INFOSYS 722 Data Mining and Big Data

Iteration 2

Semester 2, 2022

Qiong Zhou / 365217677 / qzho906

Table of Contents

1. Business / Situation Understanding

1.1 Business understanding Overview

Ischemic heart disease, stroke, and chronic obstructive pulmonary disease are the top three causes of death from disease, accounting for 16%, 11%, and 6 % of deaths worldwide respectively (WHO, 2020).

Stroke is a condition in which the blood supply to the brain is interrupted, resulting in a lack of oxygen, brain damage, and loss of function. Stroke can lead to permanent damage, including partial paralysis and impairments in speech, understanding, and memory, all of which affect the type and severity of disability depending on the part of the brain affected and the length of time the blood supply is stopped (World Stroke Organization, 2022). The prevalence of stroke has reached epidemic proportions beyond what is thought possible. Globally, one in four adults over the age of 25 will have a stroke in their lifetime (World Stroke Organization, 2022). This year 12.2 million people worldwide will have their first stroke and 6.5 million will die as a result. Worldwide, more than 110 million people have experienced a stroke (World Stroke Organization, 2022). The incidence of stroke increases significantly with age, but over 60% of strokes occur in people under the age of 70 and 16% in people under the age of 50 (World Stroke Organization, 2022).

Although stroke is an acute cerebrovascular disease with high morbidity, mortality, and disability rate. Many factors contribute to stroke, including age, race, gender, geography, and environment, which are not controllable. However, according to a review article published in the Journal of the American College of Cardiology, 90% of strokes are preventable and the key is to manage and treat controllable risk factors (Kleindorfer DO, Towfighi A, Chaturvedi S, et al., 2021). The controllable risk factors are blood pressure, blood lipids, blood sugar, and lifestyle (smoking, alcohol, etc.).

1.2 Determining Business Objectives

Although the disease itself can develop in a very short period of time and without any precursors, it is still possible to classify and predict patients based on their early personal history and to offer solutions and advice to patients while reducing the chances of a stroke.

One of the main clinical risk factors for stroke is atherosclerosis-induced hypertension, also, there are many other risk factors including smoking, physical inactivity, unhealthy diet, harmful use of alcohol, atrial fibrillation, elevated blood lipid levels, obesity, genetic predisposition, stress, and depression (World Stroke Organization, 2022).

In this case, the study was commissioned with the following data business objectives:

- Be able to predict, prevent and reduce the occurrence of stroke disease based on the patient's current status.

The expected outcome of the study:

- Be able to provide current patients with detailed information on the causes of their current stroke in terms of current health indicators, age, and life status.
- Reduces the likelihood of a potential patient having a stroke at an early stage.

1.3 Accessing the Situation

In order to achieve the goal of being able to analyze in detail the causes of stroke in stroke patients and to detect early manifestations of signs of stroke. This study should review a large number of existing patient cases and engage them as a dataset to find models, as well as relationships between signs, data mining specialists should be applied.

Personnel. It is clear and confident that it is important to consult doctors and specialists in the field of stroke research, as well as those working in the field of healthcare and rehabilitation for stroke patients, and to obtain clear and detailed information about the disease itself, including the primary, secondary and direct causes of stroke. About the data and information collected from these stroke specialists, a database specialist should be consulted. Since these specialists hope the results of the study will become part of a continuing studying and research process, data warehousing and data cleaning for analysis are required. Also, the information about the patients involves their privacy, the data management must also be considered.

Data. The data required for the study analysis will come primarily from the records of doctors and healthcare professionals, where detailed desensitized data on patients is recorded. For initial studies, the data can be easily found on the internet, furthermore, the study data can be expanded as the study is conducted.

Risks. As the project is primarily concerned with healthcare, the accuracy of the model depends to a large extent on the quality and quantity of the dataset used for the study. The model production process should also be concerned with the risk of privacy breaches to users. Once generated, the models should be further validated by a team of experts who specialize in the treatment and research of stroke patients.

1.4 Data Mining Objectives

Experts in data mining contributed by translating the project's business objectives into data mining objectives. All data mining objectives studied in the initial phase are listed below.

In this case, the study was commissioned with the following data mining objectives:

- Identify a clear relationship between the patient's current health Index (current illness and BMI) and the stroke.
- Find out the relationship between the patient's age and the chances of stroke.
- Find the relationship between the patient's current life status (smoking status, marital status, work type, living environment, etc.) and the stroke.
- Reduce the chances of stroke by providing predictions based on the patient's current illness and age.
- Use this data to train and fit the model to make it suitable for use in data prediction.

The validation phase uses a selection of healthcare records as a validation process to see the prediction results of a given model, and if the model successfully passes a pre-determined threshold, validation by healthcare professionals is continued.
The initial use of the model phase allows the healthcare professional to enter a new patient's medical history, view the patient's stroke risk rating, and have the doctor diagnose to see if the system passes the threshold.
The enhancement and optimization phase should continue to collect the required data and retrain and improve the model to increase accuracy.

The expected outcome of the study:

- Provide detailed information on the causes of the stroke to current patients.
- Provide useful information and advice to people who are at risk of stroke.
- Reduce the likelihood of stroke in potential patients at an early stage.

If the above objectives are achieved, the system should provide the ability to diagnose the patient's risk level.

1.5 Project Plan

The entire project is planned to take ten weeks of work and is planned as follows, also the detailed project schedule is shown in Gantt Chart in Figure 1.

Phase	Time	Resources	Risks
Business Understanding - Business background research - Business case study	1 week	Analysts	- Project change - Find the unsuitable case to study
Data Understanding - Dataset selection - Data exploration	2 weeks	Analysts	- Data problems - Attribute selections in the dataset
Data preparation - Data selection - Data cleaning - Data construction - Data integration - Data formatting	2 weeks	Data mining experts	- Technical problems
Modeling - Train the model - Evaluation of the performance	2 weeks	Database analyst time	- Model fitting problems - Technical problems

- improve the performance			
Validation - Validate the model by test sets	1 week	Analysts and tests	- Incompatible model - Incompatible test sets - Technical problems
Evaluation	1 week	Analysts	- Technical problems
Deployment	1 week	- Data mining consultant - Database analyst time	- Inability to deploy the whole project

Table 1. Project Plan

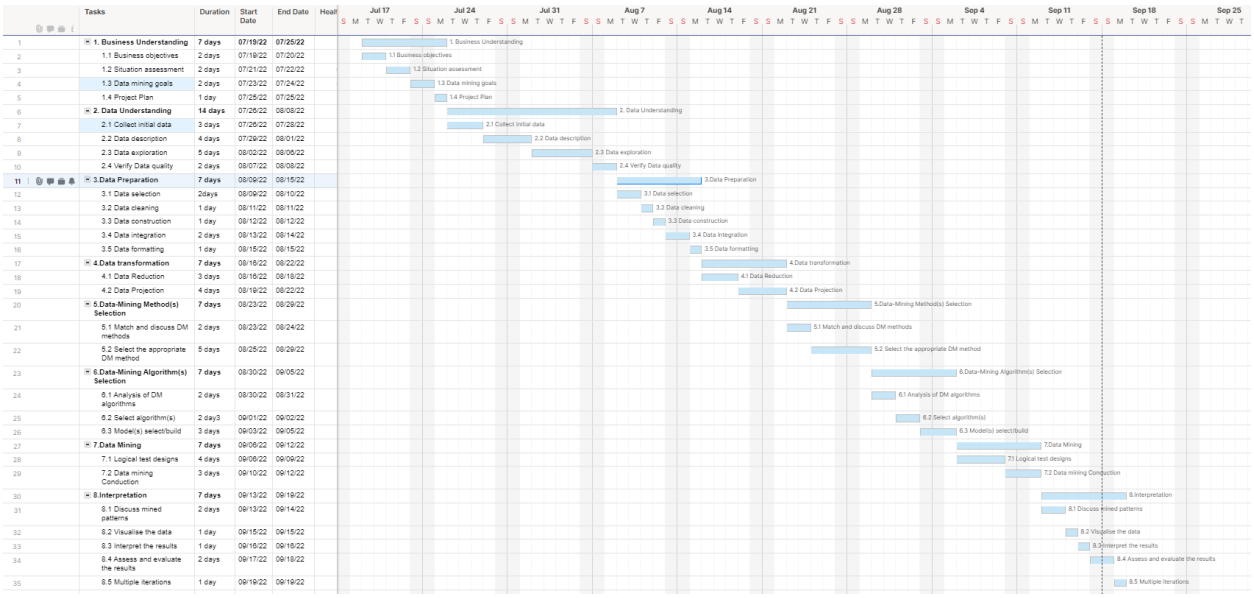


Figure 1. Gantt Chart of project schedule

2. Data Understanding

2.1 Data Understanding Overview

The healthcare information document is collected from the Internet, Kaggle. The data used will be accessed from a website link https://www.kaggle.com/datasets/lirilkumaramal/heart-stroke. The API command is kaggle datasets download -d lirilkumaramal/heart-stroke. It was last updated on 14 August 2021.

