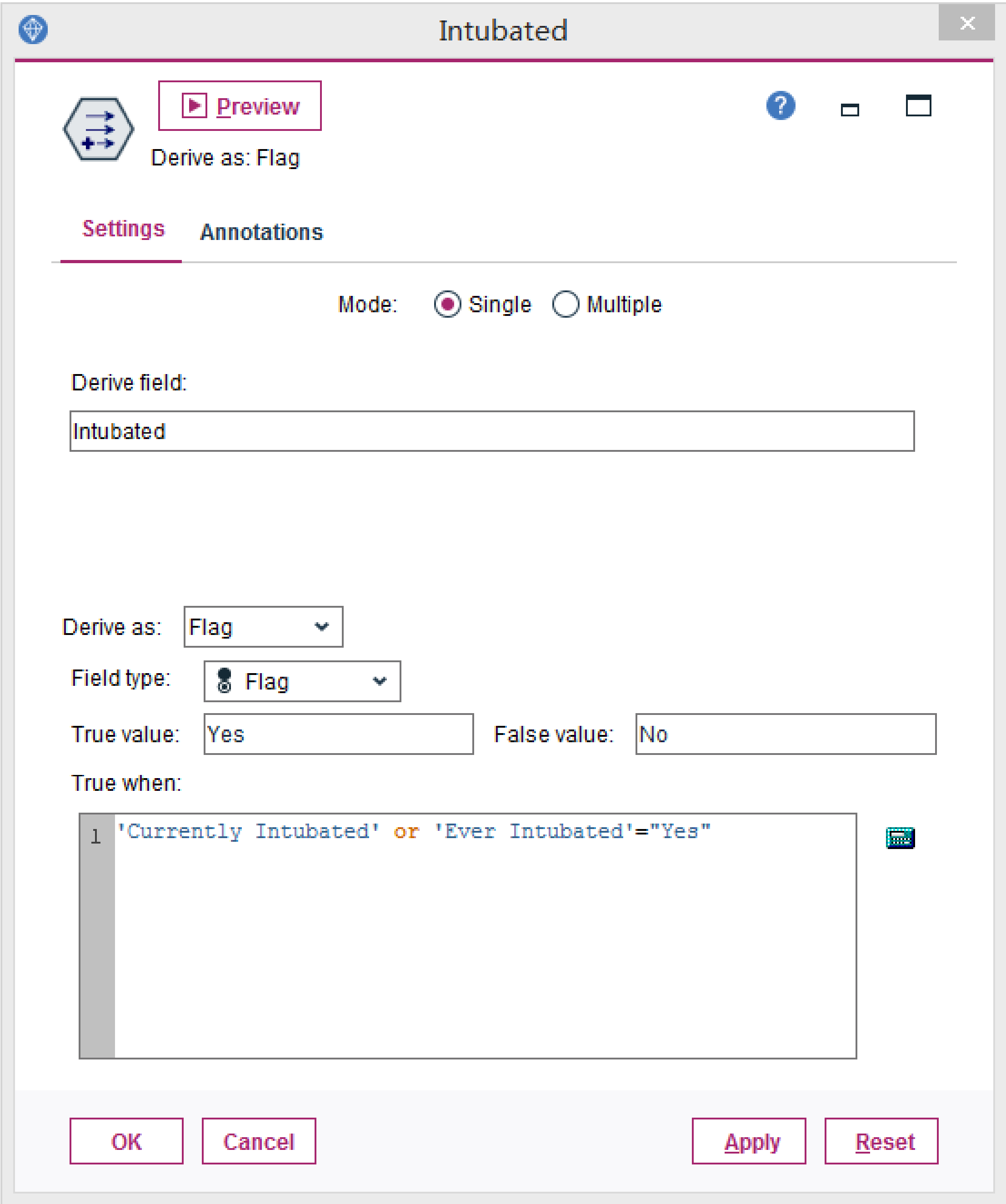


Figure12. generate new field- Intubated

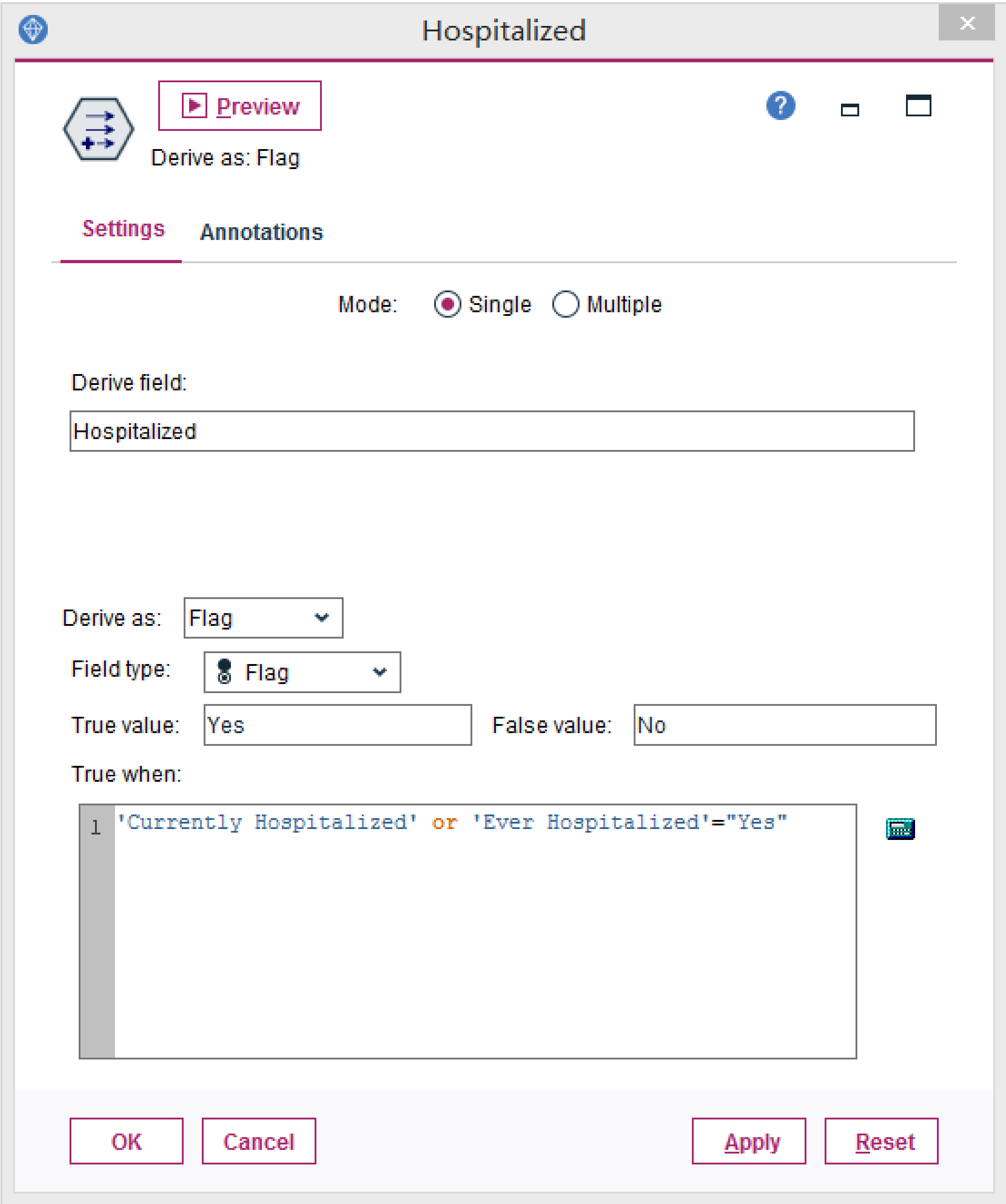


Figure13. generate new field- Hospitalized

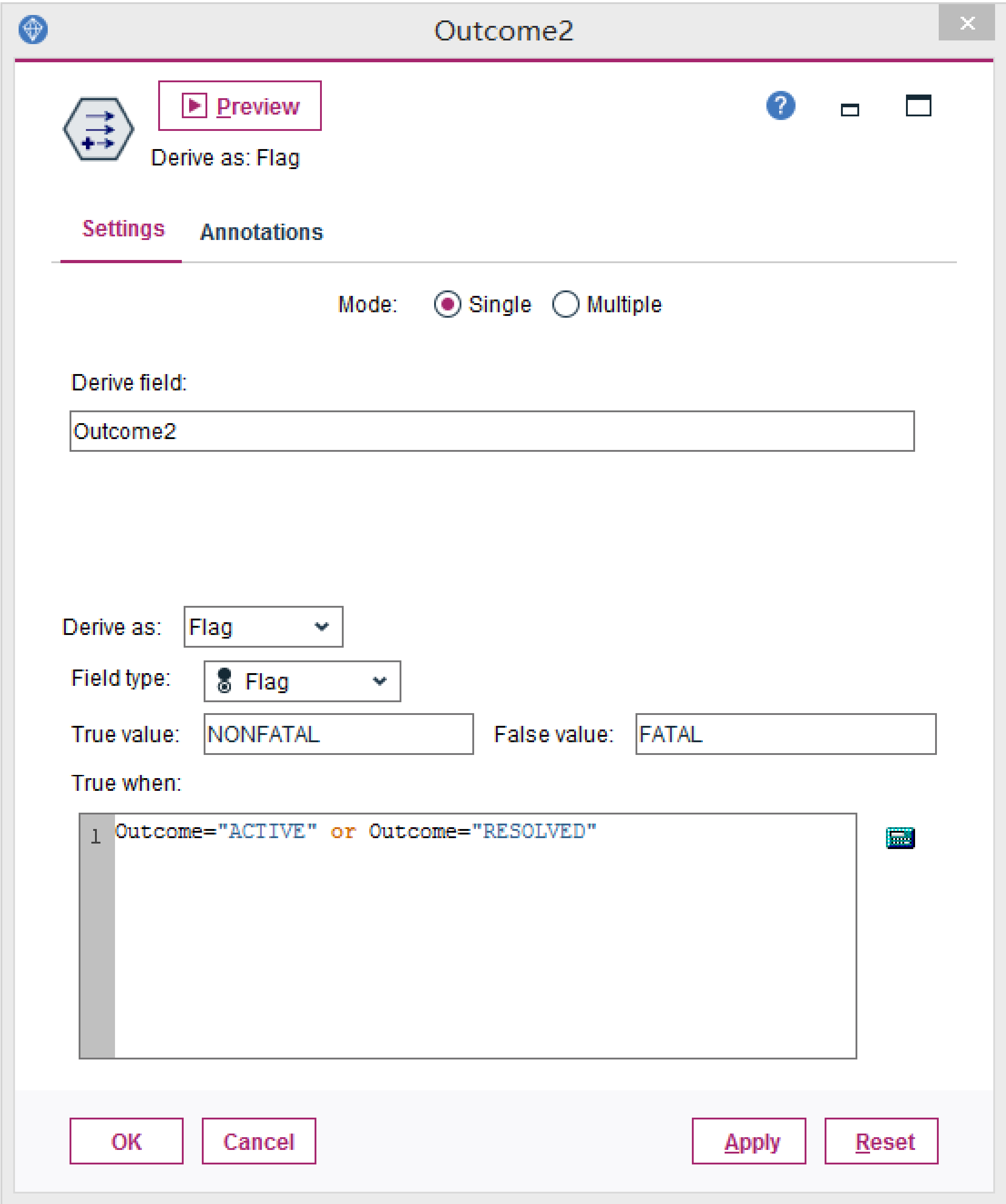


Figure14. generate new field- Cotcome2

3.4 Integrate various data sources

There is no need to integrate data because the project has only one data source, since the current data source already includes medical factors (hospitalization, source of infection) and personal factors (gender, age). There is already a comprehensive set of factors affecting cure rates, so we can use current data sources for data mining

3.5 Format the data as required

3.5.1 Format the data to fit decision tree model

(1) Decision tree requires to set the target attribute, so set the target attribute with the type node, which is Outcome2.

(2) This model requires the data type to be numeric or string type (unordered set), the data source meets the requirements

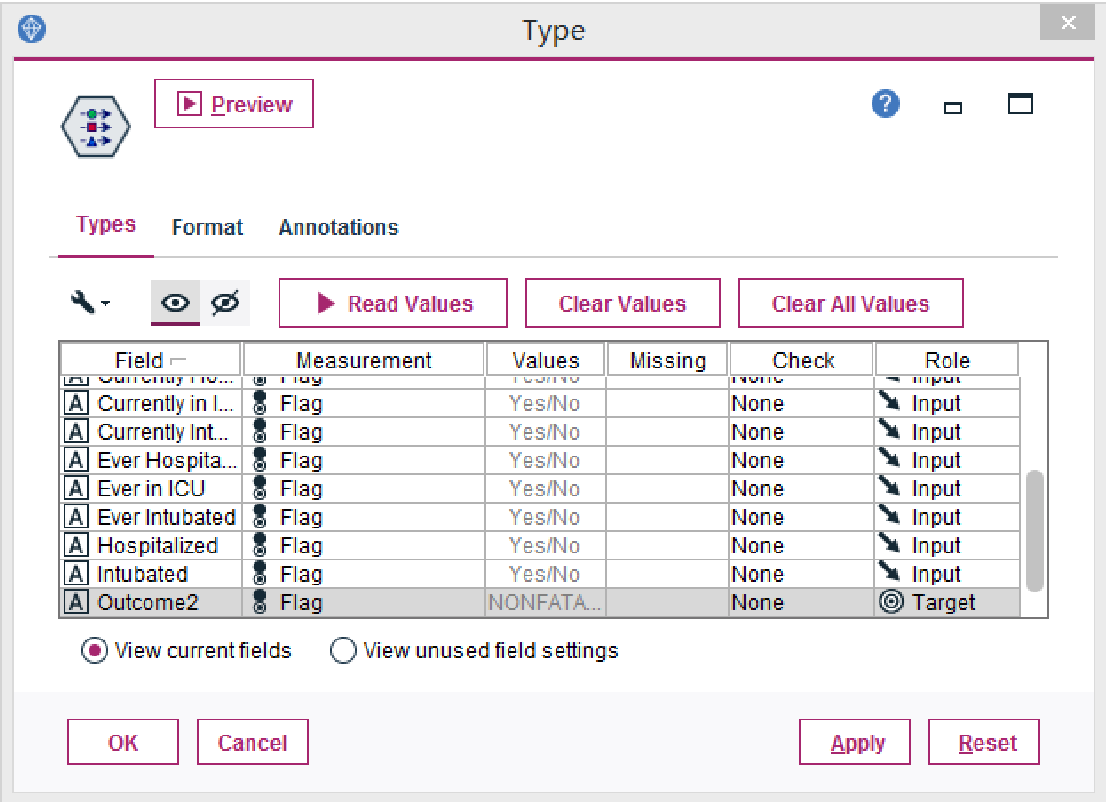


Figure15. Format the data

3.5.2 Check again for useless fields

The Hospitalized attribute is generated based on the currently hospitalized and the ‘ever hospitalized’ attributes using the derive node. The Intubated attribute is generated based on the currently intubated and the ever the currently hospitalized and the ‘ever hospitalized’ attributes using the derive node. Hospitalized, currently in ICU and, ‘ever in ICU’ are similar and repetitive attributes. So we need to filter out duplicate properties again to reduce interference.

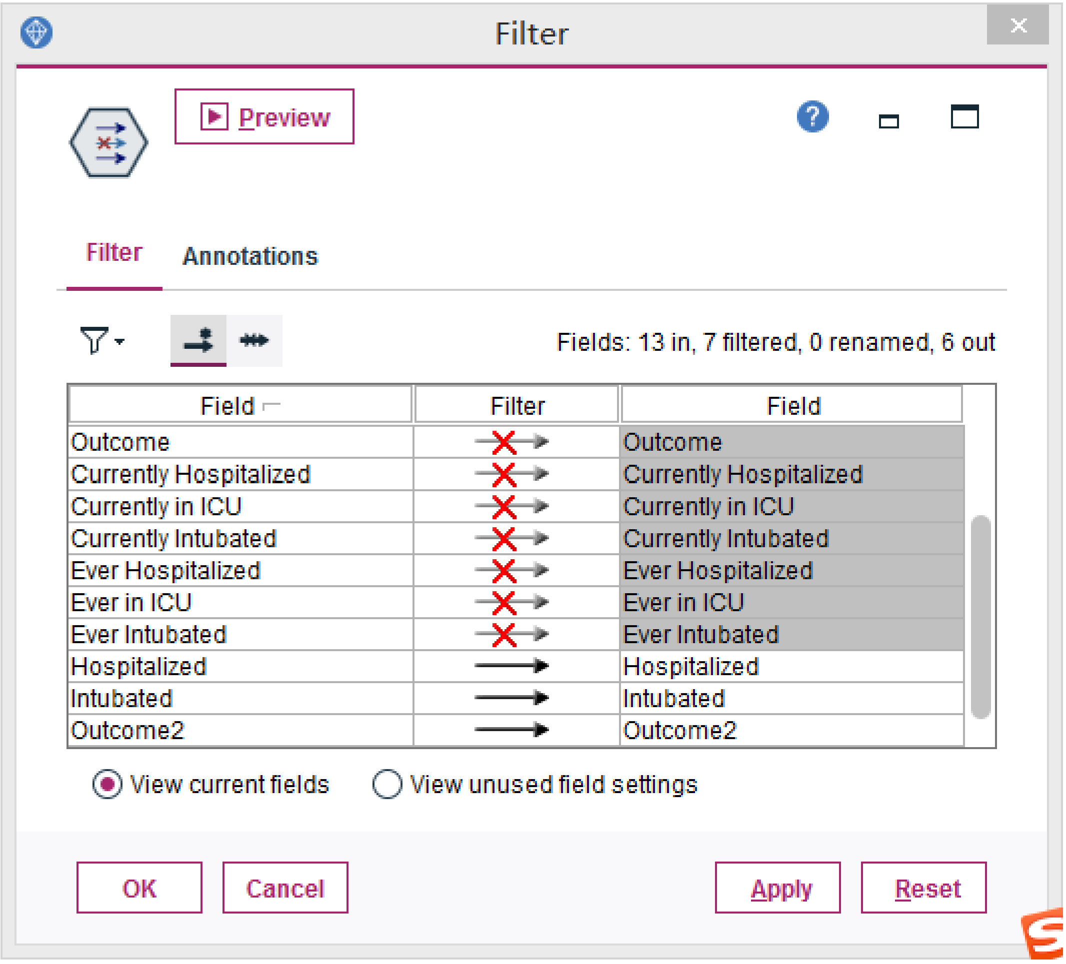


Figure16.Filter useless fields

4.1 Reduce the Data

4.1.1 Feature selection

(1) Use the feature selection node to select features related to the

predictor variable Outcome2. Use this to reduce data vertically

(2) The result is shown below, the attributes including Age Group, Hospitalized, Source of Infection and Client Genger are shown as important

(3) Intubated is shown as a coefficient of variation below threshold, so reducing attribute Intubated.