2.2 Collecting Initial Data

The detailed information covered in the dataset is as below:

- Personal information. The raw data contain the basic information about individuals but does not relate to the disease itself, which includes age and gender.
- Health Index. The raw data contain a survey of diseases and basic physical indicators. Diseases investigated include the presence of hypertension and heart disease. The underlying physical indicators include mean glucose levels and BMI.
- Life status. The raw data contains living status and habits, whether married or not, type of work, type of residence, and smoking status.

In terms of data quality assumptions, firstly, the dataset is a desensitization dataset, as obviously the dataset did not include sensitive data, for example, name, address, postal code, and so on. Secondly, the dataset is rich enough for training a model for stroke prediction. Thirdly, the dataset is correctly collected from stroke patients or potential patients.

The main problem with this dataset is that it provides limited information about the origin of the dataset itself and only states that this dataset is widely used. This makes the source of this dataset, and in which region it was collected and generated, ambiguous in the study. Once the dataset has been fabricated, all of the findings and predictions in this article will be meaningless. After searching through the discussion section of this dataset, it was found that the data was actually collected by the clinic and the source of the data was deliberately removed due to privacy concerns. After confirming the authenticity of the data sources, the reliability of this dataset was verified and approved for data mining.

2.3 Describing Data

Data Quantity

The format of the dataset has been packed as a .csv file, it can be easily accessed and converted to the desired format and used, for example, as a .xlsx file. The dataset contains 12 attributes and 43,400 records, in which, each record contains the information of individuals as configured in the SPSS Modeler below.

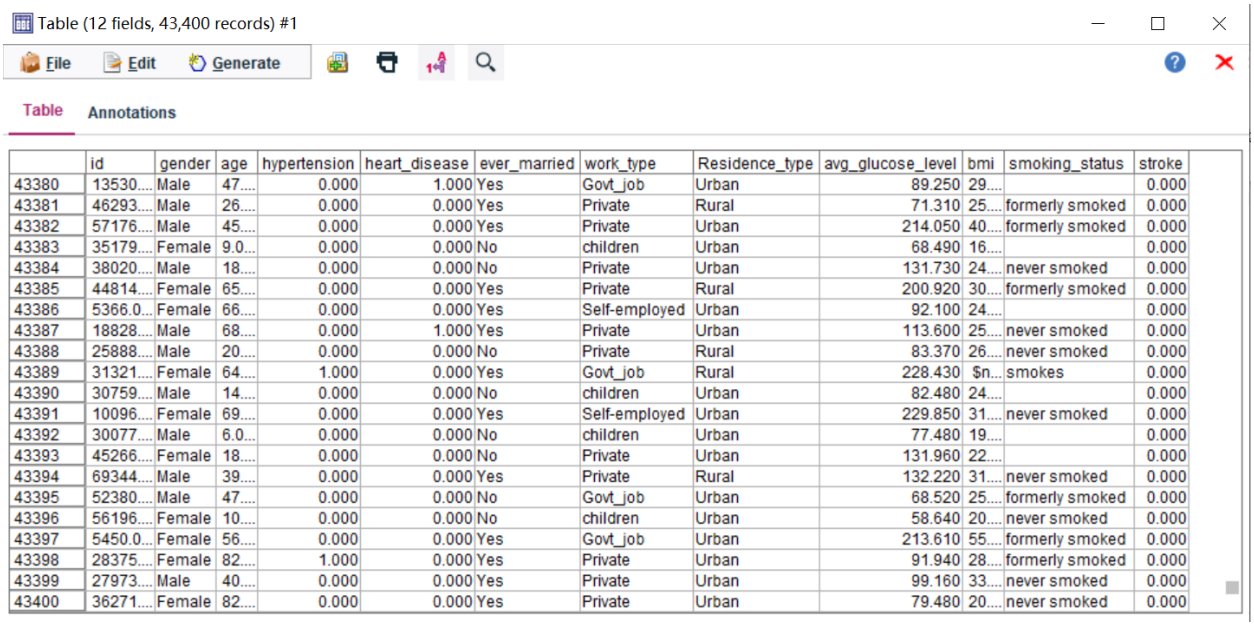


Table (12 fields, 43,400 records) #1

File Edit Generate

Table Annotations

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
43380	13530....	Male	47....	0.000	1.000	Yes	Govt_job	Urban	89.250	29....		0.000
43381	46293....	Male	26....	0.000	0.000	Yes	Private	Rural	71.310	25....	formerly smoked	0.000
43382	57176....	Male	45....	0.000	0.000	Yes	Private	Urban	214.050	40....	formerly smoked	0.000
43383	35179....	Female	9.0...	0.000	0.000	No	children	Urban	68.490	16....		0.000
43384	38020....	Male	18....	0.000	0.000	No	Private	Urban	131.730	24....	never smoked	0.000
43385	44814....	Female	65....	0.000	0.000	Yes	Private	Rural	200.920	30....	formerly smoked	0.000
43386	5366.0...	Female	66....	0.000	0.000	Yes	Self-employed	Urban	92.100	24....		0.000
43387	18828....	Male	68....	0.000	1.000	Yes	Private	Urban	113.600	25....	never smoked	0.000
43388	25888....	Male	20....	0.000	0.000	No	Private	Rural	83.370	26....	never smoked	0.000
43389	31321....	Female	64....	1.000	0.000	Yes	Govt_job	Rural	228.430	$n...	smokes	0.000
43390	30759....	Male	14....	0.000	0.000	No	children	Urban	82.480	24....		0.000
43391	10096....	Female	69....	0.000	0.000	Yes	Self-employed	Urban	229.850	31....	never smoked	0.000
43392	30077....	Male	6.0...	0.000	0.000	No	children	Urban	77.480	19....		0.000
43393	45266....	Female	18....	0.000	0.000	No	Private	Urban	131.960	22....		0.000
43394	69344....	Male	39....	0.000	0.000	Yes	Private	Rural	132.220	31....	never smoked	0.000
43395	52380....	Male	47....	0.000	0.000	No	Govt_job	Urban	68.520	25....	formerly smoked	0.000
43396	56196....	Female	10....	0.000	0.000	No	children	Urban	58.640	20....	never smoked	0.000
43397	5450.0...	Female	56....	0.000	0.000	Yes	Govt_job	Urban	213.610	55....	formerly smoked	0.000
43398	28375....	Female	82....	1.000	0.000	Yes	Private	Urban	91.940	28....	formerly smoked	0.000
43399	27973....	Male	40....	0.000	0.000	Yes	Private	Urban	99.160	33....	never smoked	0.000
43400	36271....	Female	82....	0.000	0.000	Yes	Private	Urban	79.480	20....	never smoked	0.000

Figure 2. Initial Data Overview from SPSS Modeler

Data Features

The objective of this project is to find the relationship between given information and the chance of stroke occurring. The information collected in the dataset, such as age, BMI, glucose, hypertension, history of heart disease, etc., were categorized early on as risk factors for stroke. The dataset is combining multiple formats of coding, the table 2 is showing the data type as well as the note of each attribute. Among these 12 attributes, 4 are processed as numeric data, 4 are processed as flags, and the rest 4 are processed as categories. These attributes provide sufficient information related to stroke. Therefore, based on medical research, it can be said that the information contained in the dataset has a strong relationship with stroke (Lim, 2021).

The dataset provides a realistic picture of the individual, as each person is flagged stroke status and other related attributes as shown below.

Attribute	Value Type	Usage	Example
ID	Numeric	Unique Identifier	30468
Gender	Categorical (string)	Male, Female, Other	Male
Age	Numeric	Continues number	58
Hypertension	Boolean (Flag)	0 = No hypertension before; 1 = Yes	1 (Yes)

Heart_disease	Boolean (Flag)	0 = No heart disease before; 1 = Yes	0 (No)
Ever_married	Boolean (Flag)	0 = Never married; 1 = Married	Yes
Work_type	Categorical (string)	Type of work the individual has	Private
Residence_type	Categorical (string)	Rural/Urban	Urban
Avg_glucose_level	Numeric (Decimal)	The average glucose level of an individual	87.96
BMI	Numeric (Decimal)	The body mass index of an individual	39.2
Smoking_status	Categorical (string)	Formerly smoked/ smokes/never smoke	never smoked
Stroke	Boolean (Flag)	0 = No stoke happened before; 1 = Yes	0

Table 2. Attribute and Value types of Analysis

Inside this dataset, 41% of samples were recorded from male patients, and 59% of samples were from females. The average age of all patients was 42.2±22.5 with an average BMI of 28.6-±7.77, 64% of patients were ever married. 57% of patients worked for private companies, and 16% were self-employed. 50% of them lived in a rural place while others were urban. 90.64% of patients had a record of hypertension, and 95.25% of all patients were diagnosed with heart diseases and stroked happened on 98.20% of patients. The average glucose level of all patients was 104±43.1. In the records of smoking status for patients, 37% of them never smoked, the rest of them either did provide any records or were grouped into others.

At the current stage, the preference attributes associated with stroke are not known. Further exploration of the dataset is required.

2.4 Exploring Data

After observing the relationship between different attributes with target attribute stroke by plots in the SPSS application as shown in Figure 3. With these attributes, we do not consider with the relationship between id and stroke, as it is obvious that ID is not an influential attribute toward stroke.

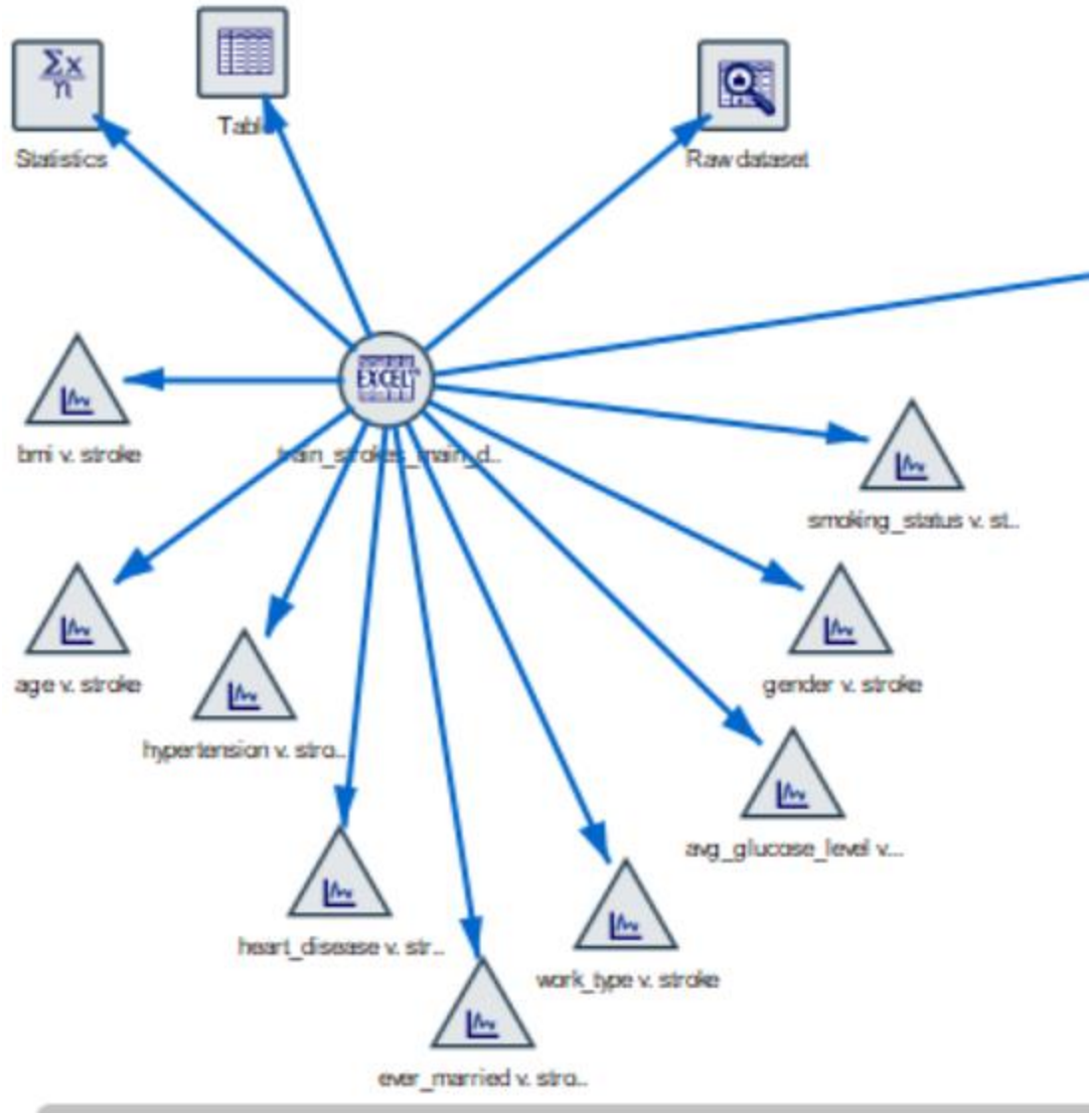


Figure 3. Plot relationship between different attributes with stroke

Then we take the first glance of the relationship between BMI (Figure 4), age (Figure 5), hypertension (Figure 6), heart disease (Figure 7), marriage status (Figure 8), work type (Figure 9), average glucose (Figure 10), gender (Figure 11), smoking status (Figure 12) and residence type (Figure 13) with target attribute stroke respectively.