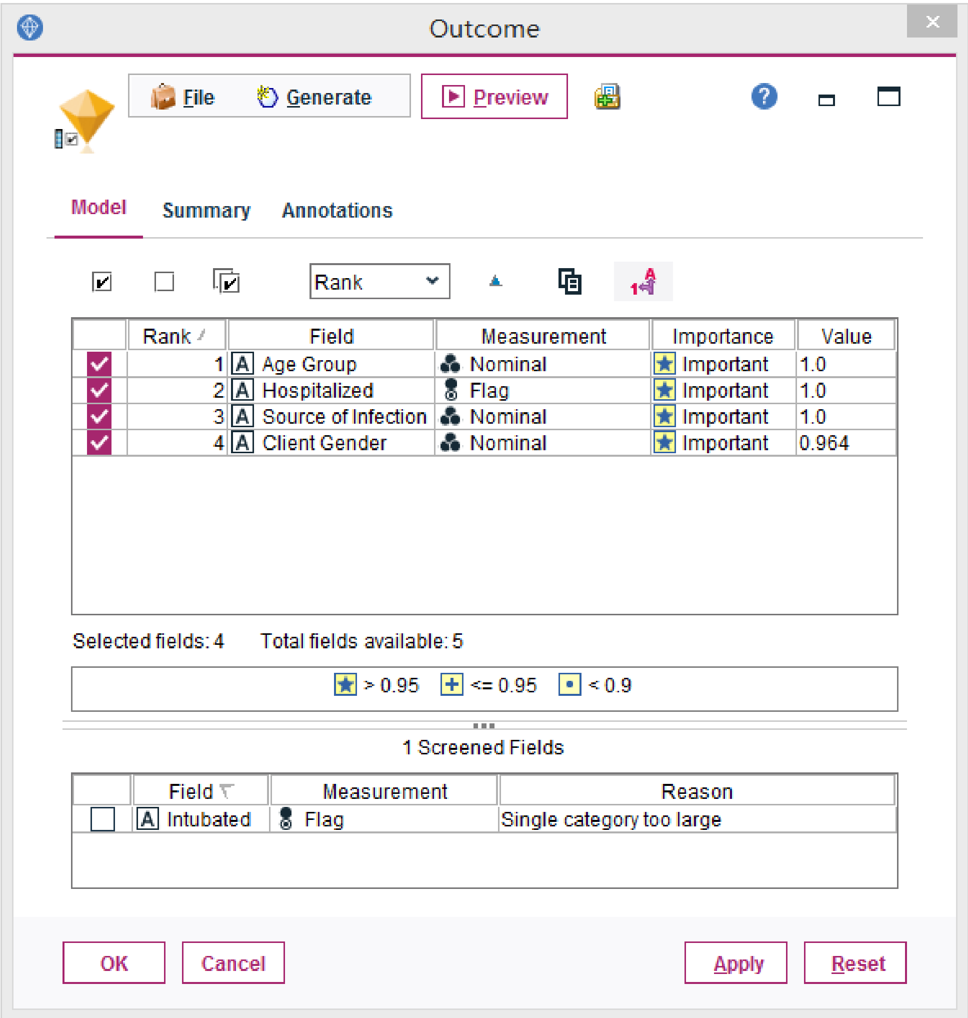


Figure17. Feature selection

4.1.2 Reduce unimportant attribute

Use the Outcome2 model to generate a filter which help to delete unimportant variation ‘Intubated’ selected during feature selection

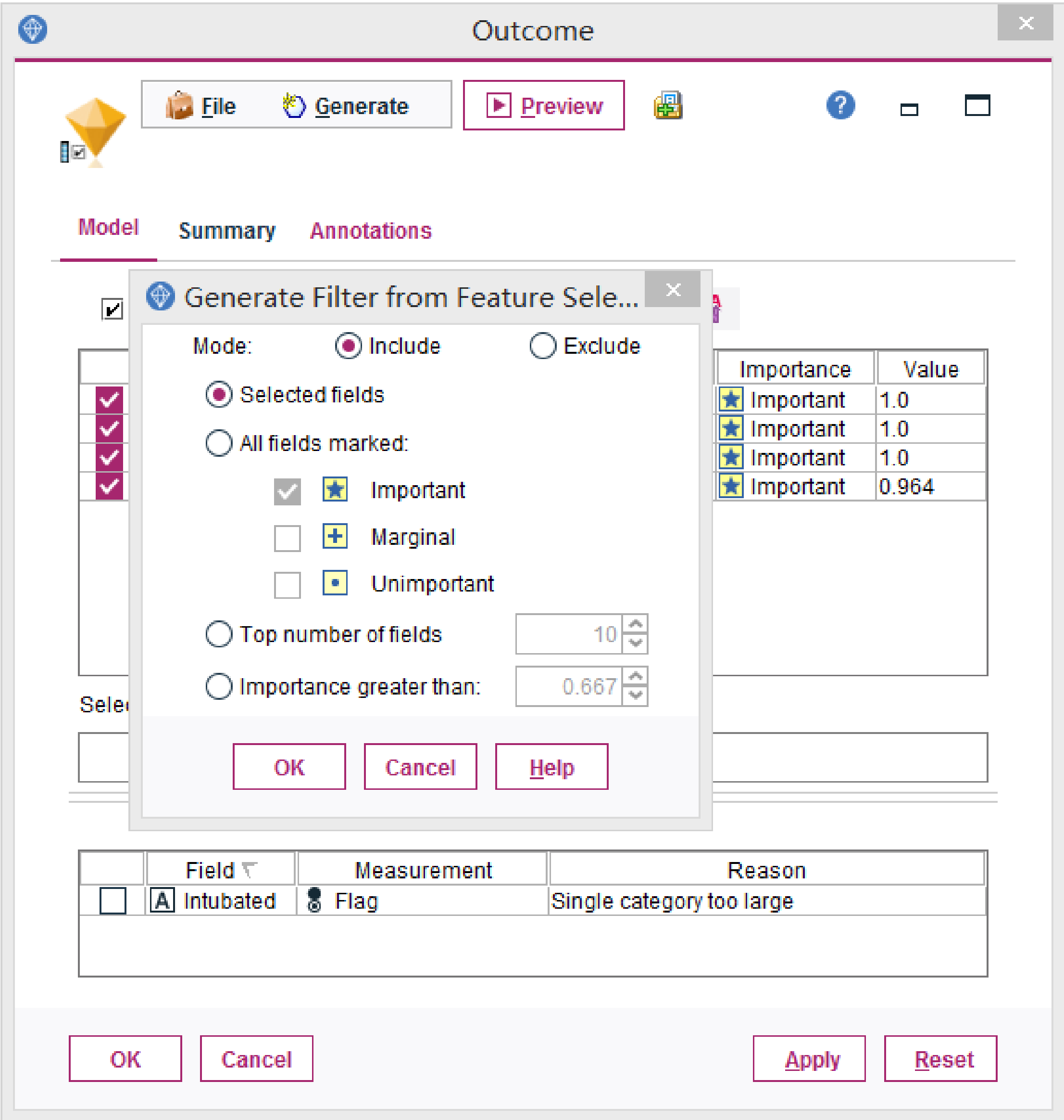


Figure18. reduce unimportant attribute

4.2 Project the Data

4.2.1 Distribution of target attribute

(1) Use the Distributed node to find the distribution of value ‘NONFATAL’ and ‘FATAL’ of attribute ‘Outcome2’ and select this attribute as the target one to display its distribution.

(2) The results are shown in the following figure. In the Outcome attribute, value ‘FATAL’ accounts for 7.52%, value ‘NONFATAL’ accounts for 92.48 %.

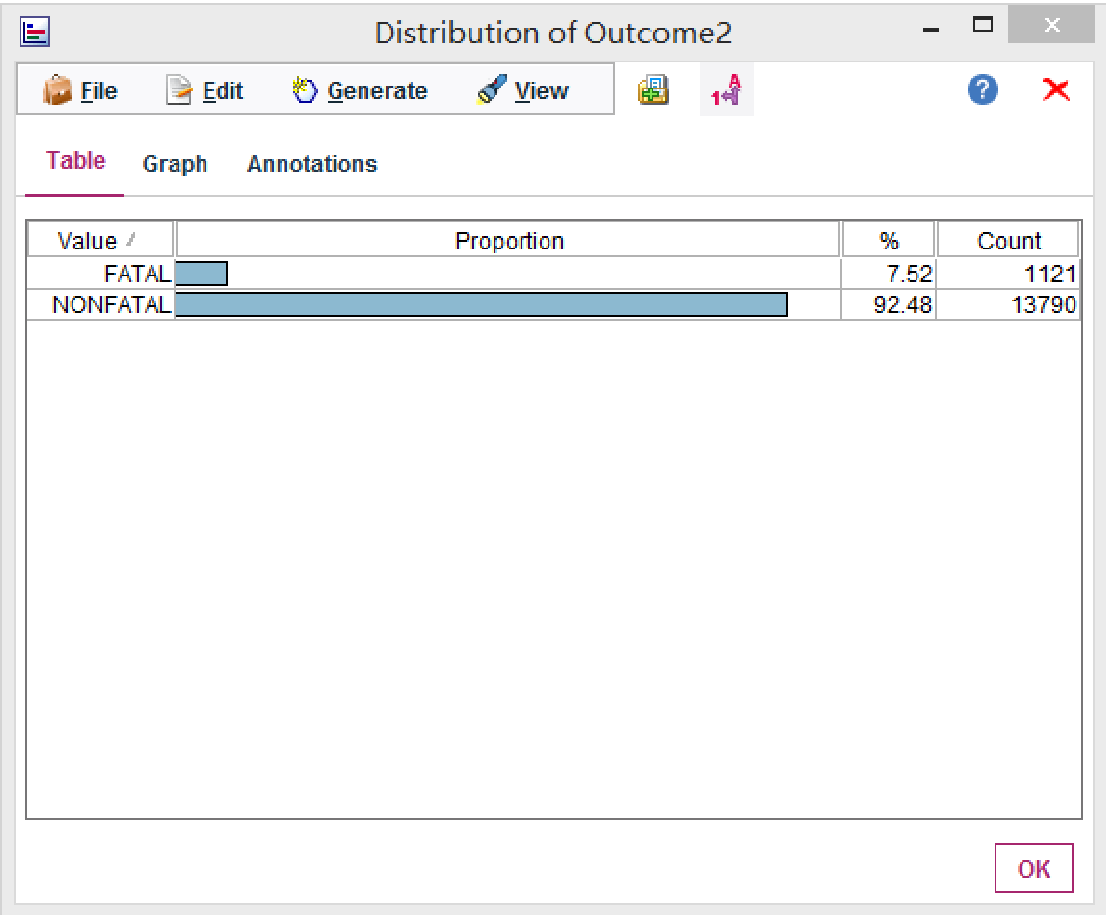


Figure19. Distribution of target attribute

4.2.2 Balance the Data

Select and apply generated balance node. Using the distribution node again, observe the data distribution of the target attribute, balanced Data is shown below, value ‘FATAL’ accounts for

50.05%, value ‘NONFATAL’ accounts for 49.95 %. Data balance is achieved.

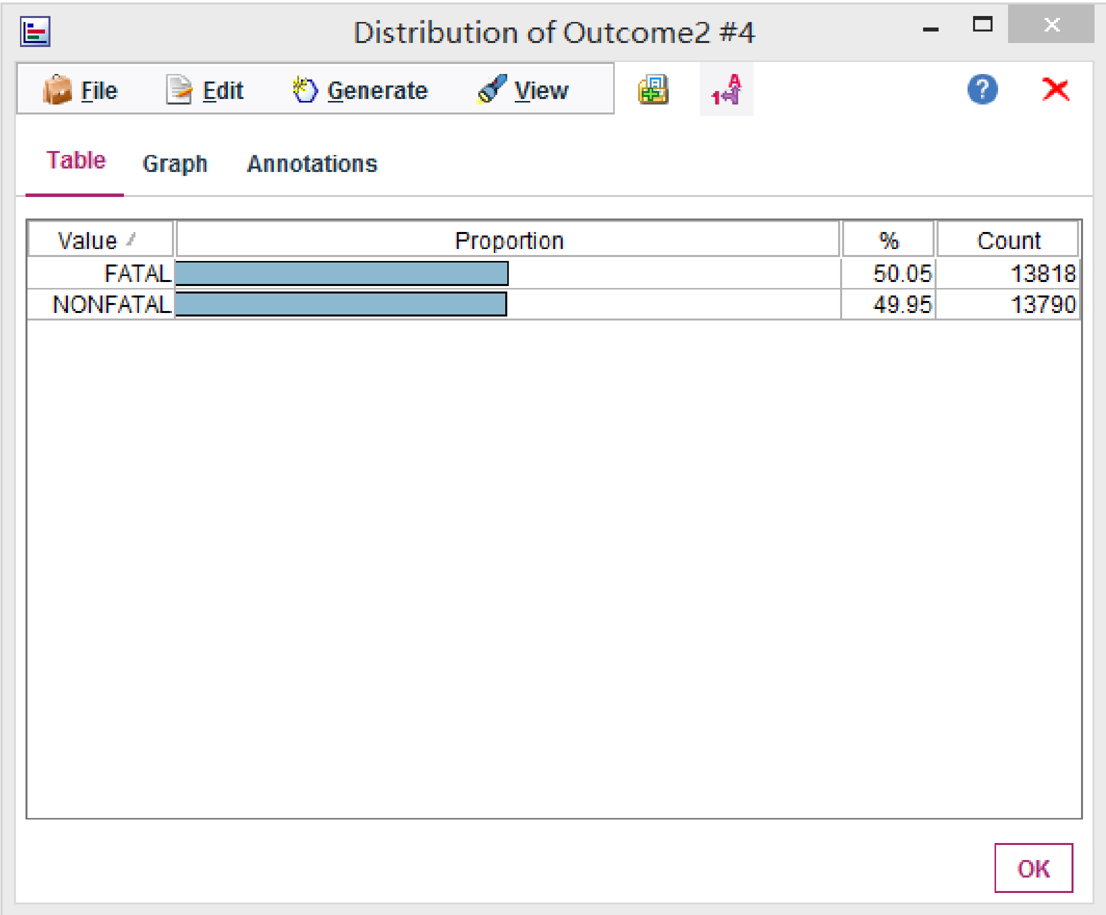


Figure 20. Distribution of target attribute after balance

4.2.3 Reclassified node to view the output

（1）Pick Client Gender as the variable I would like to view. There are five categories. I want to reclassify this variable so that there are only three categories.

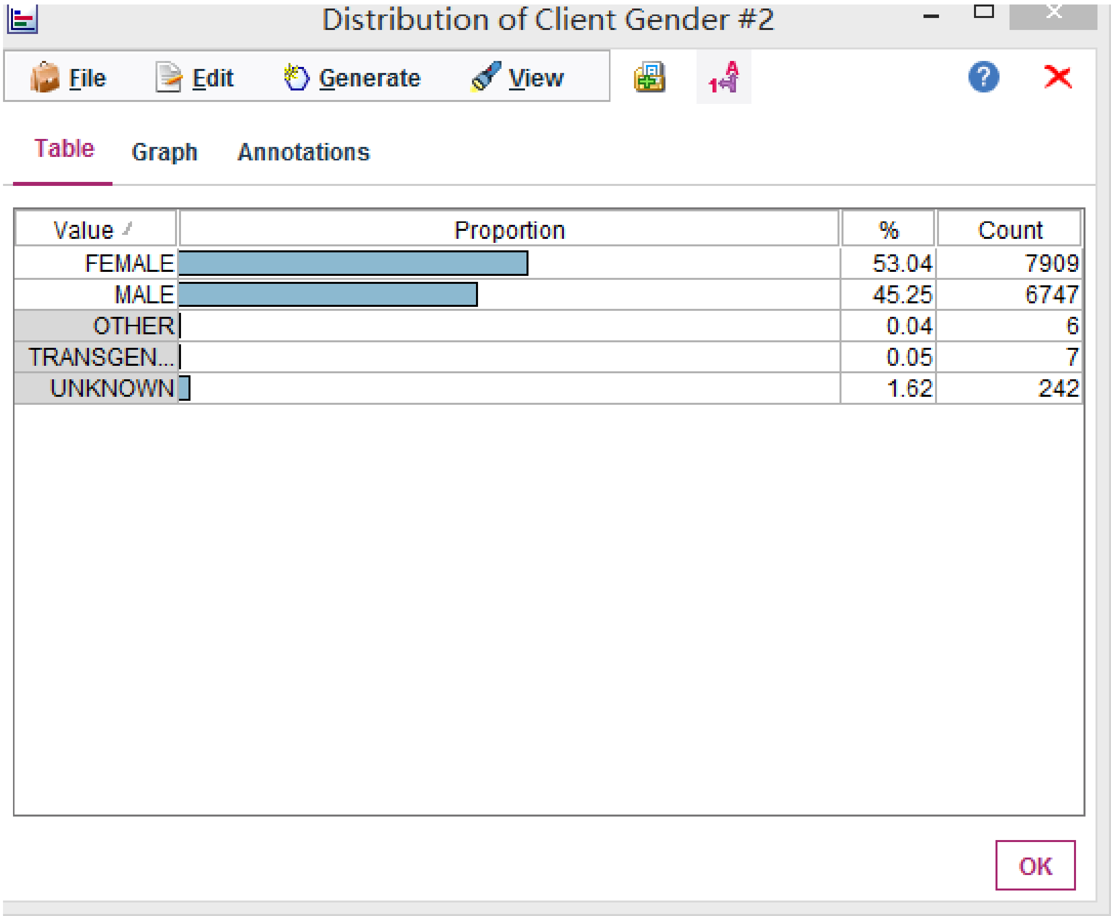


Figure21. Pick variable and group it

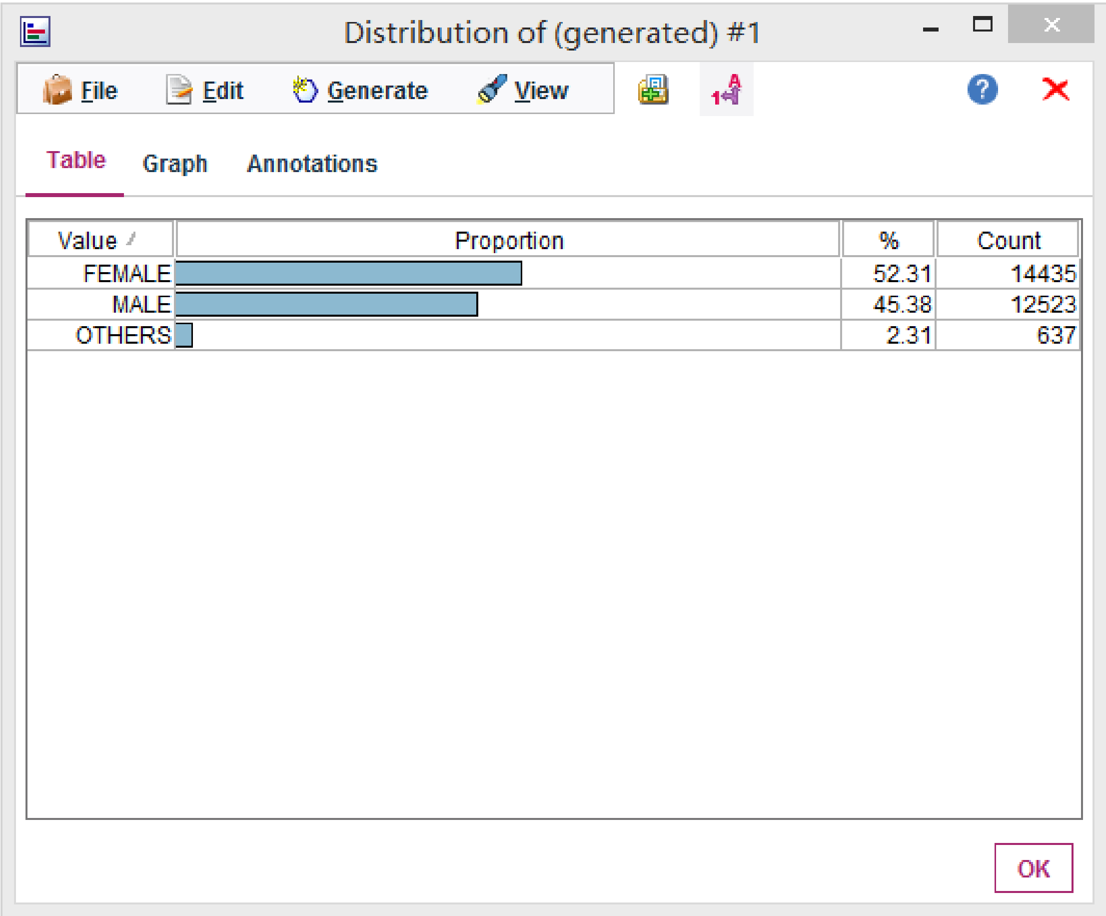
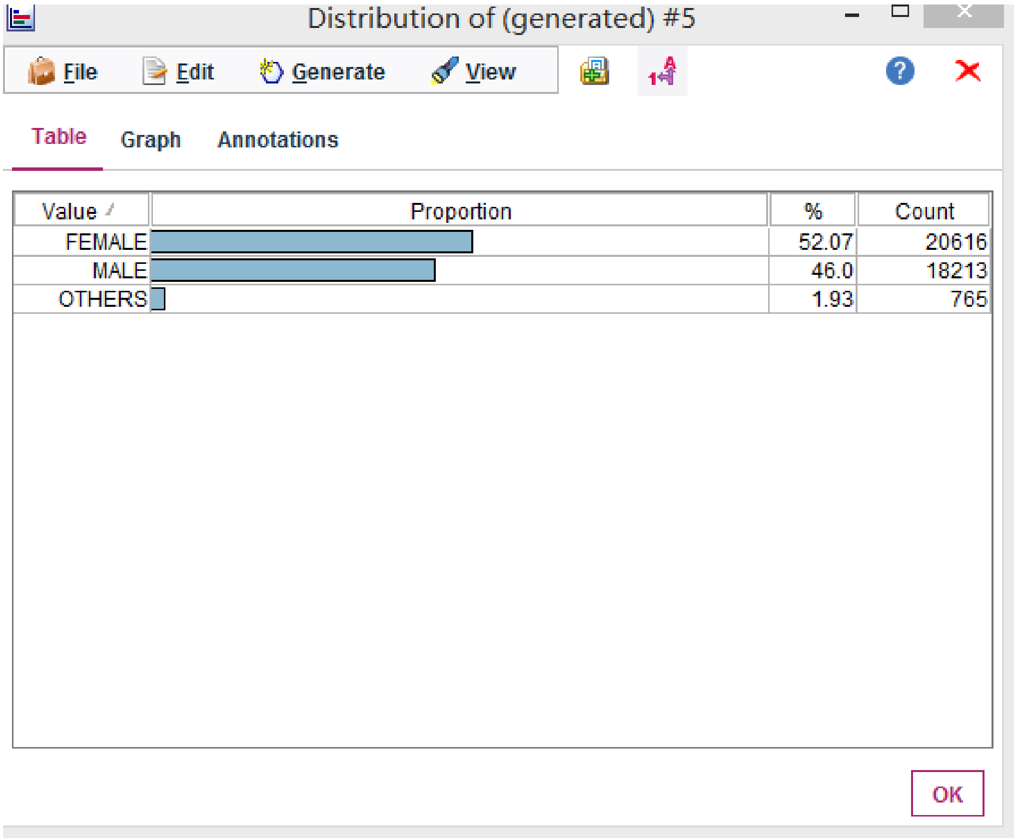


Figure22. Distribution after group

(2) Attach the Reclassification Node to the Generated Rebalancing Node. Attach a Distribution Graph to the reclassified node to view the output. Now, we can only see three groups, MALE, FEMALE and OTHERS.



Value	Proportion	%	Count
FEMALE		52.07	20616
MALE		46.0	18213
OTHERS		1.93	765

Figure 23. Distribution after rebalancing

5.1 Match and discuss the objectives of data mining to data mining methods

5.1.1 Supervised and Unsupervised Learning

(1) How to select learning Method

Whether the data set use labeled Data

(2) Supervised learning

The supervisory model uses the values of one or more input fields to predict the values of one or more output or target fields. Some examples of these techniques are: decision trees (C &R trees, CHAID, QUEST and

C5.0 algorithms), regression (linear, logical, generalized linear and Cox regression algorithms), neural networks, support vector machines, and Bayesian networks.

Oversight models help organizations predict known outcomes, such as whether customers will buy or leave, or whether transactions conform to known fraud patterns. Model technology includes machine learning, rule induction, subgroup identification, statistical method and multi model generation.

（3） Unsupervised learning

Unsupervised learning algorithm groups data into unlabeled data sets according to potential hidden features. Because there is no label, the results cannot be evaluated (this is the key difference between supervised learning algorithms). By grouping the data by unsupervised learning, you can learn about some raw data that may not be visible in other situations. In high-dimensional data sets or large data sets, this problem is more obvious (Jones,2017).

5.1.2 Classification, Association and Segmentation

(1) Classification: Find the common characteristics of a group of data objects in the database and divide them into different categories according to the classification mode. The purpose is to map the data items in the database to a given category through the classification model (Data for Data Mining, n.d.).

(2) Association: The association model finds a pattern in the data in which one or more entities (such as events, purchases, or properties) are associated with one or more other entities. The model builds the set of

rules that define these relationships. Here, the fields in the data can be either input or target. You can find these associations manually, but the association rules algorithm finds them faster and explores more complex patterns. (IBM Knowledge Center, n.d.).

(3) Segmentation: The segmentation model divides the data into segments or clusters with similar input field patterns. Because they are only interested in input fields, the segmentation model has no concept of output fields or target fields. Examples of segmentation models include Kohonen network, K-means clustering, two-step clustering and anomaly detection (IBM Knowledge Center, n.d.) Subdivision models (also known as "clustering models") are useful when specific results are unknown (for example. The clustering model focuses on identifying groups of similar records and labeling them according to the group they belong to. This is done without prior knowledge of the group and its characteristics, and it distinguishes the clustering model from other modeling techniques because the model does not have predefined output or target fields for prediction (IBM Knowledge Center, n.d.).

5.2 Select the appropriate data-mining method (s) based on discussion

5.2.1 Choose supervised learning

Because there are both input variables and output variables (Outcome), so choose to use supervised learning methods. Because the prediction target is continuous, the regression method is appropriate.

5.2.2 Choose classification

Classification can be used for forecasting. The purpose of classification is to automatically derive the general description of given data from historical data records, so as to predict future data. Different from regression, the output of classification is discrete, while the output of regression is continuous.

Because the target attribute Outcome is categorical value, so we choose the classification model (IBM Knowledge Center, n.d.).

6.1 Conduct exploratory analysis and discuss

6.1.1 Algorithm discussion

(1) C & R tree

Requirements: The C & R tree model requires one or more input fields and one target field. The target field and input field can be continuous (numeric range) or classified. Fields set to all or none are ignored. Fields used in the model must fully instantiate their type, and any ordinal (ordered set) fields used in the model must have a numeric store (not a string) (IBM Knowledge Center, n.d.).

Strengths: C & R tree model is very powerful when there are problems such as data loss and a large number of fields. They usually don't need a long training time to estimate. In addition, the C & R tree model is easier to understand than some other model types - rules derived from the model have very simple explanations. Unlike C5.0, C & R trees can accommodate continuous and classified output fields(IBM Knowledge Center, n.d.).

(2) CHAID

Requirements: The target field and input field can be continuous or classified. Nodes can be divided into two or more subgroups at each level. Any ordinal field used in the model must have a numeric store (not a string). If necessary, you can use the reclassification node to transform it (IBM Knowledge Center, n.d.).

Strengths: Unlike C & R trees and quest nodes, CHAID can generate non binary trees, which means that some splits have more than two branches. Therefore, it tends to create larger trees than the binary growth method. CHAID applies to all types of input and accepts case weight and frequency variables (IBM Knowledge Center, n.d.).

(3) QUEST

Requirements: The input fields can be contiguous (numeric range), but the target fields must be classified. All splits are binary. The weight field cannot be used. Any ordinal (ordered set) fields used in the model must have a numeric store (not a string).If necessary, you can use the Reclassification node to transform it(IBM Knowledge Center, n.d.).

Strengths: Quest uses a series of rules based on the importance test to evaluate the input fields on the node. For selection purposes, you may only need to perform a test once on each input on the node. Unlike C & R trees, all splits are not checked; unlike C & R trees and CHAID, category combinations are not tested when evaluating input fields for selection. This can speed up the analysis (IBM Knowledge Center, n.d.).

(4) C 5.0

Requirements: To practice the C5.0 model, there must be a classified (i.e., nominal or ordered) target field and one or more input fields of any type. Fields set to all or none are ignored. Fields used in models must fully instantiate their types. You can also specify a weight field (IBM Knowledge Center, n.d.).

Strengths: The C5.0 model is very powerful when there are problems such as data loss and a large number of input fields. They usually don't need a long training time to estimate. In addition, because the rules derived from this model have very direct interpretations, the C5.0 model is easier to understand than some other model types. C5.0 also provides a powerful enhancement method to improve the accuracy of classification (IBM Knowledge Center, n.d.).

(5) Bayesian Network

Requirements: The target field must be classified and can have nominal, *Ordinal* or tagged measurement levels. The input can be any type of field. Continuous (numeric range) input fields are automatically discarded; however, if the distribution is skewed, you can get better results by manually binding fields with a binding node before the Bayesian network node.

Strengths: It helps you understand causality. As a result, it enables you to understand the problem area and predict the consequences of any intervention. The network provides an effective method to avoid data over fitting. It is easy to observe a clear visualization of the relationships involved (IBM Knowledge Center, n.d.).

6.2 Select data-mining algorithms based on discussion

6.2.1 Algorithm requirements :

(1) Objective: Find out the relationship between the target attribute and the predictors, and use it to predict the probability of the target variable occurring.

(2) Target attribute requirement: flag.

(3) Data type requirements: nominal and flag(string)

(4) Result requirements: prefer model results which are easy to demonstrate

requirements / model	Objective	Target attribute	Input data type	Result
	relationship between the target and predictors	flag	nominal and flag(string)	easy to demonstrate

CHAID	√	×	×	√
QUEST	√	×	×	√
C 5.0	√	√	√	√
Bayesian Network	√	√	√	√
C & R tree	√	×	×	√

Table 3

6.2.2 Select data-mining algorithms

Through the above analysis, considering the data characteristics of this data set, which is the some attributes contain more than two variables. So we need to use the Multiple Classification method. finally choosing two classification models, C5.0 and Bayesian Network .

6.3 Select appropriate model(s) and choose relevant parameter(s)

6.3.1 C 5.0 decision tree model

(1) Build C5.0 decision tree model

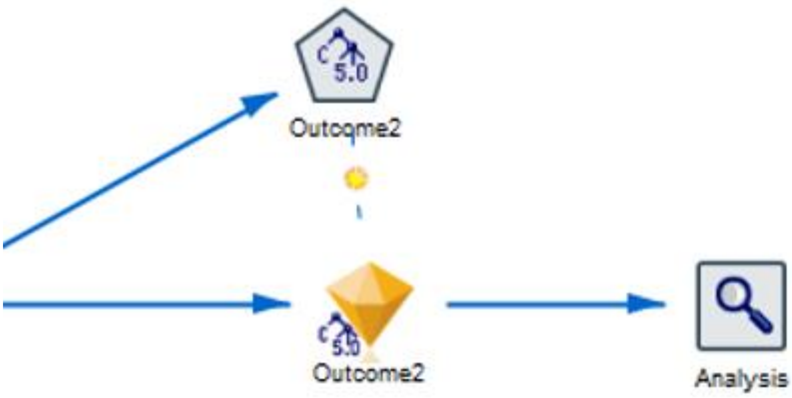


Figure 24. C5.0 decision tree model

(2) Set relevant parameters

If winnow attributes is selected, C5.0 will check the usefulness of the predictor before starting to build the model. Unrelated predictors are excluded from the model building process. This option is useful for models with many prediction fields and can help prevent over fitting.