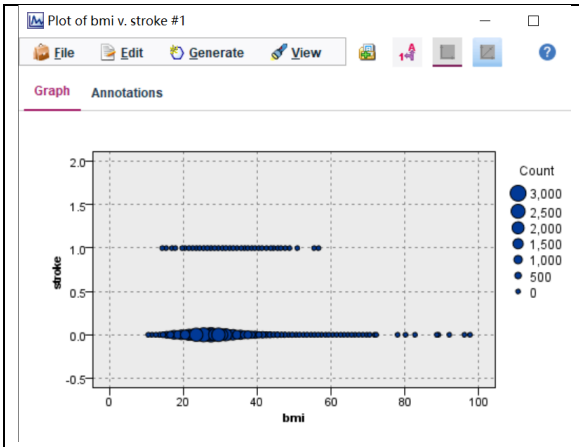


Figure 4. The relationship plot between BMI and stroke

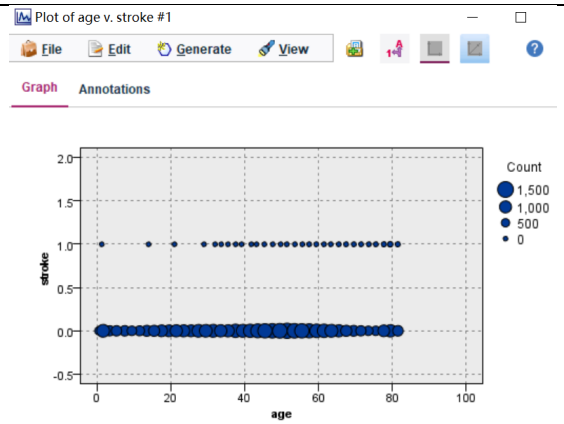


Figure 5. The relationship plot between age and stroke

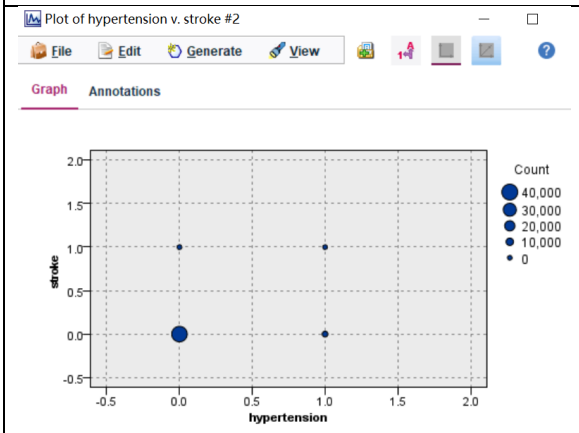


Figure 6. The relationship plot between hypertension and stroke

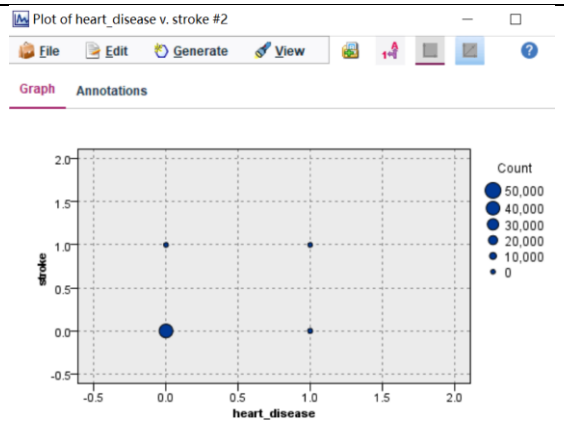


Figure 7. The relationship plot between heart disease and stroke

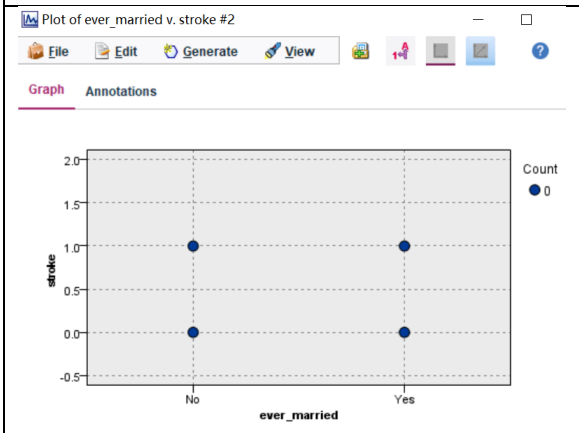


Figure 8. The relationship plot between marriage status and stroke

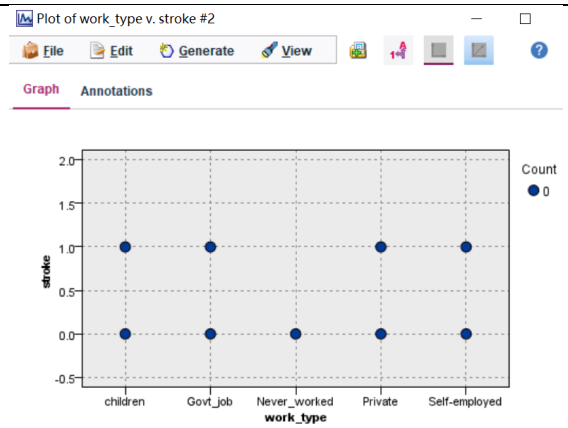


Figure 9. The relationship plot between work type and stroke

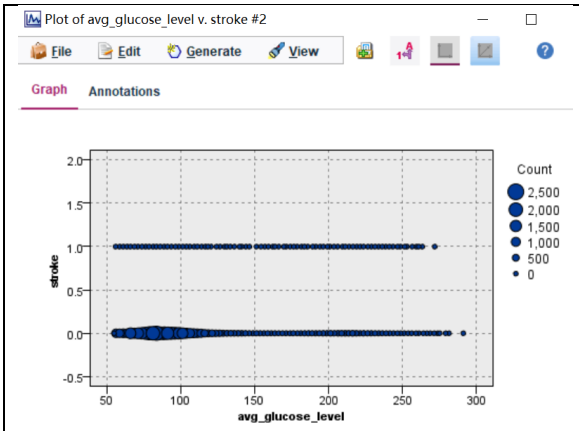


Figure 10. The relationship plot between average glucose and stroke

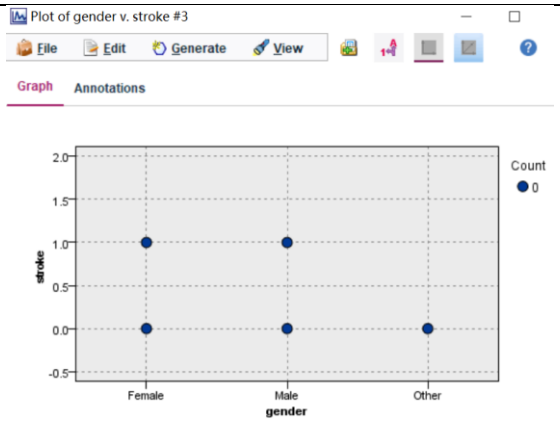


Figure 11. The relationship plot between gender and stroke

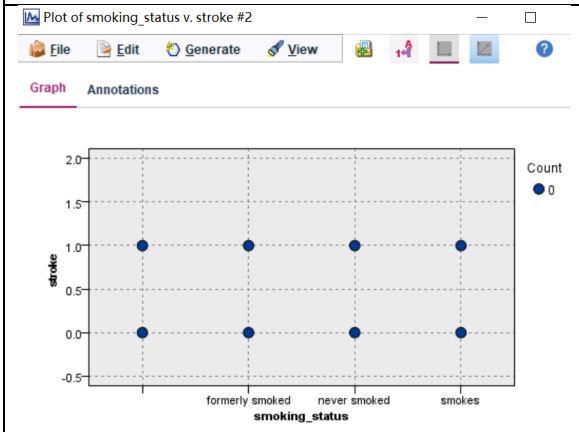


Figure 12. The relationship plot between smoking status and stroke

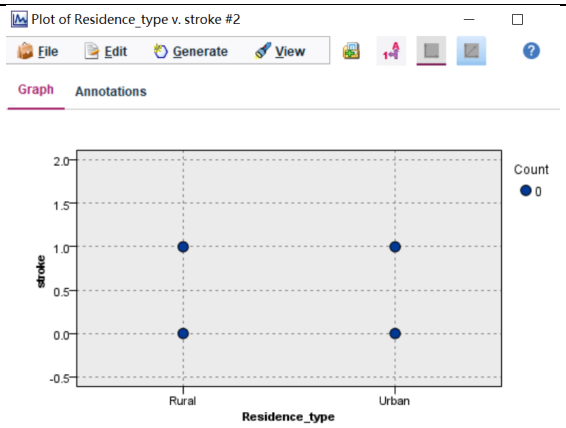


Figure 13. The relationship plot between residence type and stroke

From Figure 4, the number of stroke cases were analyzed in patients with different BMI. From the statistical result, patients with from 14.3 to 56.6 showed more cases of a stroke happening, while patients with a lower BMI (<15) or higher BMI (>55) showed less cases of stroke.

From Figure 5, the number of stroke cases among all different age patients were presented. A larger count circle indicates a higher records of stroke cases. From the figure, stroke can happen at any age but most of stroke cases were happened when age (>30) years old.