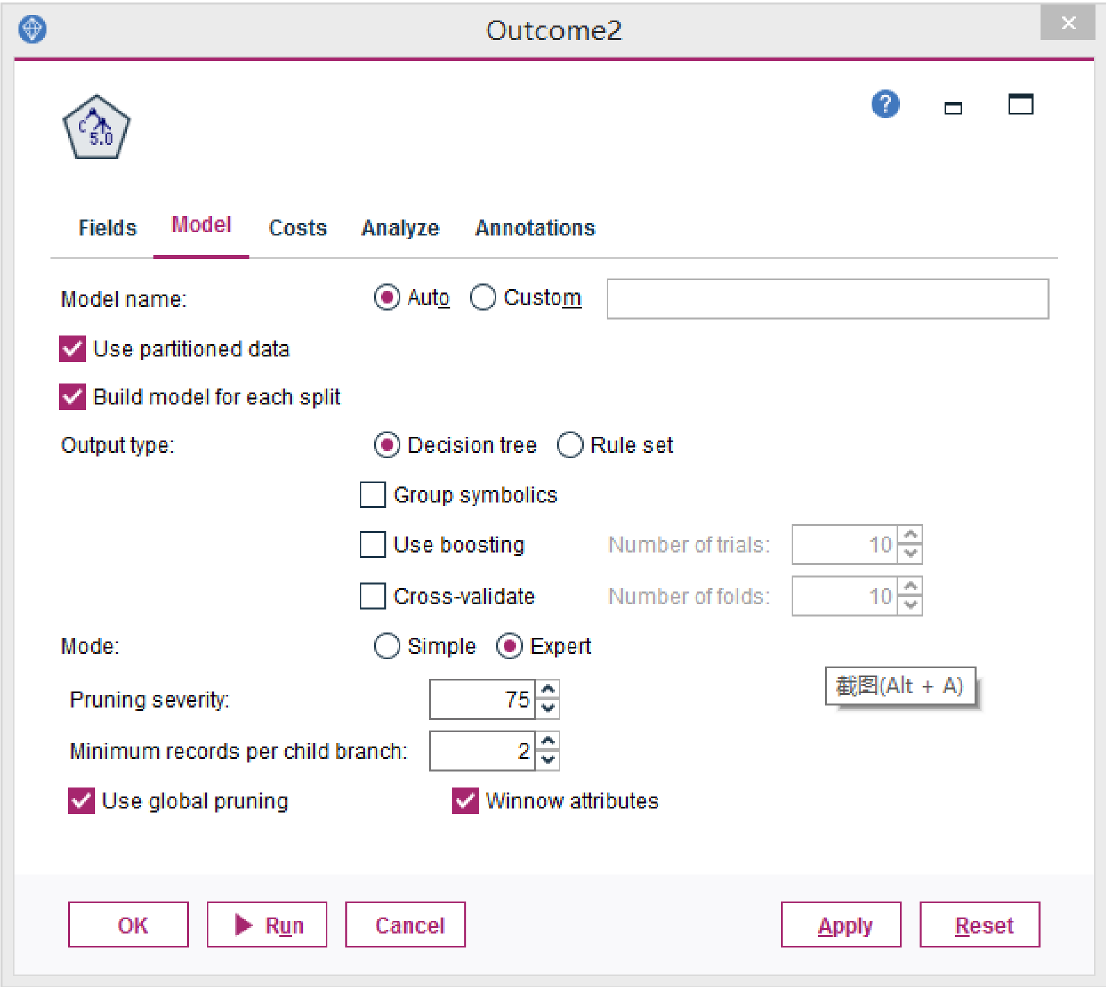


Figure 25. parameter setting

6.3.2 Bayesian Network model

(1) Build Bayesian Network model

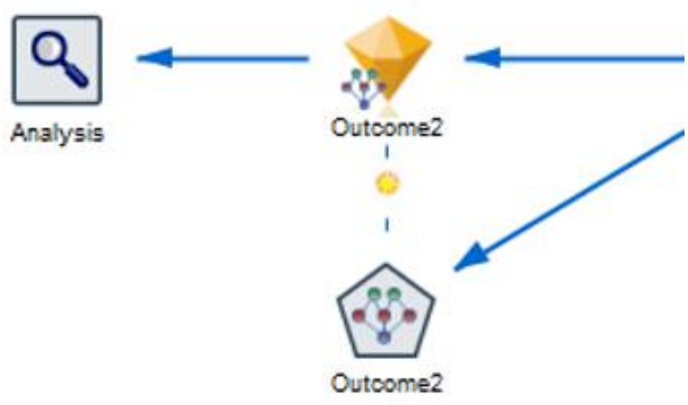


Figure 26. Bayesian Network model

(2) Set relevant parameters

Calculate raw propensity scores. For a model with a flag target Outcome2, we can request a propensity score to indicate the likelihood of a real result specified for the target field.

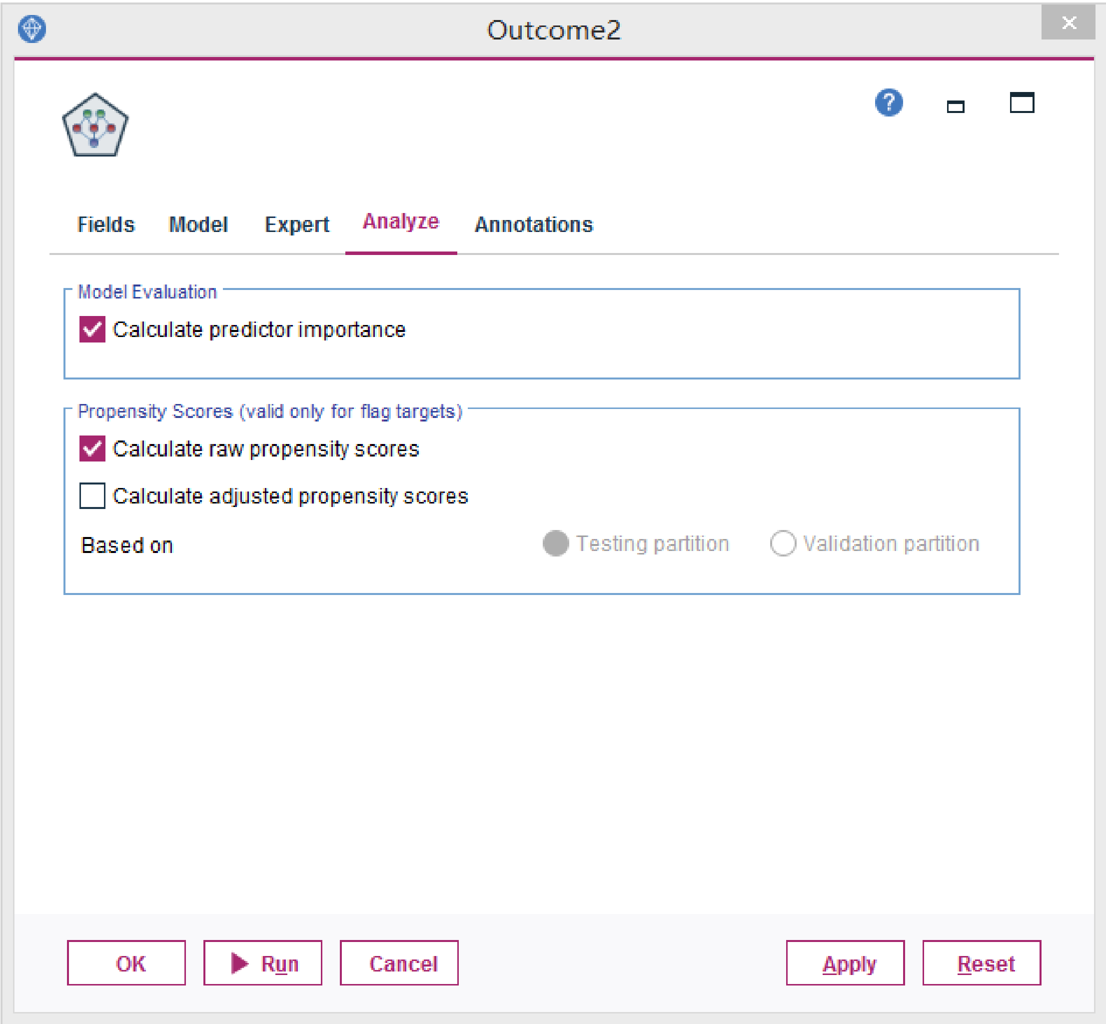


Figure 27. setting parameters

7.1 Logical test designs

7.1.1 create logical test design

(1) When an algorithm use a limited data set to search for the best parameters for a particular model, it may model not only the general pattern in the data, but also any noise specific to that data set, resulting in poor performance of the model on the test data, that is, overfitting (Usama, Gregory & Padhraic,1996). So we chose cross-validation to solve this problem

(2) Due to the small number of data samples and the need for sufficient data to test results, we set 70% training set and 30% testing set. Use partition node to divide the data into 70% training data and 30% test data

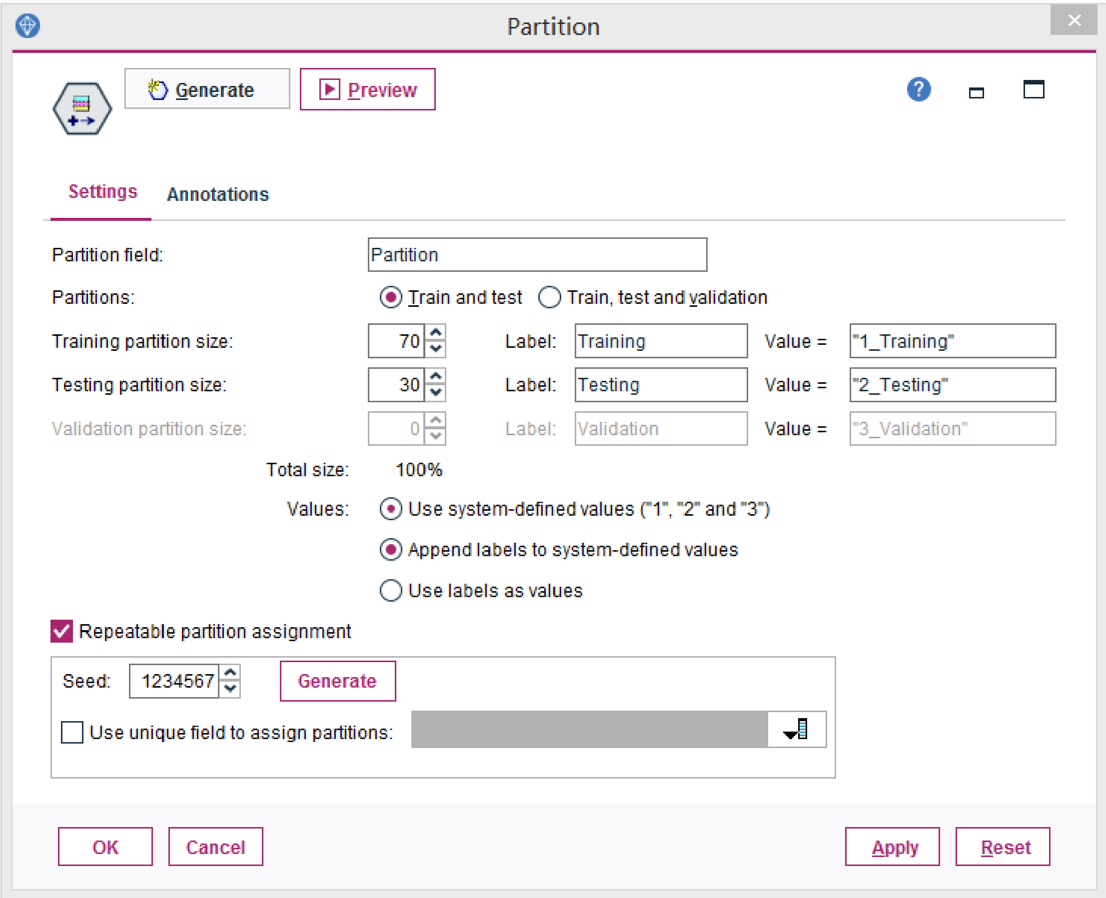
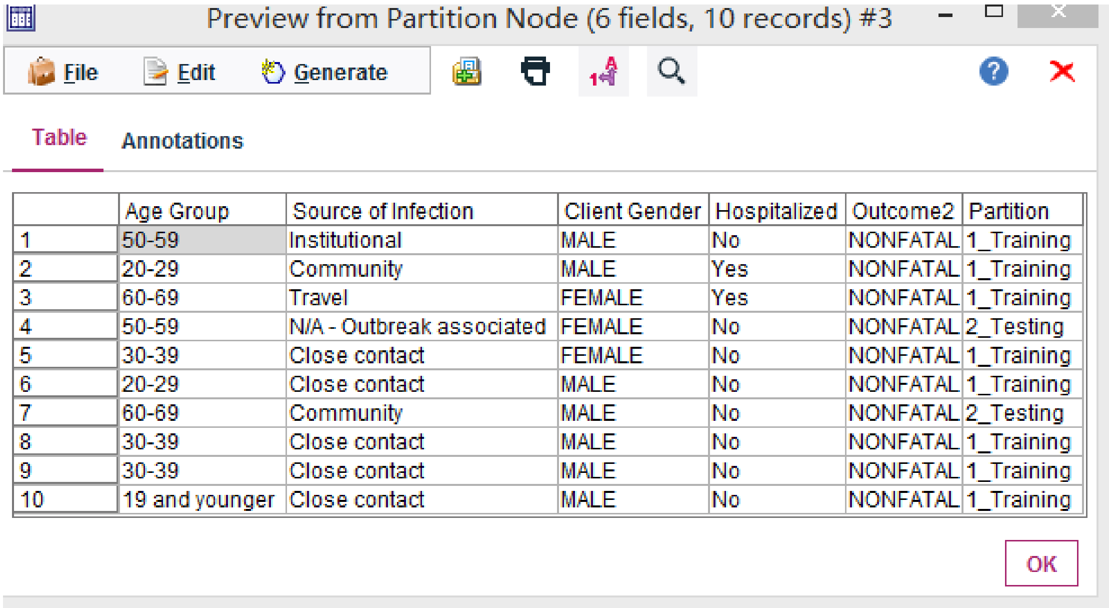


Figure 28. training data and test data

(3)"1_Training" represents the training group and "2_Testing" represents the training group



	Age Group	Source of Infection	Client Gender	Hospitalized	Outcome2	Partition
1	50-59	Institutional	MALE	No	NONFATAL	1_Training
2	20-29	Community	MALE	Yes	NONFATAL	1_Training
3	60-69	Travel	FEMALE	Yes	NONFATAL	1_Training
4	50-59	N/A - Outbreak associated	FEMALE	No	NONFATAL	2_Testing
5	30-39	Close contact	FEMALE	No	NONFATAL	1_Training
6	20-29	Close contact	MALE	No	NONFATAL	1_Training
7	60-69	Community	MALE	No	NONFATAL	2_Testing
8	30-39	Close contact	MALE	No	NONFATAL	1_Training
9	30-39	Close contact	MALE	No	NONFATAL	1_Training
10	19 and younger	Close contact	MALE	No	NONFATAL	1_Training

Figure 29. training data and test data preview

7.2 Data mining must be conducted (the model must run).

7.2.1 Run C5.0 model and Bayesian Network model

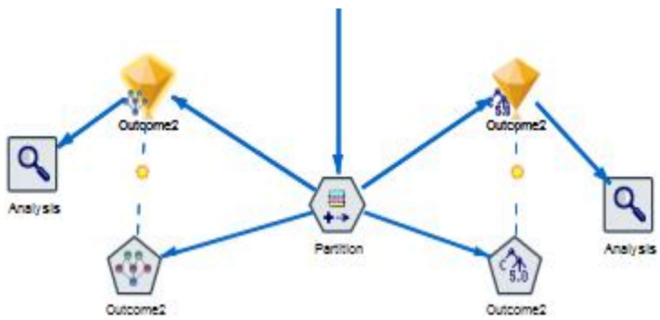


Figure 32. C5.0 model and Bayesian Network model

(1) result and predictor importance

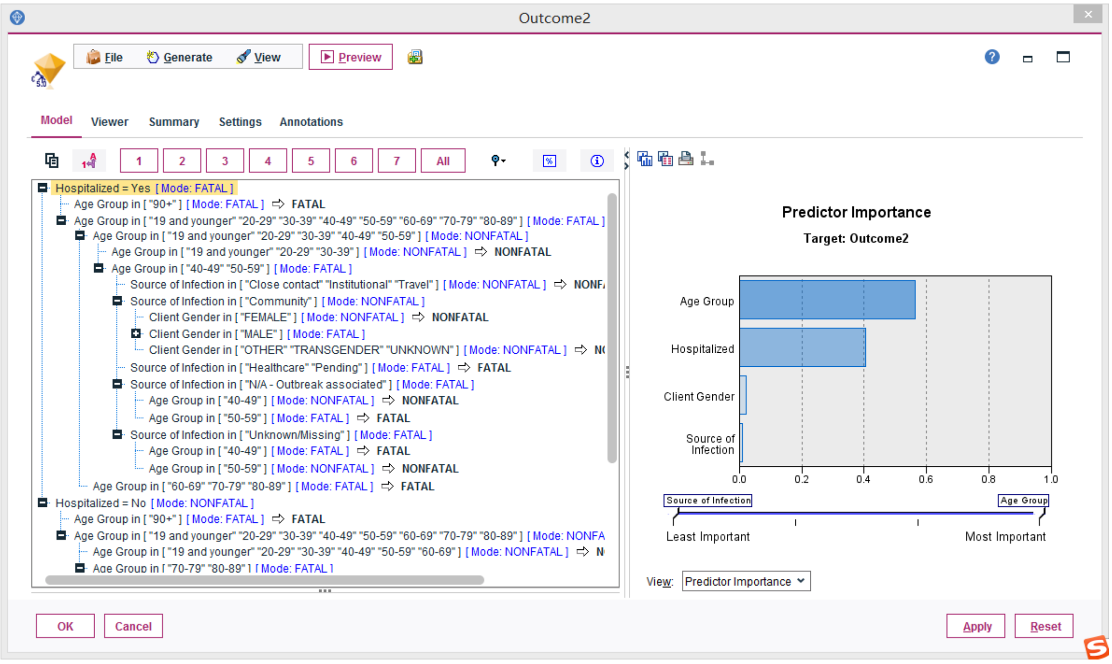


Figure 30. C5.0 model result

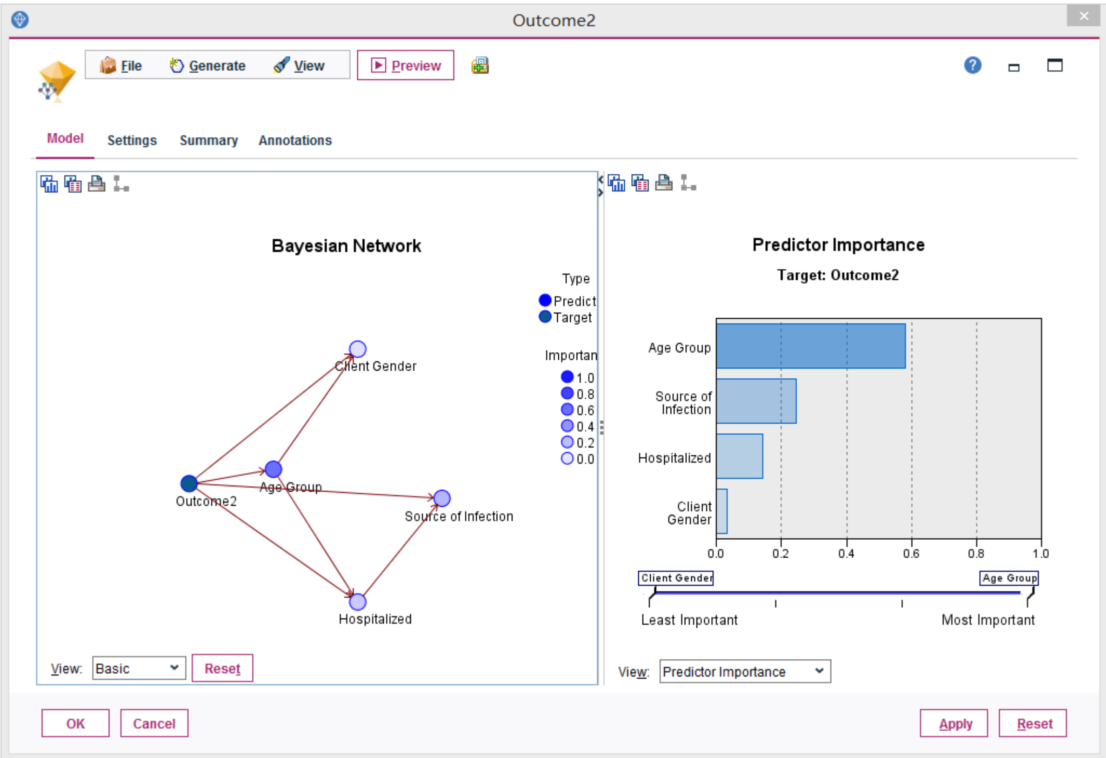


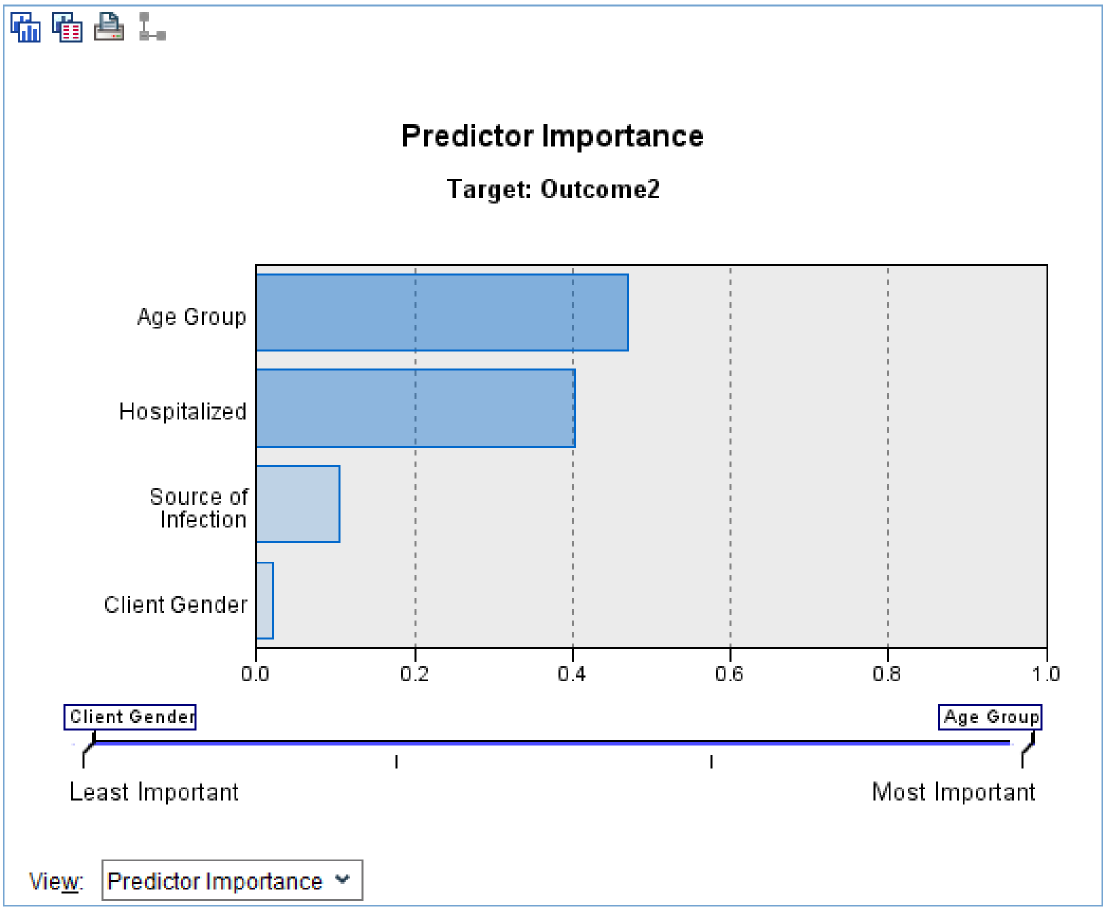
Figure 31. Bayesian Network model result

7.3 Search for patterns and document the model's output.

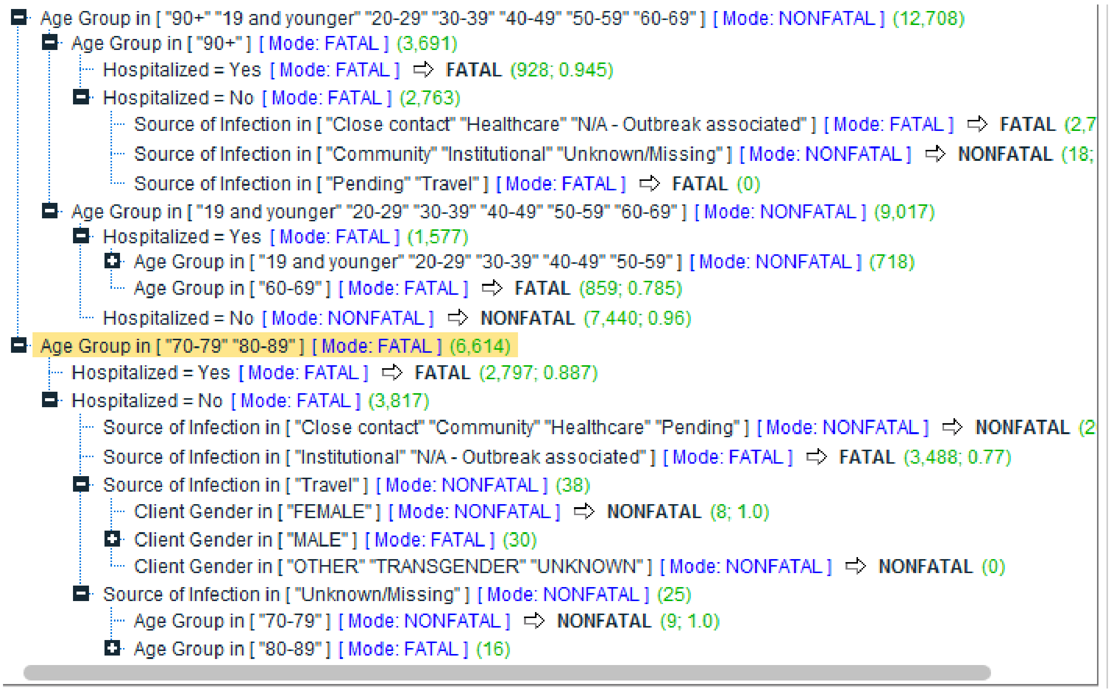
There are many patterns to choose from. In order to make the model as simple and understandable as possible, I only chose two models.C5.0 and Bayesian Network Model, both of which meet the requirements of the data set, cooperate with each other so that I have a lot of new discoveries.

C5.0 model Bar chart of predictor importance

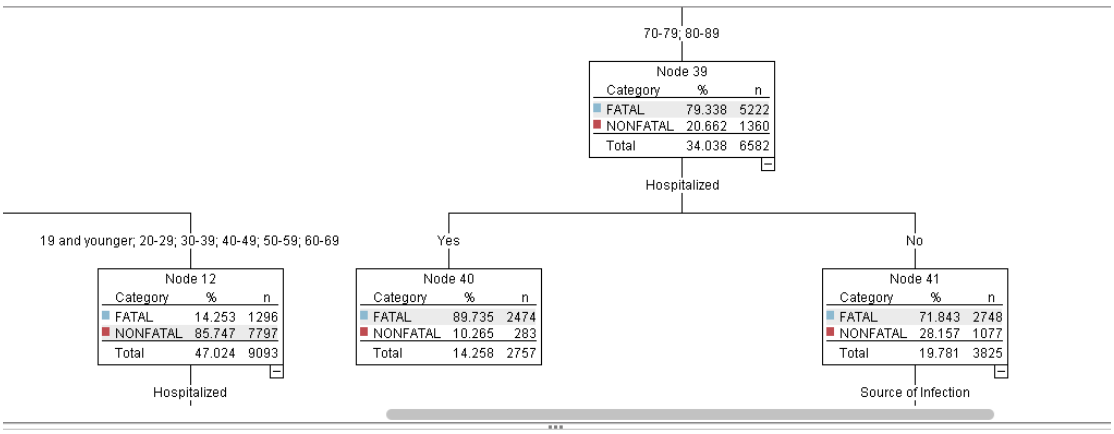
The Age Group was the most important model in C 5.0, with Hospitalized following.



- Text tree

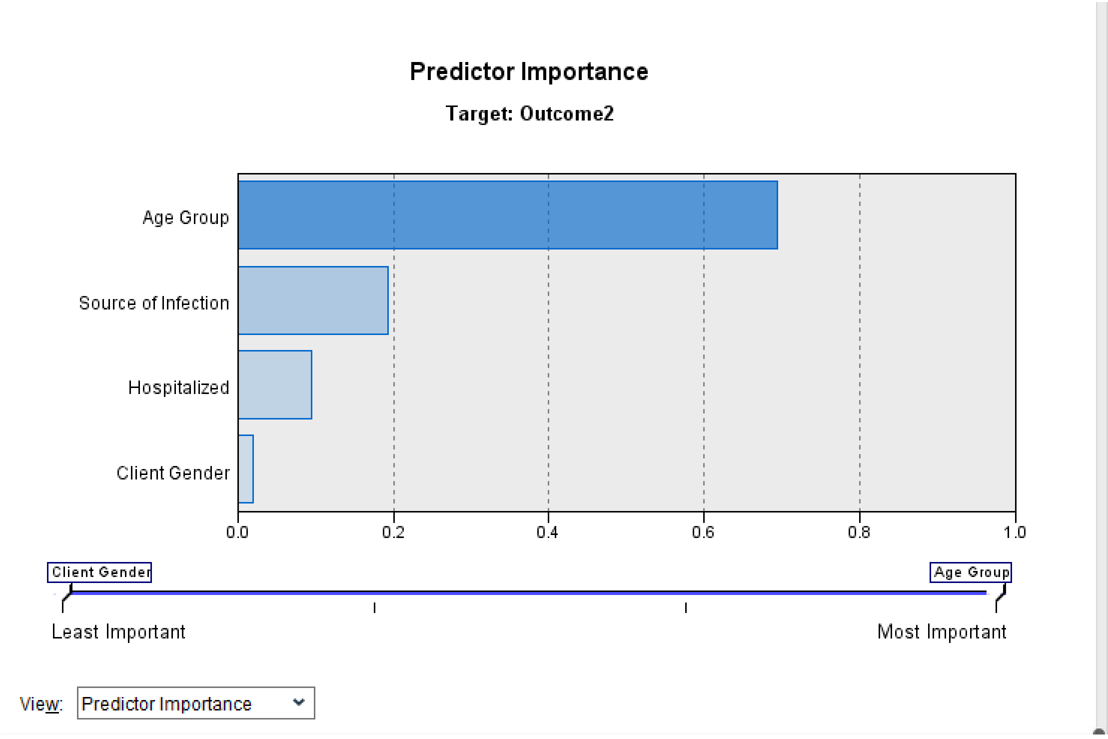


- Tree map

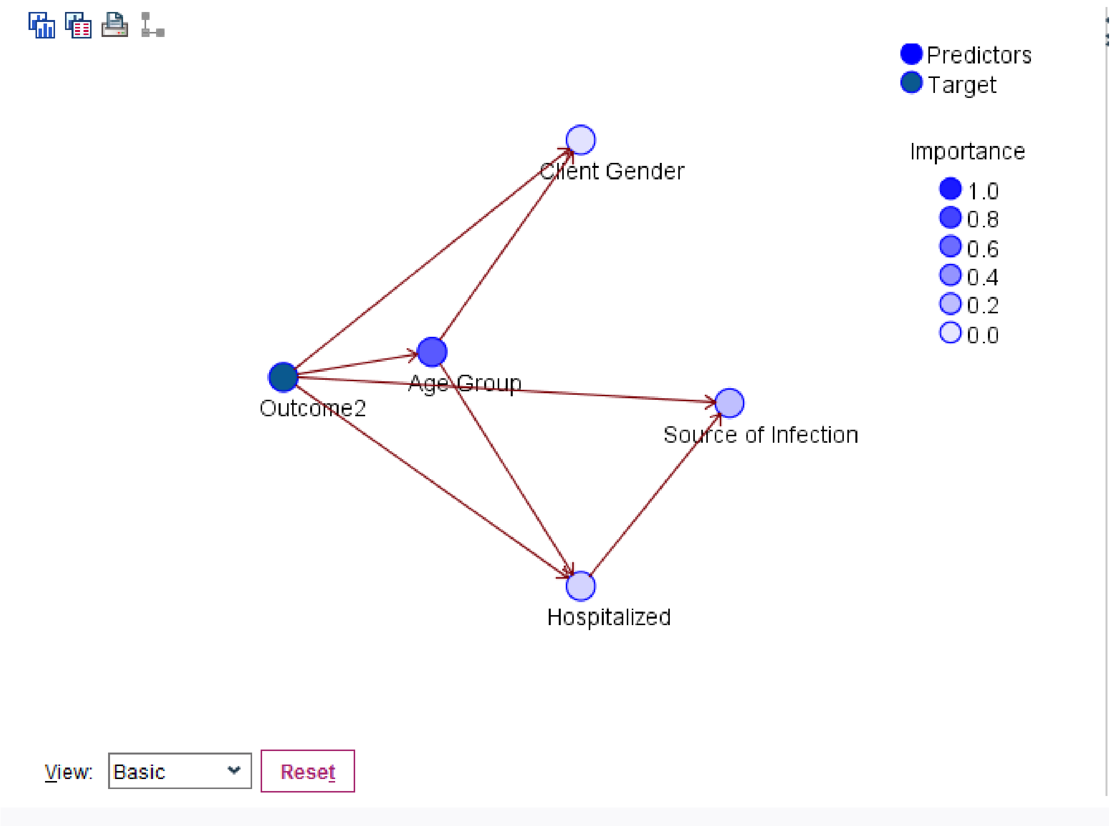


(2) Bayesian Network Model,

Bar chart of predictor importance



Network graph



Conditional Probabilities of Age Group

Parents	Probability								
Outcome2	19 and younger	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90+
NONFATAL	0.07	0.16	0.16	0.15	0.16	0.11	0.06	0.08	0.05
FATAL	0.00	0.00	0.00	0.01	0.03	0.09	0.18	0.36	0.33

8.1 Study and discuss the mined patterns.

Carry out an in-depth discussion about the data, results, models and patterns.