From Figure 6, patients without underlying conditions of hypertension are less likely to suffer a stroke.

From Figure 7, patients without underlying heart disease are less likely to have a stroke.

From Figure 8, we cannot observe whether marriage was a factor in the stroke.

From Figure 9, we can see that never worked patients exclude children has less chance to get stroke.

From Figure 10, the number of stroke cases among all different glucose level (56 - 260) patients were presented, but glucose level from 55 to 116 has most patients who did not get stroke before. And the patients who has glucose level higher than 271 did not get stroke in this dataset.

From Figure 11, male and female seems get same chance to get stroke, and other gender has no one get stroke in this dataset.

From Figure 12 and Figure 13, we cannot observe whether smoking status and residence type are factors towards the stroke.

Furthermore, we can see from Figure 14, the stroke cases are extremely imbalance, only 783 (1.8%) patients got stroke, other 42,617 (98.2%) patients did not get stroke. Though the relationship between the different attributes is not yet clear, the attributes gender, age, health index (current illness: hypertension, heart disease, average glucose level; BMI) and work type needs more investigation as they may not be removed due to the imbalance in stroke cases.

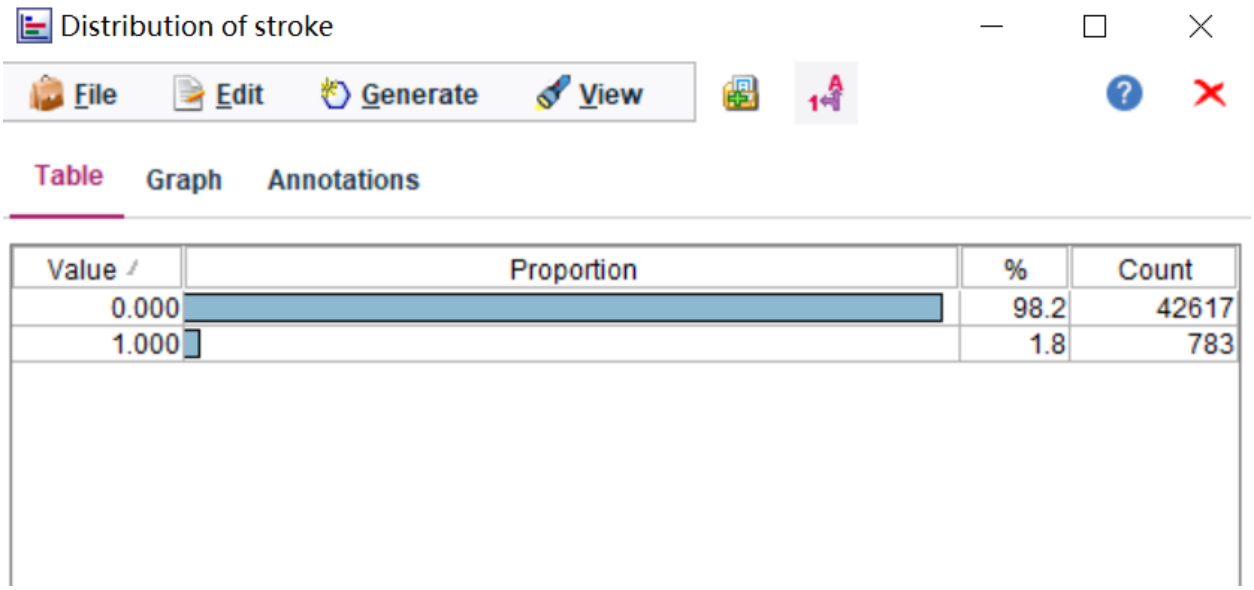
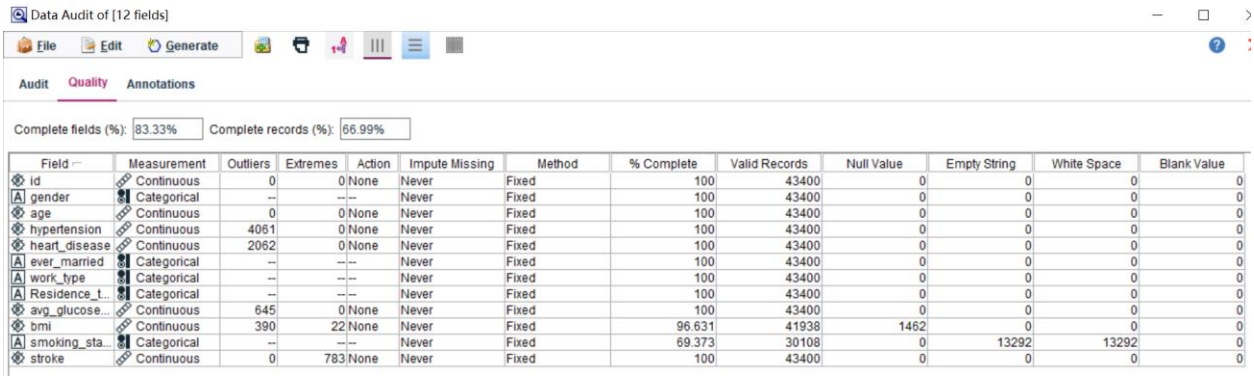


Figure 14. Distribution of stroke cases

2.5 Verifying Data Quality

Identifying the quality of the entire dataset, the percentage of attribute completion is 83.33%, and the percentage of records completion is 66.99%. The main reason for lowering the records completion rate is the incompletion of attribute smoking status of individuals, since only 69% of the patients provided their status of smoking. Also, 96.631% completion rate of BMI for

individuals also lowers the total completion rate compared to the 100% completion rate for other attributes. Because this data set is a median size dataset, for later identifying the relationship between the attribute of smoking status and the possibility of stroke, we could only use these 69 % datasets (with smoking status completed) to train the related prediction model.


Data Audit of [12 fields]

File Edit Generate

Audit Quality Annotations

Complete fields (%): 83.33% Complete records (%): 66.99%

Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value	Empty String	White Space	Blank Value
id	Continuous	0	0	None	Never	Fixed	100	43400	0	0	0	0
gender	Categorical	--	--	--	Never	Fixed	100	43400	0	0	0	0
age	Continuous	0	0	None	Never	Fixed	100	43400	0	0	0	0
hypertension	Continuous	4061	0	None	Never	Fixed	100	43400	0	0	0	0
heart_disease	Continuous	2062	0	None	Never	Fixed	100	43400	0	0	0	0
ever_married	Categorical	--	--	--	Never	Fixed	100	43400	0	0	0	0
work_type	Categorical	--	--	--	Never	Fixed	100	43400	0	0	0	0
Residence_t...	Categorical	--	--	--	Never	Fixed	100	43400	0	0	0	0
avg_glucose...	Continuous	645	0	None	Never	Fixed	100	43400	0	0	0	0
bmi	Continuous	390	22	None	Never	Fixed	96.631	41938	1462	0	0	0
smoking_sta...	Categorical	--	--	--	Never	Fixed	69.373	30108	0	13292	13292	0
stroke	Continuous	0	783	None	Never	Fixed	100	43400	0	0	0	0

Figure 15. Dataset Quality from SPSS Modeler

Based on the observation of data quality inside SPSS, we found missing data and extreme data (most likely data errors) in this database. It is not difficult to understand that in the cycle of information collection it is unlikely that perfect data will be available, due to privacy or sensitivity issues etc. And problems with the data will often be discovered when the data is analyzed as the project progresses. Here are some of the potential issues that can arise during the execution of a project.

- **Missing data.** As can be seen from this database, the missing data is obvious and expected. In the questionnaires used to collect data, privacy concerns and other concerns caused patients to fill in questionnaires without filling in detailed personal information such as BMI, marital status and place of work, among other things.
- **Data errors.** There are two possibilities for causing data errors and incorrect data entry at the time of recording. In the medical dataset used in this study, each record was manually recorded into the system by the healthcare professionals and a small amount of incorrect data entry was inevitable. In addition, machine misinterpretation, resulting in incorrect data being read in, as different machines use different ways of reading data. And most of these small amounts of incorrect data can be eliminated in most cases by filtering/processing outliers.
- **Measurement errors.** When data is measured without strict adherence to the measurement rules, small errors do occur from time to time or there is also a risk that units are reported incorrectly.

3. Data Preparation

3.1 Data Selection

Data selection, as the initial step in the data cleaning process, aims to remove some unnecessary attributes and eliminate the number of instances. It can be divided into two main parts, selecting items (row selection) and selecting attributes (column selection).

Items (row) selection: The first 3,000 instances of the initial dataset will be used to have a first taste of how biased the dataset is as below in Figure 16. The results are categorized as 0 and 1, with 0 indicating no stroke and 1 indicating the occurrence of a stroke. However, there is a significant imbalance between two categories, as only 48 instances were marked as stoke and the remaining 2,952 records were cited as no stoke as below in Figure 17.

Because the overall dataset was too large, it was considered that running the entire dataset and building the model later on would add unnecessary expense (time and cost of purchasing computer memory). We therefore selected a representative sample of 3,000 records from this to represent the characteristics of the entire dataset. The data analysis and data mined from these 3,000 cases is sufficiently educational that if the original dataset needs to be run, we will use a more powerful computer and server later, but with a model that is consistent with the 3,000 data.

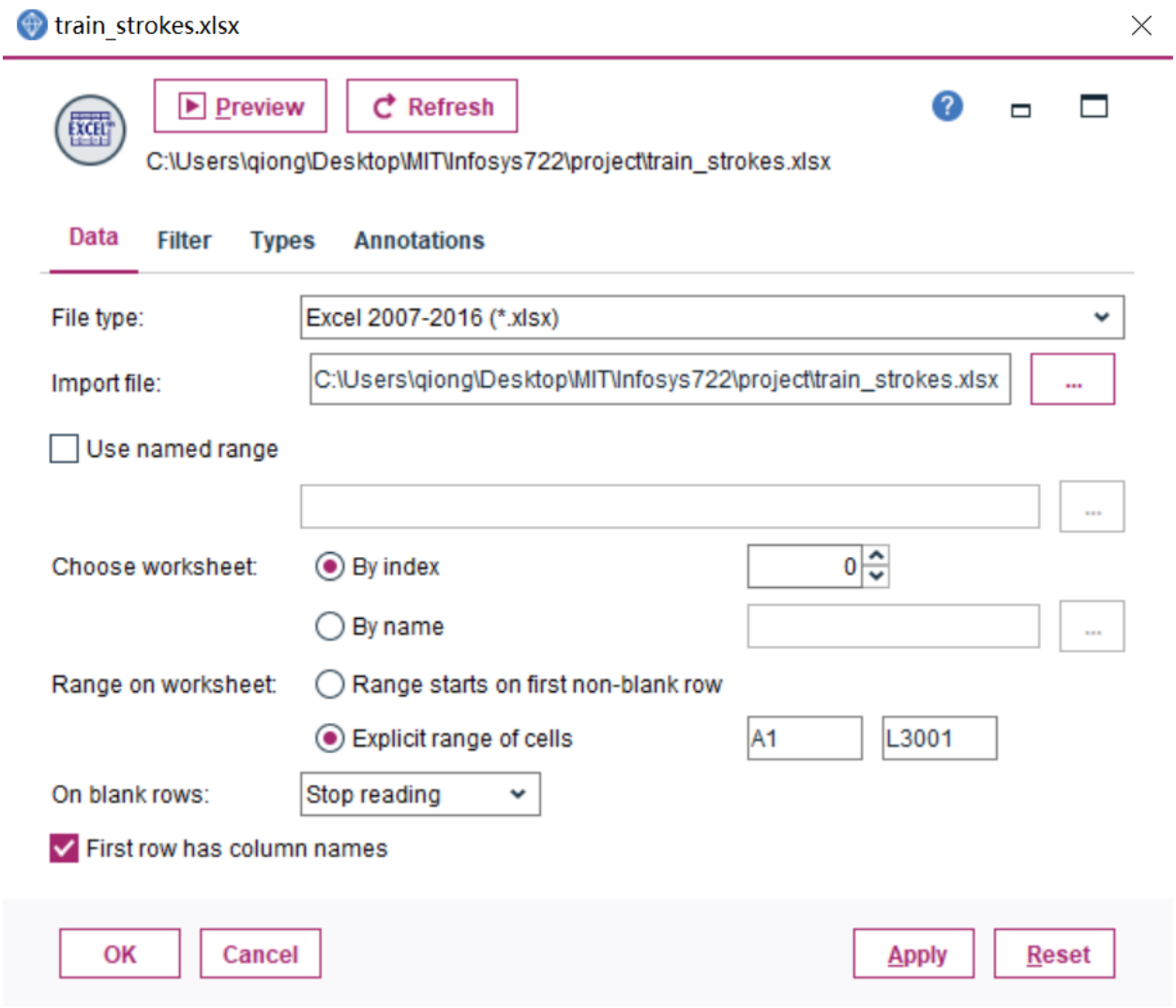


Figure 16. Items (row) selection of the dataset

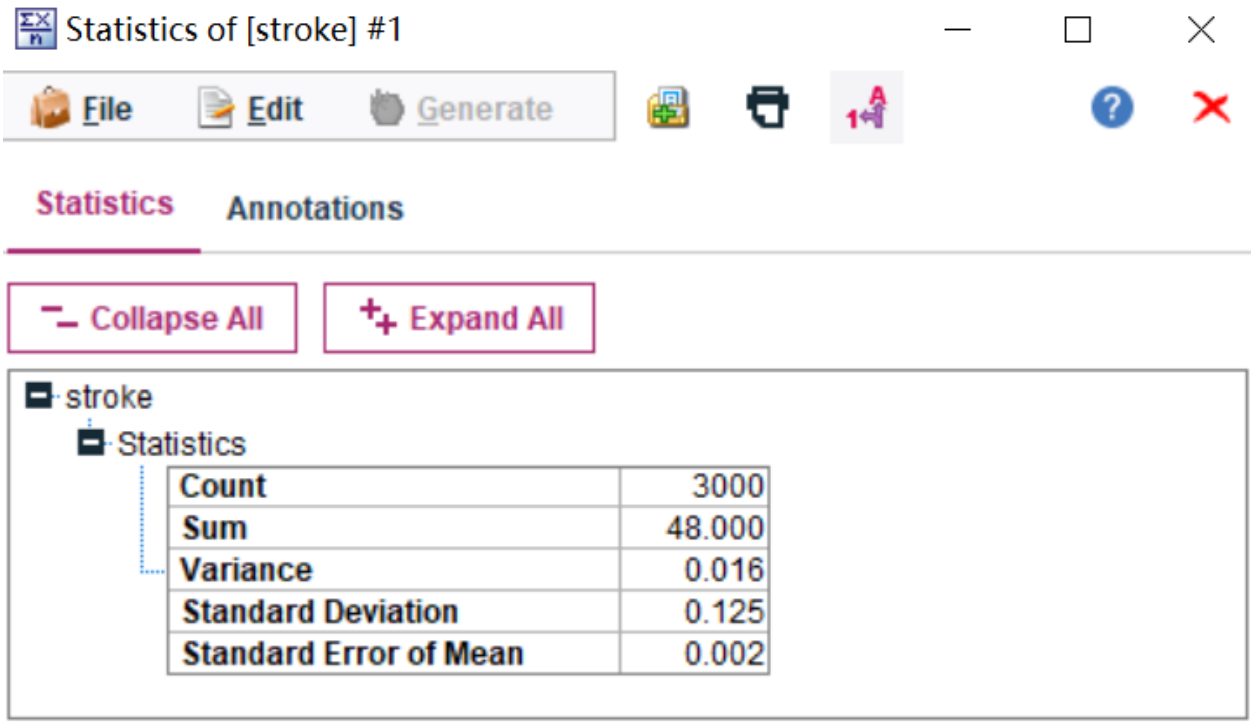


Figure 17. Analyze the target attribute based on Items (row) selection

Attribute (column) selection: According to (Barnett & H, 2005; Chong, 2020; Wajngarten & Silva, 2019), risk factors for stroke include heart disease, hypertension, smoking habits, diabetes, and obesity, but of these risk factors, the type of dwelling is not one of them.