8.1.1 data and result

The data range is only from the Toronto area, and the geographical scope is relatively small. In addition, due to the different cultural habits, living habits and protection measures for coVID-19, the infected population may be different. The outcome of the forecast may change depending on the location. The selection of data increases the spatial limitation of the prediction.

In addition, this data set mainly includes data from January 2020. In the months when the epidemic is very severe, data statistics may be affected by the epidemic, resulting in information loss or incorrect data input, which may increase the error.

The projections clearly show that mortality rates are very high for older coVID-19 cases, especially those between the ages of 70 and 90.But the number of deaths over the age of 90 is low, perhaps because the sample of cases in their 90s is already small.

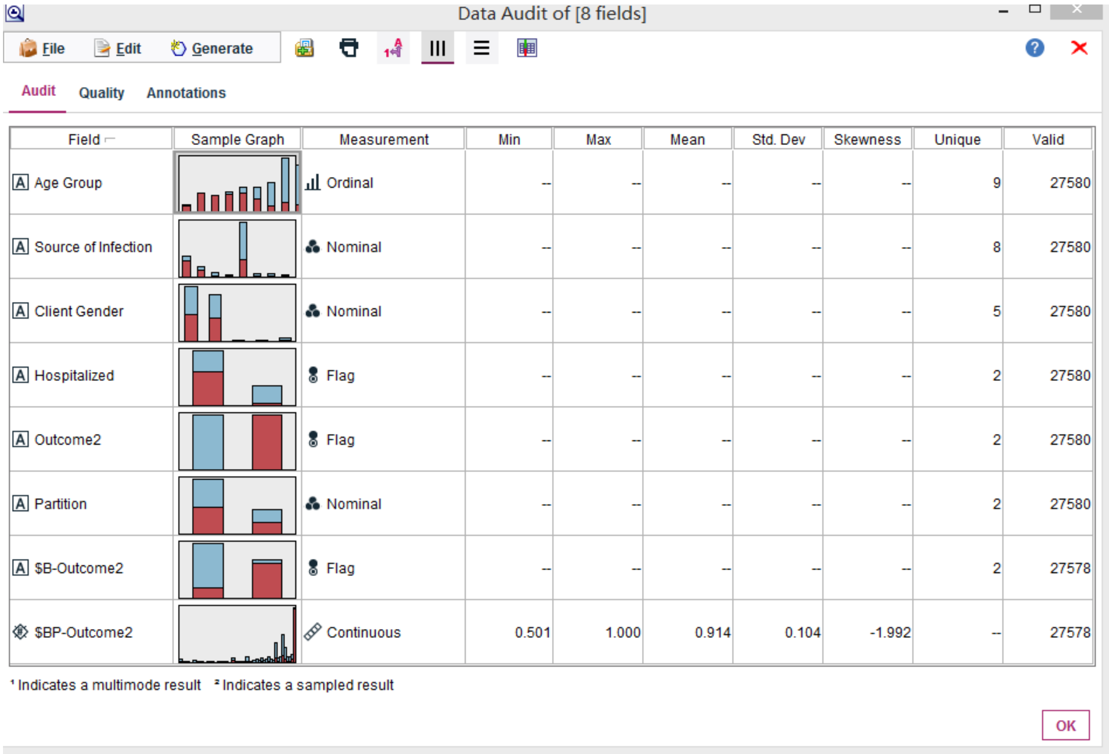
```
Age Group in [ "90+" "19 and younger" "20-29" "30-39" "40-49" "50-59" "60-69" ] [ Mode: NONFATAL ] (12,708)
Age Group in [ "70-79" "80-89" ] [ Mode: FATAL ] (6,614)
```

8.1.2 models and patterns

My data set is dominated by Oridinal, Nominal and Flag data and the target value is a value in a flag format. Identify the factors that influence the mortality or survival of coVID-19 by distinguishing between the high mortality and the low mortality groups. This can help predict future patterns, predict which populations are vulnerable, and help people increase their survival rates. So I chose Classification. C 5.0 decision Tree Model and Bayesian Network Model all meet my data requirements very well, and there is no operation error in the operation of these two models.

8.2 - Visualize the data, results, models and patterns in a clear and effective manner.

data



Data Audit of [8 fields]

File Edit Generate

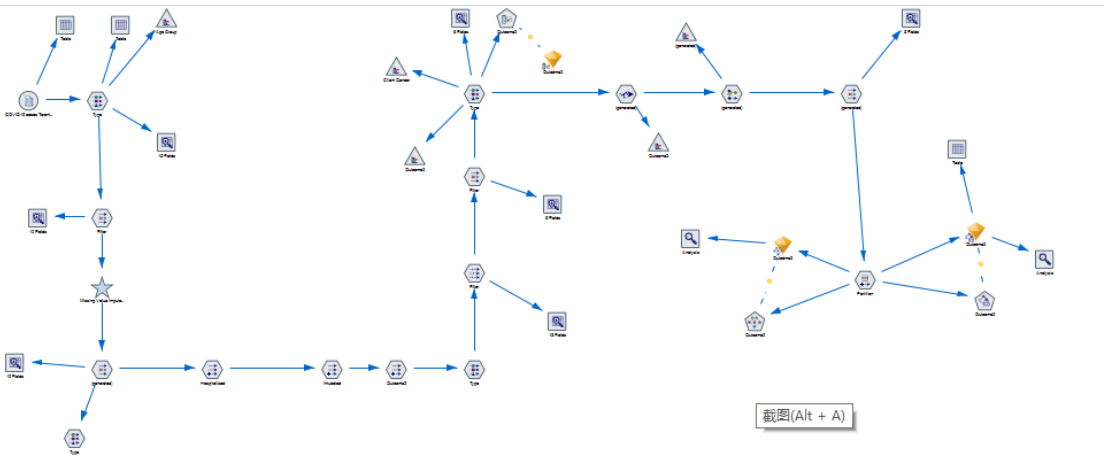
Audit Quality Annotations

Field	Sample Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
Age Group		Ordinal	--	--	--	--	--	9	27580
Source of Infection		Nominal	--	--	--	--	--	8	27580
Client Gender		Nominal	--	--	--	--	--	5	27580
Hospitalized		Flag	--	--	--	--	--	2	27580
Outcome2		Flag	--	--	--	--	--	2	27580
Partition		Nominal	--	--	--	--	--	2	27580
$B-Outcome2		Flag	--	--	--	--	--	2	27578
$BP-Outcome2		Continuous	0.501	1.000	0.914	0.104	-1.992	--	27578

[1] Indicates a multimode result [2] Indicates a sampled result

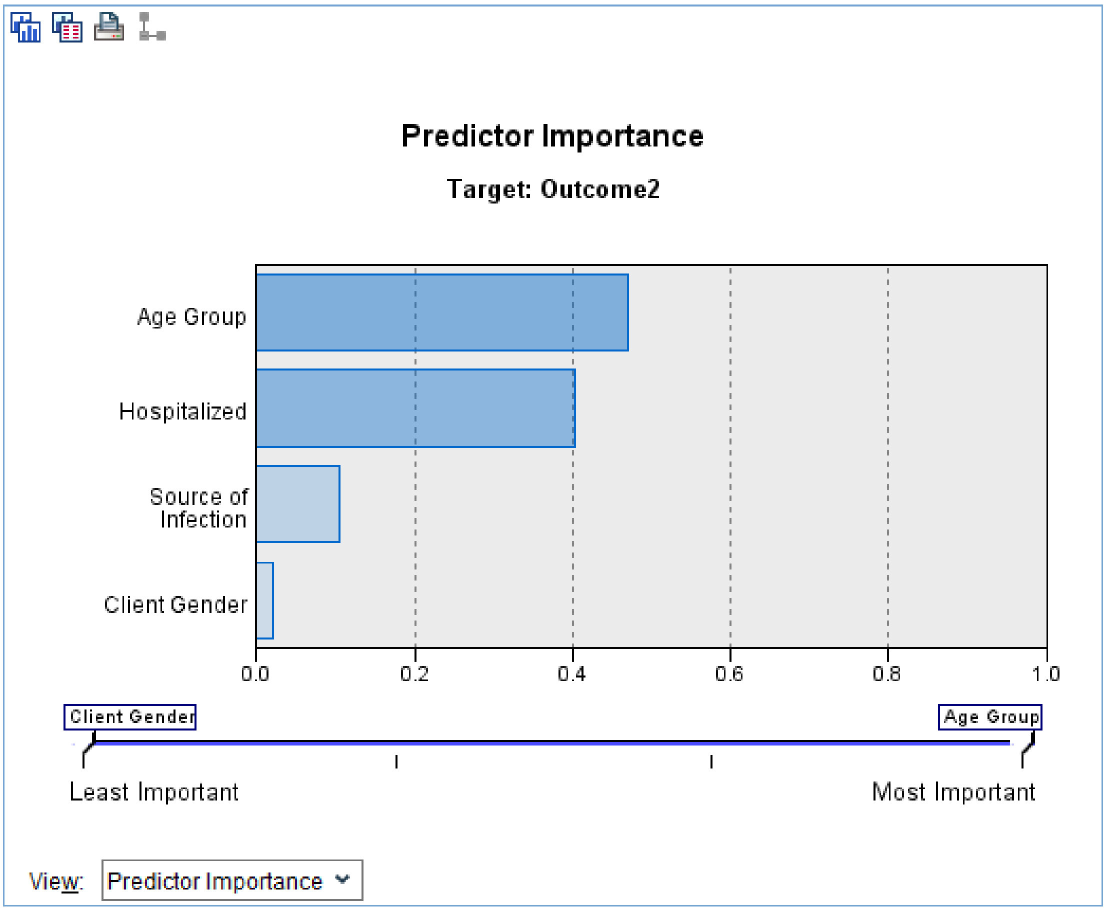
OK

Model



8.2.1 C5.0 model

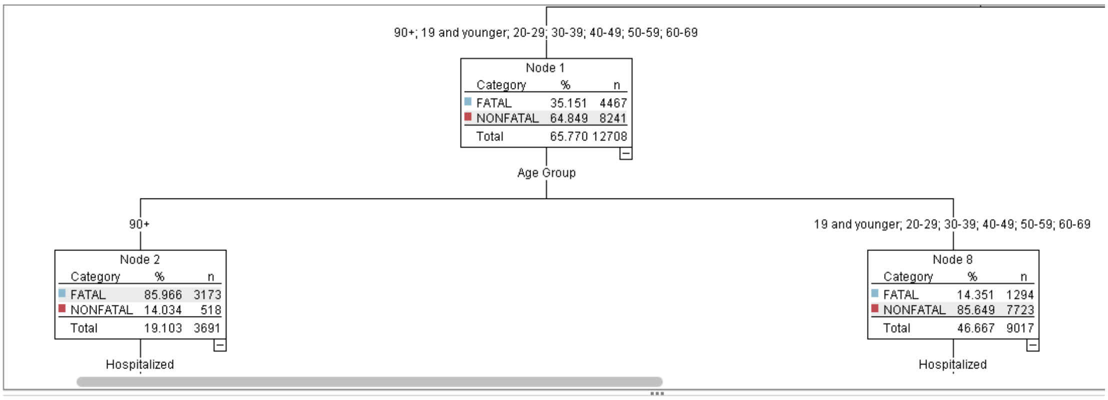
Bar chart of predictor importance



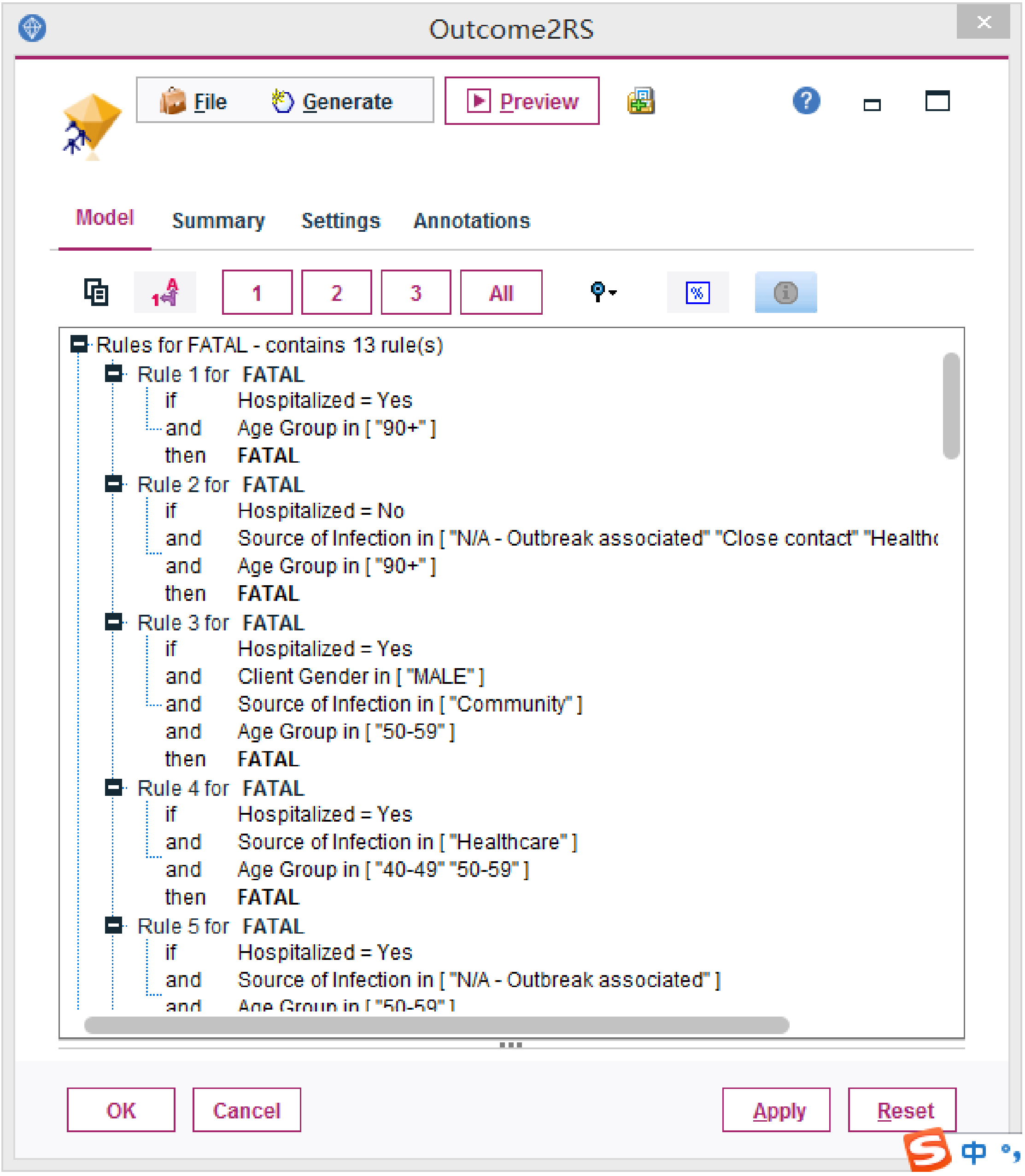
Text tree

```
Age Group in [ "90+" "19 and younger" "20-29" "30-39" "40-49" "50-59" "60-69" ] [ Mode: NONFATAL ] (12,708)
  Age Group in [ "90+" ] [ Mode: FATAL ] (3,691)
    Hospitalized = Yes [ Mode: FATAL ] ⇨ FATAL (928; 0.945)
    Hospitalized = No [ Mode: FATAL ] (2,763)
      Source of Infection in [ "Close contact" "Healthcare" "N/A - Outbreak associated" ] [ Mode: FATAL ] ⇨ FATAL (2,7
      Source of Infection in [ "Community" "Institutional" "Unknown/Missing" ] [ Mode: NONFATAL ] ⇨ NONFATAL (18;
      Source of Infection in [ "Pending" "Travel" ] [ Mode: FATAL ] ⇨ FATAL (0)
  Age Group in [ "19 and younger" "20-29" "30-39" "40-49" "50-59" "60-69" ] [ Mode: NONFATAL ] (9,017)
    Hospitalized = Yes [ Mode: FATAL ] (1,577)
      Age Group in [ "19 and younger" "20-29" "30-39" "40-49" "50-59" ] [ Mode: NONFATAL ] (718)
      Age Group in [ "60-69" ] [ Mode: FATAL ] ⇨ FATAL (859; 0.785)
    Hospitalized = No [ Mode: NONFATAL ] ⇨ NONFATAL (7,440; 0.96)
Age Group in [ "70-79" "80-89" ] [ Mode: FATAL ] (6,614)
  Hospitalized = Yes [ Mode: FATAL ] ⇨ FATAL (2,797; 0.887)
  Hospitalized = No [ Mode: FATAL ] (3,817)
    Source of Infection in [ "Close contact" "Community" "Healthcare" "Pending" ] [ Mode: NONFATAL ] ⇨ NONFATAL (2
    Source of Infection in [ "Institutional" "N/A - Outbreak associated" ] [ Mode: FATAL ] ⇨ FATAL (3,488; 0.77)
    Source of Infection in [ "Travel" ] [ Mode: NONFATAL ] (38)
      Client Gender in [ "FEMALE" ] [ Mode: NONFATAL ] ⇨ NONFATAL (8; 1.0)
      Client Gender in [ "MALE" ] [ Mode: FATAL ] (30)
      Client Gender in [ "OTHER" "TRANSGENDER" "UNKNOWN" ] [ Mode: NONFATAL ] ⇨ NONFATAL (0)
    Source of Infection in [ "Unknown/Missing" ] [ Mode: NONFATAL ] (25)
      Age Group in [ "70-79" ] [ Mode: NONFATAL ] ⇨ NONFATAL (9; 1.0)
      Age Group in [ "80-89" ] [ Mode: FATAL ] (16)
```