Furthermore, according to section 2.4, some attributes contribute less to stroke, which means that they are less critical attributes. The presence of these values has the potential to influence the judgements and patterns of the algorithm. Once these tributes have been removed, the algorithm will start to move along the right track towards the correct and robust pattern.

As we did data exploration in section 2.4, residence type is not a risk, which has no effect on the outcome that ultimately leads to stroke. Therefore, it can be removed from further studies. Figure 18 below illustrates the attribute selection process.

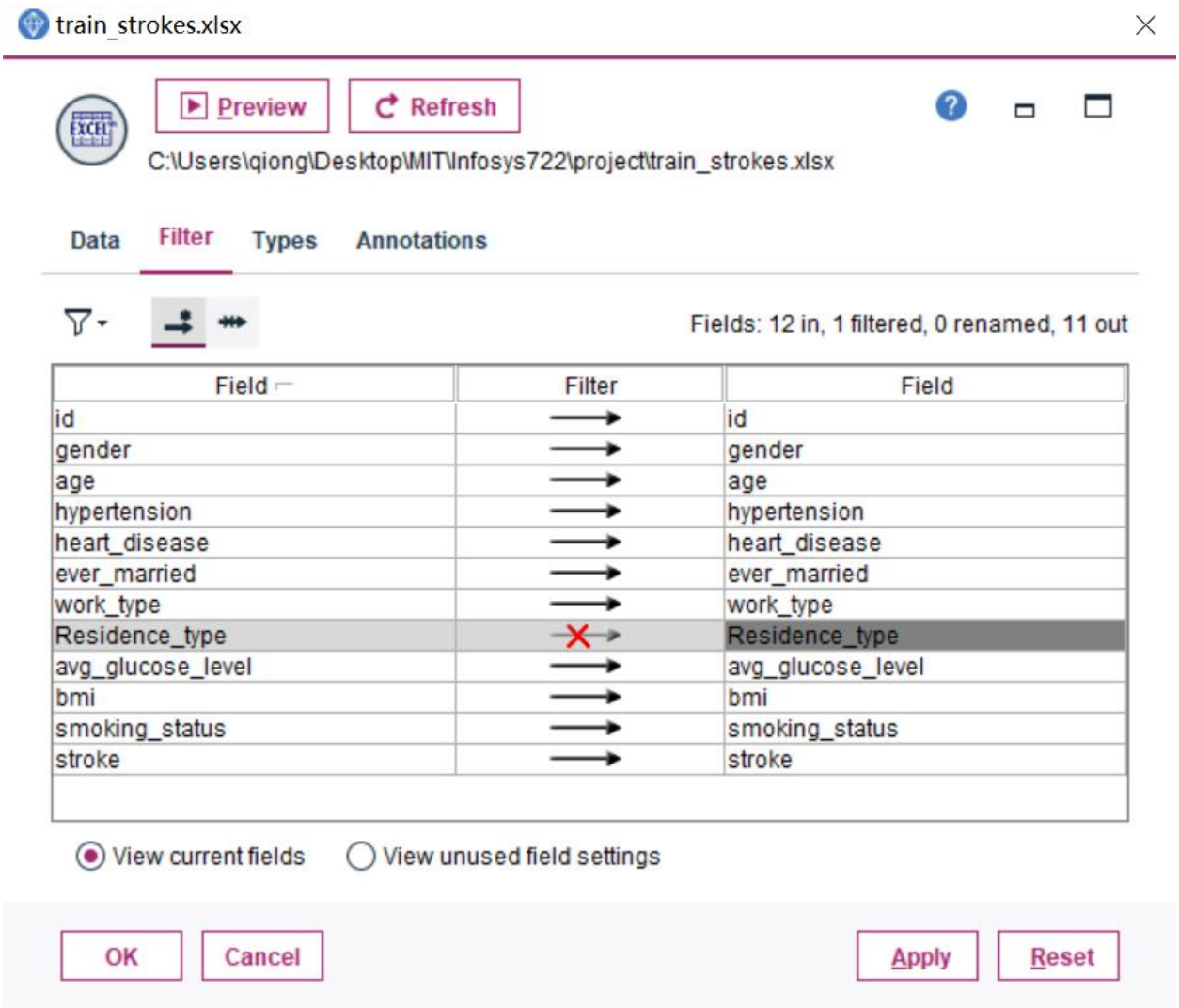
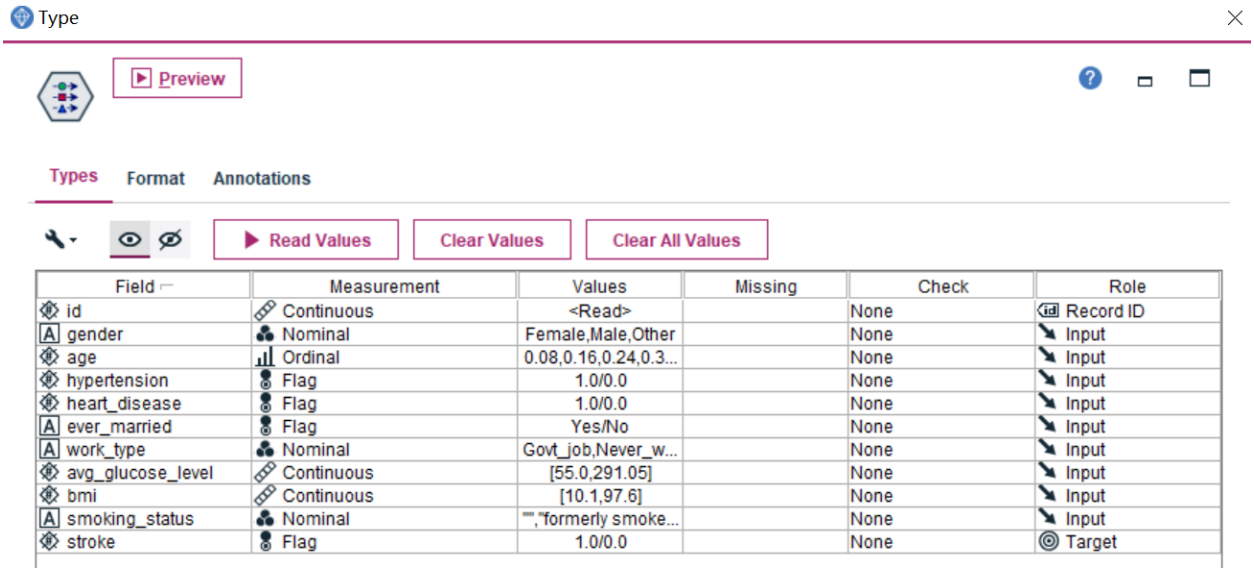


Figure 18. Attribute (column) selection

3.2 Cleaning Data

The original files of the dataset had gaps and unfinished areas and these errors were often missed or deliberately ignored by those who did not record them in the dataset due to privacy concerns or other reasons of consideration. Two types of errors were found in the processed data, missing values and data errors. These values will be removed or replaced in preparation for further use of the data.

Firstly, the measurement of each attribute is re-identified to clarify the accurate processing of dataset as Figure 19 below.



Field	Measurement	Values	Missing	Check	Role
id	Continuous	<Read>		None	Record ID
gender	Nominal	Female,Male,Other		None	Input
age	Ordinal	0.08,0.16,0.24,0.3...		None	Input
hypertension	Flag	1.0/0.0		None	Input
heart_disease	Flag	1.0/0.0		None	Input
ever_married	Flag	Yes/No		None	Input
work_type	Nominal	Govt_job,Never_w...		None	Input
avg_glucose_level	Continuous	[55.0,291.05]		None	Input
bmi	Continuous	[10.1,97.6]		None	Input
smoking_status	Nominal	"","formerly smoke...		None	Input
stroke	Flag	1.0/0.0		None	Target

Figure 19. Re-identify measurement for each attribute

Missing values:

A total of two attributes were found to contain missing values, with integrity values of 96.631% for BMI and 69.373% for smoking status. The inclusion of 14 missing values can reduce the efficiency of the module's execution. Therefore, they need to be removed before being applied to any model. This step eliminates the problem when applying data to certain algorithms, making the algorithm less skewed in the wrong direction.