Tree map

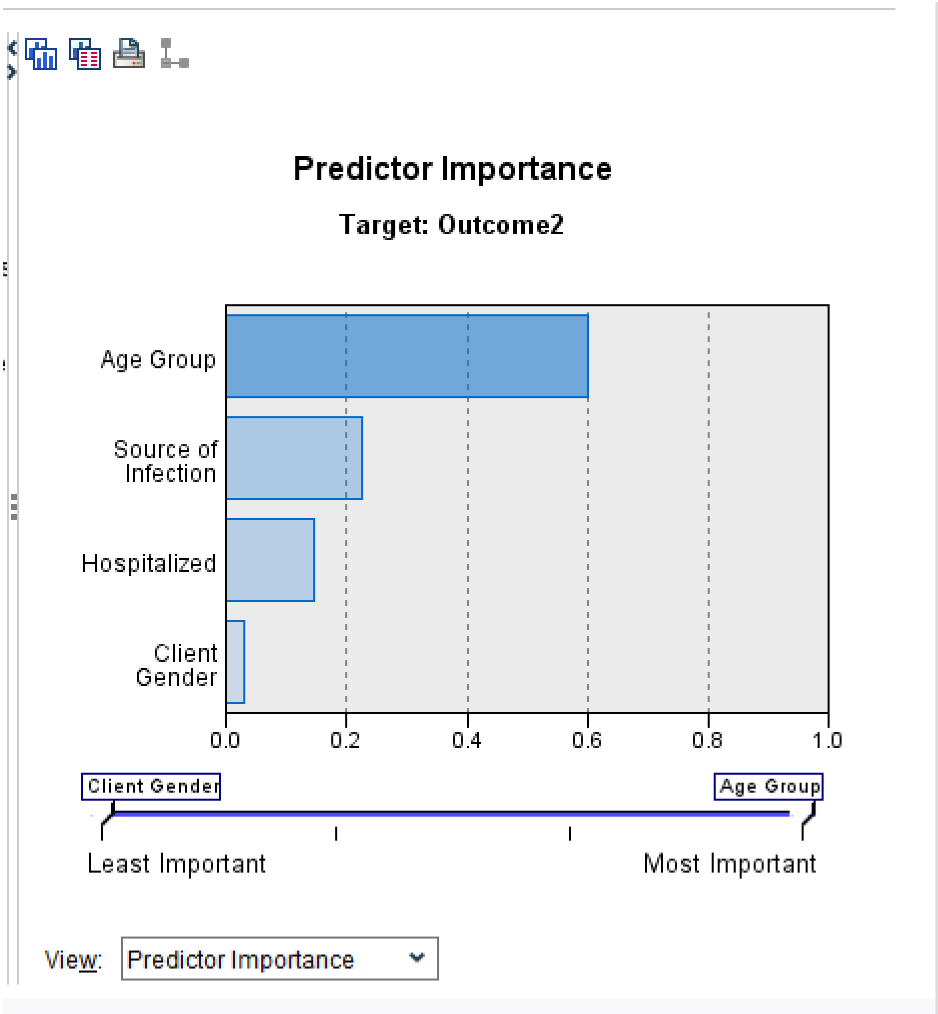


Rule set

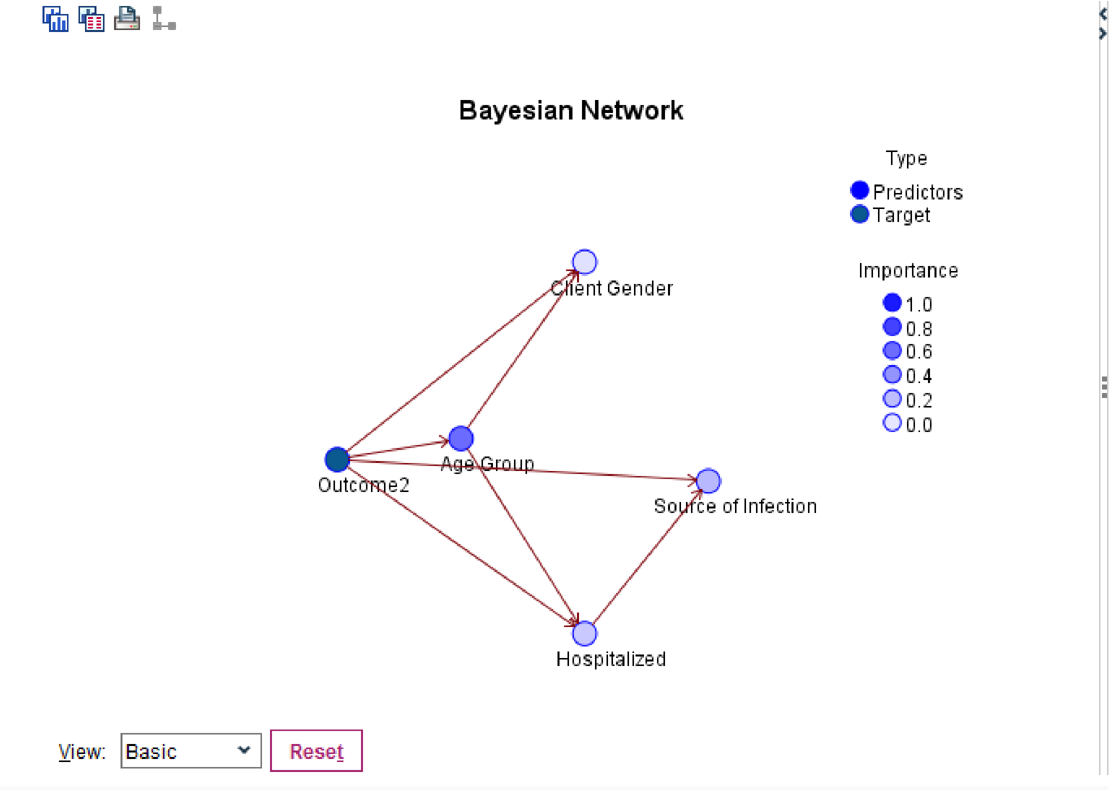


8.2.2 Bayesian Network model

Bar chart of predictor importance


Predictor Importance
Target: Outcome2
Age Group
Source of Infection
Hospitalized
Client Gender
0.0
0.2
0.4
0.6
0.8
1.0
Client Gender
Age Group
Least Important
Most Important
View:
Predictor Importance

network graph



8.3 Interpret the results, models and patterns showing a clear understanding of the results.

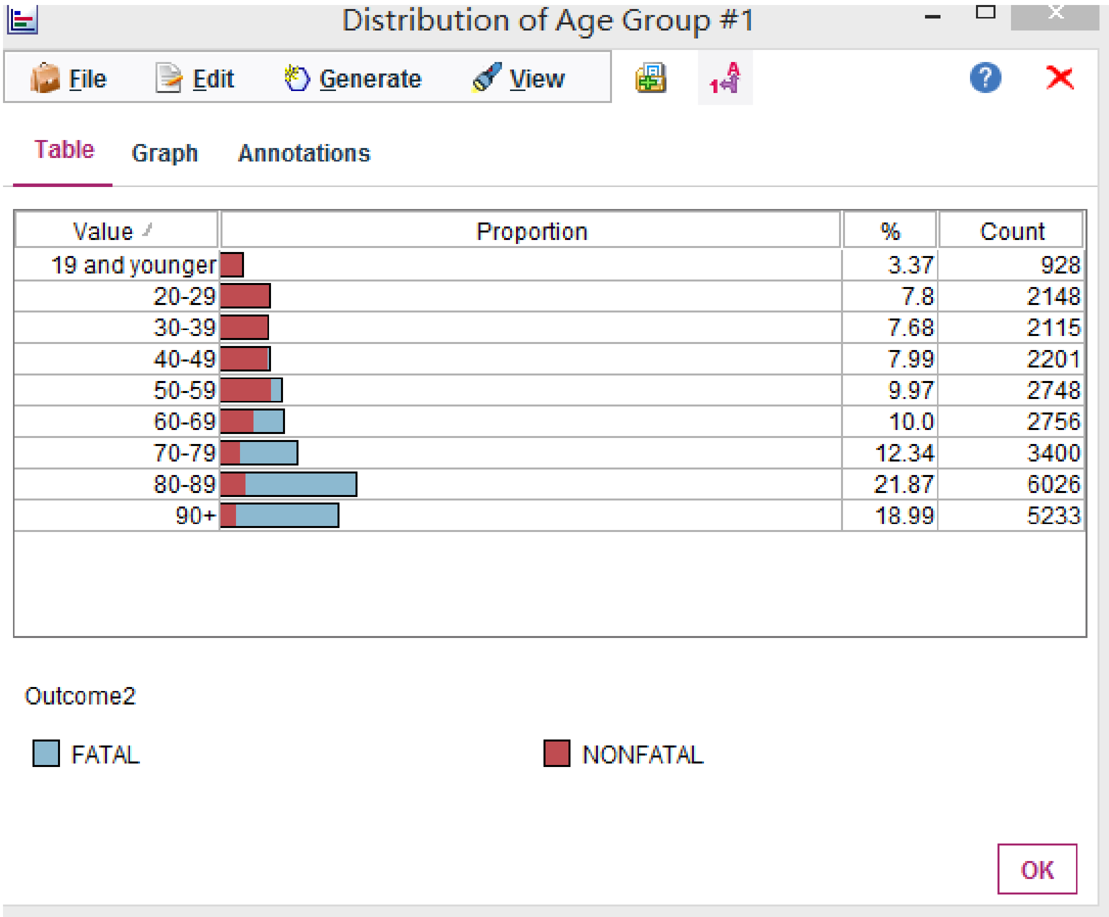
Through the above steps, I have a clear understanding of the project's results, models, and patterns.

8.3.1 predictor importance

From highest to lowest ranking, Age Group, Hospitalized, Source of Infection and Client Gender are the priorities of The C5.0 model. This is very different from the Bayesian Network Model. The Bayesian Network Model ranked Age Group, Source of Infection, Hospitalized and Client Gender from highest to lowest. Combining these two models, it can be seen that age is the most important predictor, while gender in either model, the importance of the predictor tends to 0, which has little predictive significance.

Conditional Probabilities of Age Group

Parents	Probability								
Outcome2	19 and younger	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90+
NONFATAL	0.07	0.16	0.16	0.15	0.16	0.11	0.06	0.08	0.05
FATAL	0.00	0.00	0.00	0.01	0.03	0.09	0.18	0.36	0.33



Value	Proportion	%	Count
19 and younger		3.37	928
20-29		7.8	2148
30-39		7.68	2115
40-49		7.99	2201
50-59		9.97	2748
60-69		10.0	2756
70-79		12.34	3400
80-89		21.87	6026
90+		18.99	5233

The Conditional Probabilities of The Age Group, according to The results of The elderly, especially in The phase of 80-89 and 90 + COVID - 19 has The highest mortality, FATAL condition possibility is more than 30%.On the other hand, young people, especially those in age groups 20-29 and 30-39, are most likely to be fatal.

8.3.2 Achieve business goals

According to the bar chart of predictor importance, the characteristics that affect suicide rate are Age Group, Source of Infection, Hospitalized.

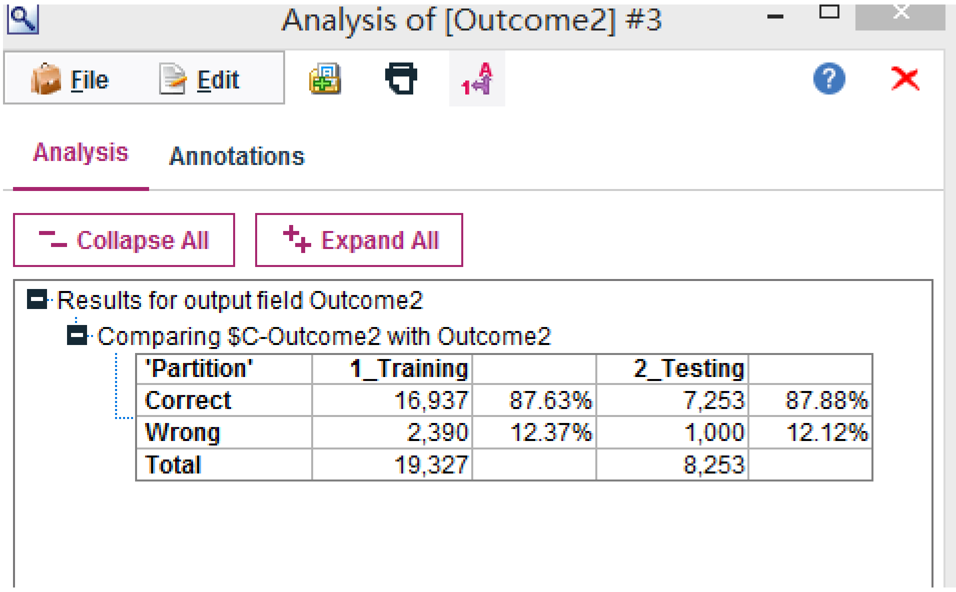
The rule set can be used to guide the state, institutions or families to focus on protecting groups of older persons.These groups with lower survival rates are regularly checked and equipped with more effective protective measures.

8.4 - Assess and evaluate the results, models and patterns using the appropriate methods/processes.

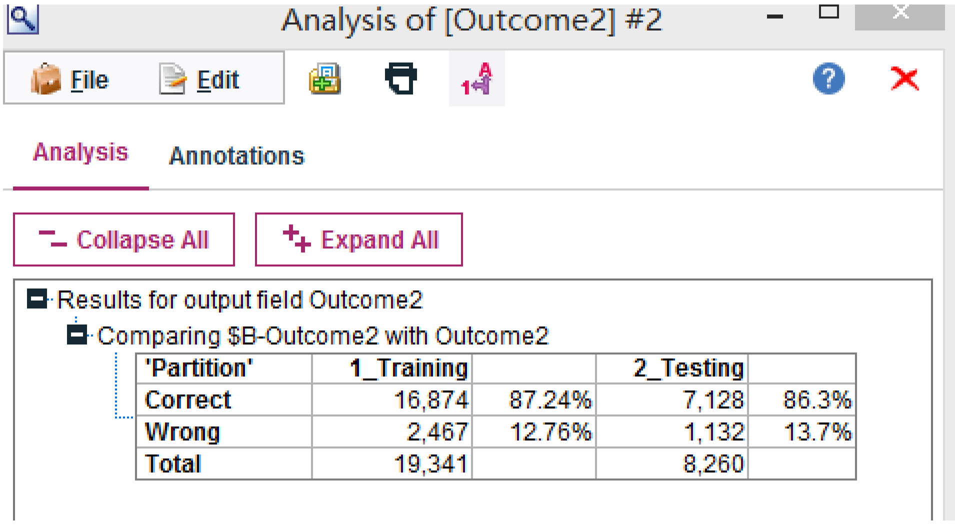
8.4.1 assess the results, models and patterns

(1) Test Accuracy

Use the analysis node to evaluate the model, and the results are shown below. Both of the model have a high accuracy rate. The accuracy rate of C5.0 testing set is 87.88%. The accuracy rate of Bayesian Network model testing set is 86.3%.

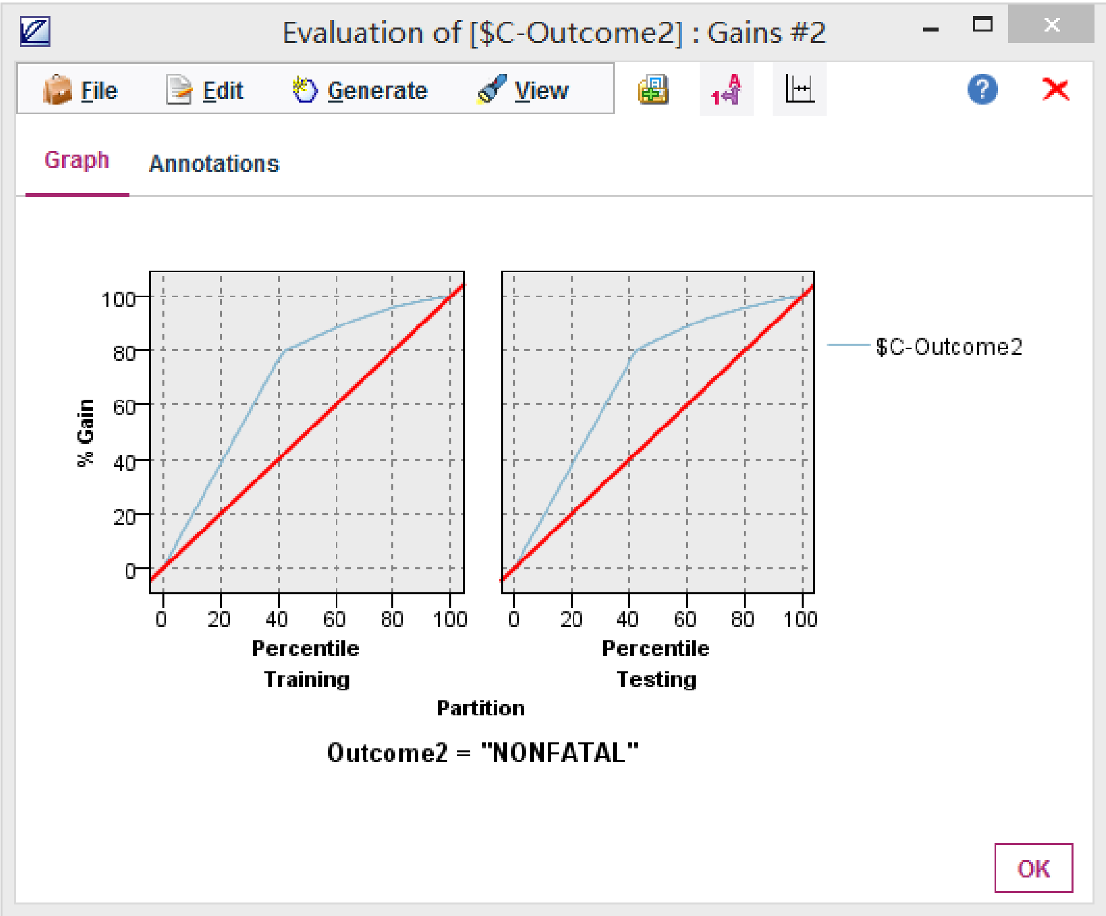


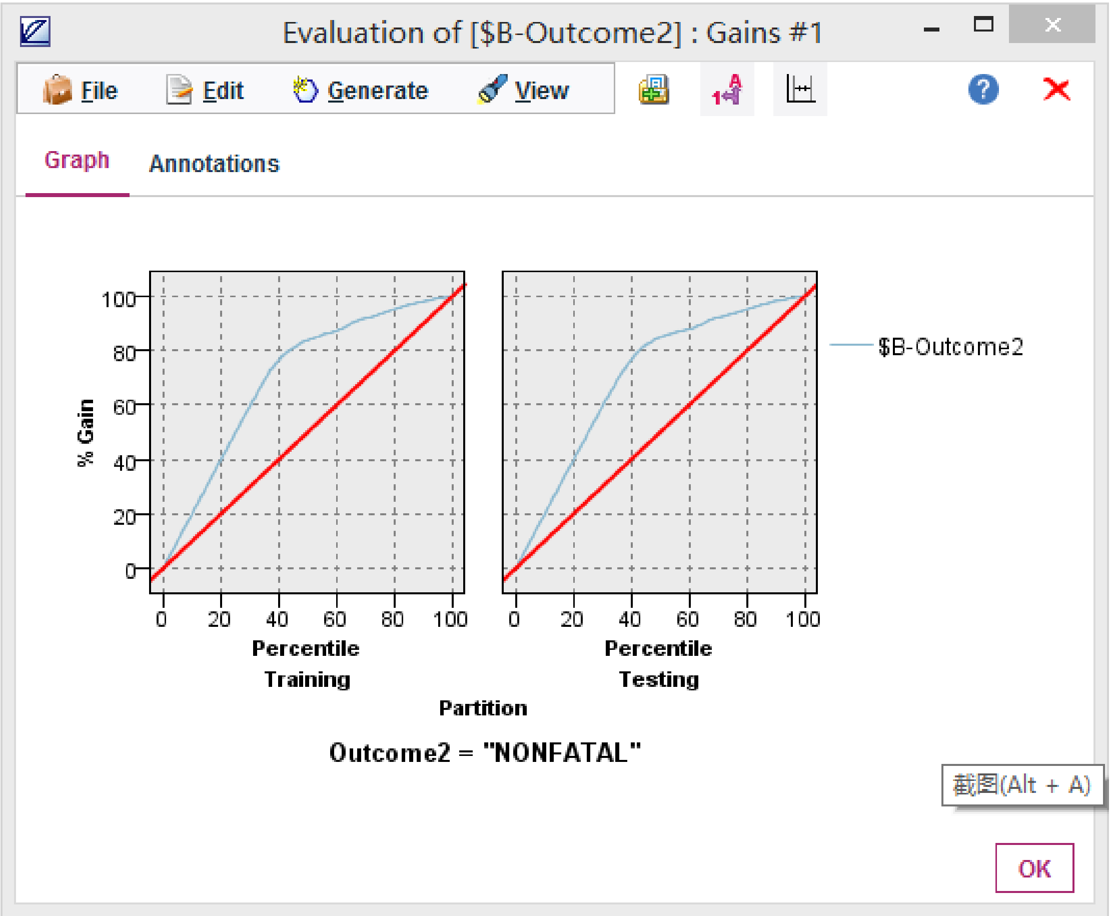
'Partition'	1_Training		2_Testing	
Correct	16,937	87.63%	7,253	87.88%
Wrong	2,390	12.37%	1,000	12.12%
Total	19,327		8,253	



'Partition'	1_Training		2_Testing	
Correct	16,874	87.24%	7,128	86.3%
Wrong	2,467	12.76%	1,132	13.7%
Total	19,341		8,260	

(2) Evaluation Graph


Evaluation of [$C-Outcome2] : Gains #2
File
Edit
Generate
View
Graph
Annotations
100
80
60
40
20
0
% Gain
0
20
40
60
80
100
Percentile
Training
Percentile
Testing
Partition
$C-Outcome2
Outcome2 = "NONFATAL"
OK



8.4.2 evaluation the results, models and patterns

(1) Business goals

People at high risk of COVID-19 can be protected by pre-locating and taking preventive measures.

(2) Data mining goals

Decision rules can be used to predict or classify a group of people with high coVID-19 survival rates, achieving the required model accuracy, but the data quality is poor and may affect the use of the model. The result of the model is logical. The results are easy to understand and easy to deploy

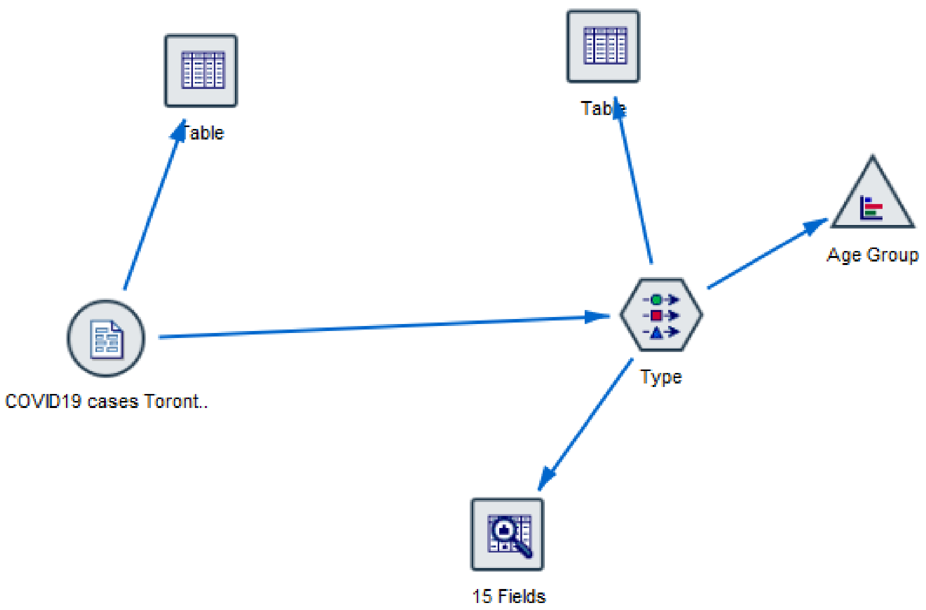
8.5 Iterate prior steps (1 – 7) as required

8.5.1 Business understanding

Covid-19 is one of the most difficult diseases in the world, and there are still tens of thousands of living cases every day in countries that are at risk of death every day. However, the vaccine development process still needs a lot of time. How to reduce the increase of cases in the meantime, who should we care about most. To find out who is more deserving of additional protection and special care, we used a dataset of COVID-19 cases in the Toronto area as a data source to analyze which factors contribute to the survival of infected persons and which factors are associated with case fatality.

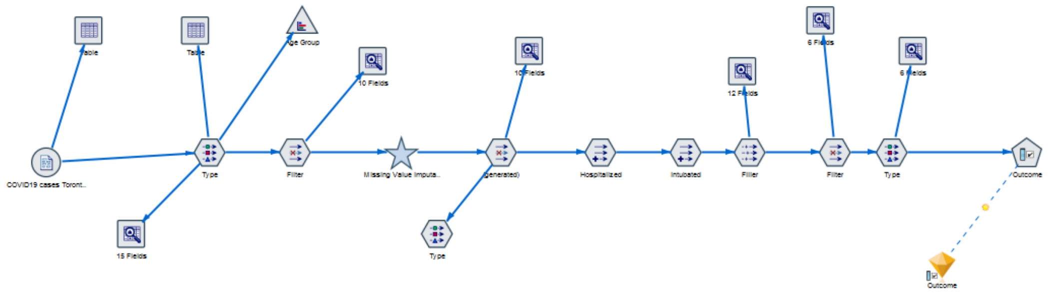
8.5.2 data understanding

Data collection, preliminary cognition and processing of data, and certain cognition of data type, data size and processing method. There are assumptions and analysis.



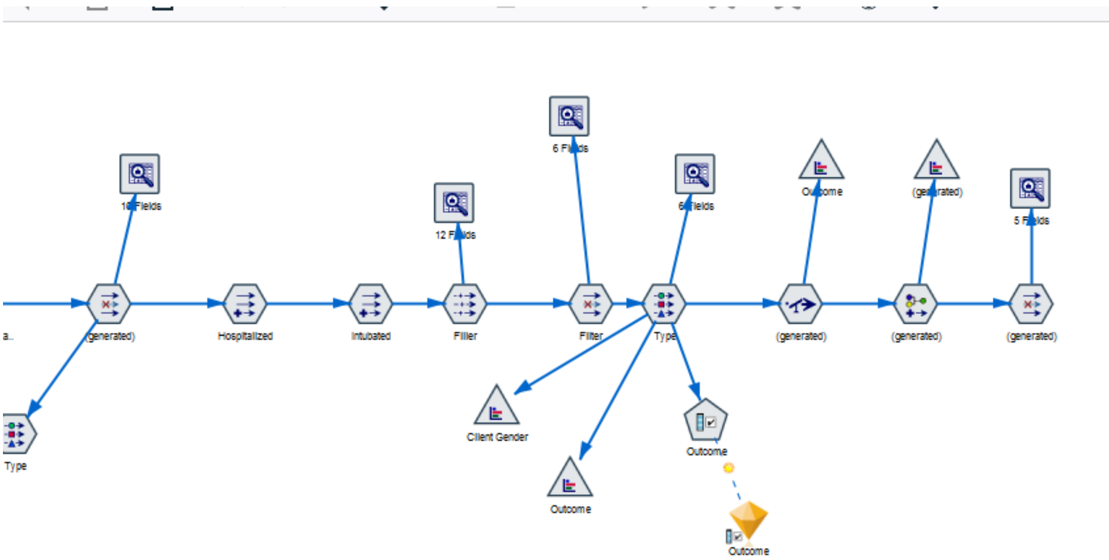
8.5.3 Data preparation

Select data, clean data, merge data, deal with missing value, and format data. Prepare the data required for subsequent modeling.



8.5.4 Data transformation

To further simplify the data set and remove the unimportant and disturbing data, balance the data.



8.5.5 Data-mining method selection

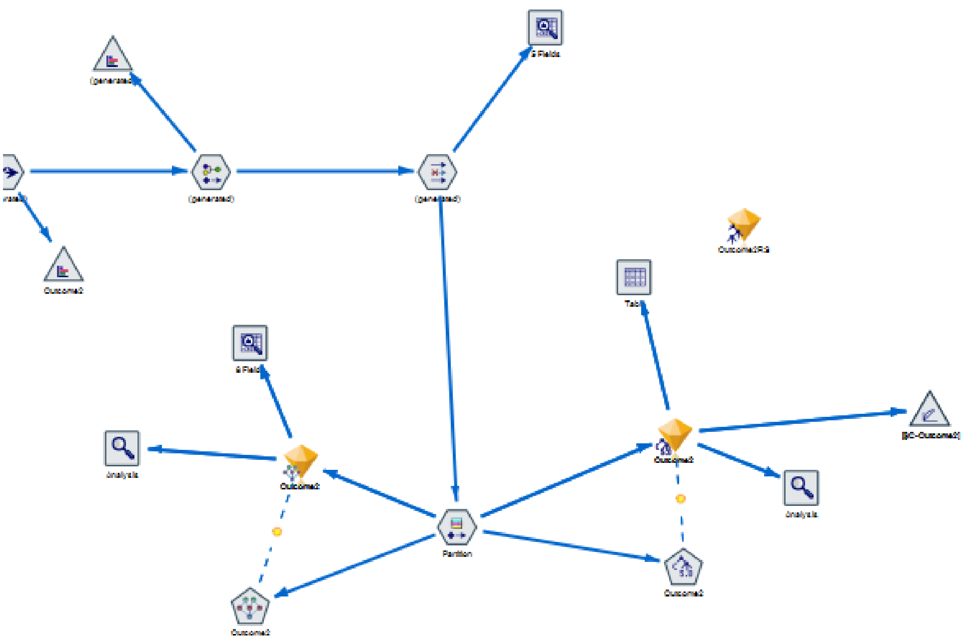
The output variable and the input variable are both String values. In consideration of the target values and methods to predict, I choose supervisory learning and classification.

8.5.6 Data-mining algorithms selection

SPSS Modeler provides many reliable algorithms. After investigation, it was found that these algorithms have their own advantages and disadvantages, so I finally chose To use C 5.0.And the Bayesian network model.

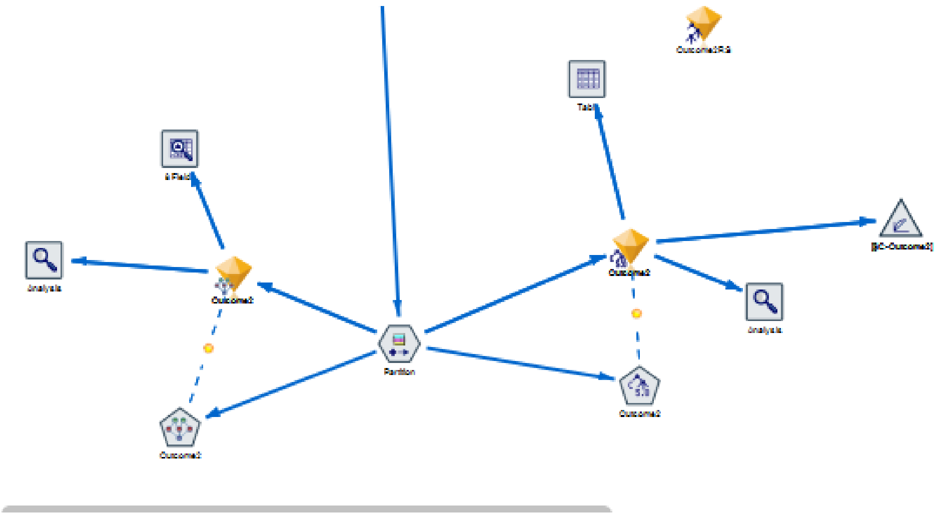
8.5.7 Data-mining

Through the construction of the model, let me have a feeling of suddenly enlightened. The model clearly shows the importance of the attributes and their relationship to the target value, shaping the prediction.



8.5.8 Interpretation

This step can also involve visualizing the extracted patterns and models, and we evaluate and evaluate the models, results, and their reliability.



Reference

Jones, M.(2017). Unsupervised learning for data classification. Retrieved from https://developer.ibm.com/articles/ccunsupervised-learning-data-classification/

IBM Knowledge Center (n.d.) Overview of modeling nodes**.** Retrieved from https://www.ibm.com/support/knowledgecenter/en/SS3RA7_18.1.0/modeler_mainhelp_client_ddita/clementine/modeling_nodes.html

Data for Data Mining. (n.d.). *Principles of Data Mining*, 11–21. doi:10.1007/978-1-84628-766-4_1.

Usama, F., Gregory, P.-S., & Padhraic, S. (1996) Knowledge discovery and data mining: toward a unifying framework. In *Proceeding of The Second Int. Conference on Knowledge Discovery and Data Mining*, pages 82-88, 1996.

"I acknowledge that the submitted work is my own original work in accordance with the University of Auckland guidelines and policies on academic integrity and copyright.
(See: https://www.auckland.ac.nz/en/students/forms-policies-and-guidelines/student-policies-and-guidelines/academic-integrity-copyright.htmlLinks to an external site.).

I also acknowledge that I have appropriate permission to use the data that I have utilised in this project. (For example, if the data belongs to an organisation and the data has not been published in the public domain then the data must be approved by the rights holder.) This includes permission to upload the data file to Canvas. The University of Auckland bears no responsibility for the student's misuse of data."