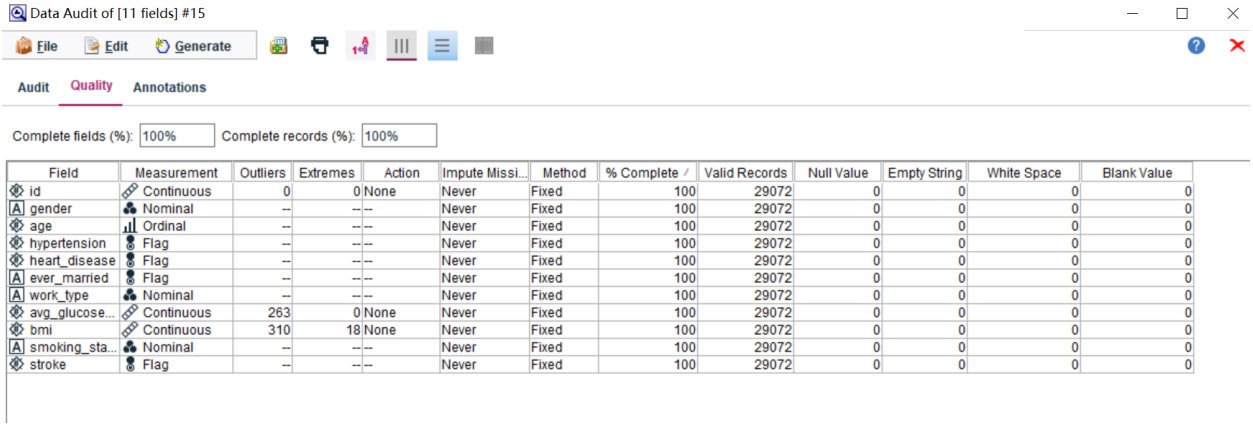
A null or empty value is not very compatible with some algorithms. It is usually represented in many different forms, such as empty cells as well as N/A markers. In this dataset, both smoking status and BMI contain these missing values. There are three ways to remove them: directly from the record or populated according to some pattern (usually copied from the last populated instance) or populated with a fixed number derived from the averaging algorithm.

For the missing BMI and smoking status values in this project, the best approach is to delete these instances directly, because there are three smoking statuses and the missing values occupy 31% of the existing dataset, which is too large a proportion, and filling the values randomly or copying from later instances would likely cause the dataset to lose its original characteristics and lead to much less accurate prediction results.

The BMI value is also difficult to fill with a fixed number and it is also difficult to fill with the following pattern. Since most of the instances in this empty subset were marked as non-stroke, deleting these instances would not have a significant impact on the shortage of stroke cases.

Therefore, it was decided to remove these instances where the missing values for BMI and smoking status were located. Figure 20 show the results of removing the null values from the BMI and smoking status attribute.

Selected dataset after removing missing values as Figure 20 below:



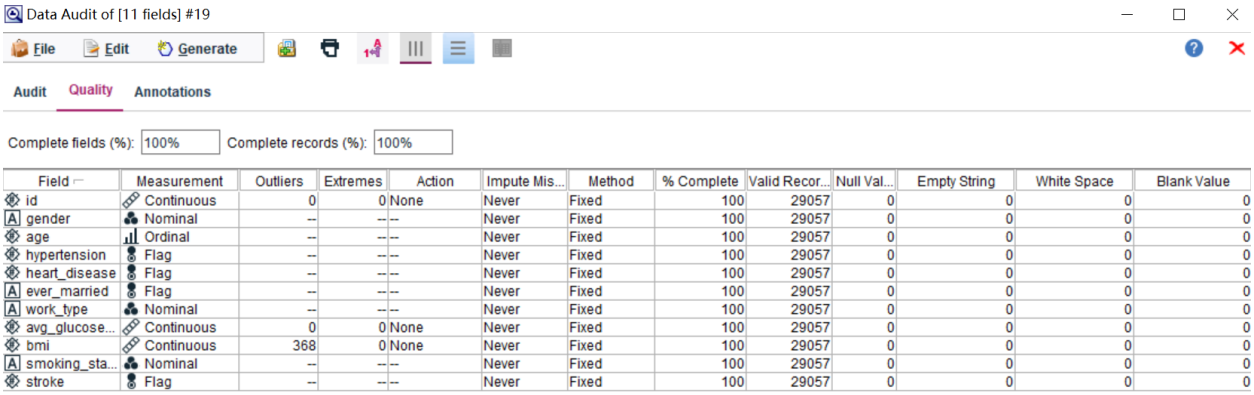
Field	Measurement	Outliers	Extremes	Action	Impute Missi...	Method	% Complete	Valid Records	Null Value	Empty String	White Space	Blank Value
id	Continuous	0	0	None	Never	Fixed	100	29072	0	0	0	0
gender	Nominal	--	--	--	Never	Fixed	100	29072	0	0	0	0
age	Ordinal	--	--	--	Never	Fixed	100	29072	0	0	0	0
hypertension	Flag	--	--	--	Never	Fixed	100	29072	0	0	0	0
heart_disease	Flag	--	--	--	Never	Fixed	100	29072	0	0	0	0
ever_married	Flag	--	--	--	Never	Fixed	100	29072	0	0	0	0
work_type	Nominal	--	--	--	Never	Fixed	100	29072	0	0	0	0
avg_glucose...	Continuous	263	0	None	Never	Fixed	100	29072	0	0	0	0
bmi	Continuous	310	18	None	Never	Fixed	100	29072	0	0	0	0
smoking_sta...	Nominal	--	--	--	Never	Fixed	100	29072	0	0	0	0
stroke	Flag	--	--	--	Never	Fixed	100	29072	0	0	0	0

Figure 20. Dataset after cleaning missing values

The method used to remove the missing values is to generate a missing select node, which selects all valid values and filters out those uncompleted values in BMI and smoking status values(rows). After this step, there are 2,009 records left, which means 991 records are reduced. When this algorithm is applied, the completeness of all attributes is increased to 100%.

Data Errors:

The only data errors contained in the file are extreme values, which can affect the module's predictions as the module may be biased towards outliers as it tries to cover many different possible values. From Figure 6, we know that there are 18 extremes. These extreme values are removed using forced deletion. This method will attempt to drag the values from the extreme range to a reasonable range within the range. The dataset after deleting the extreme values as Figure 21 below.



Field	Measurement	Outliers	Extremes	Action	Impute Mis...	Method	% Complete	Valid Recor...	Null Val...	Empty String	White Space	Blank Value
id	Continuous	0	0	None	Never	Fixed	100	29057	0	0	0	0
gender	Nominal	--	--	--	Never	Fixed	100	29057	0	0	0	0
age	Ordinal	--	--	--	Never	Fixed	100	29057	0	0	0	0
hypertension	Flag	--	--	--	Never	Fixed	100	29057	0	0	0	0
heart_disease	Flag	--	--	--	Never	Fixed	100	29057	0	0	0	0
ever_married	Flag	--	--	--	Never	Fixed	100	29057	0	0	0	0
work_type	Nominal	--	--	--	Never	Fixed	100	29057	0	0	0	0
avg_glucose...	Continuous	0	0	None	Never	Fixed	100	29057	0	0	0	0
bmi	Continuous	368	0	None	Never	Fixed	100	29057	0	0	0	0
smoking_sta...	Nominal	--	--	--	Never	Fixed	100	29057	0	0	0	0
stroke	Flag	--	--	--	Never	Fixed	100	29057	0	0	0	0

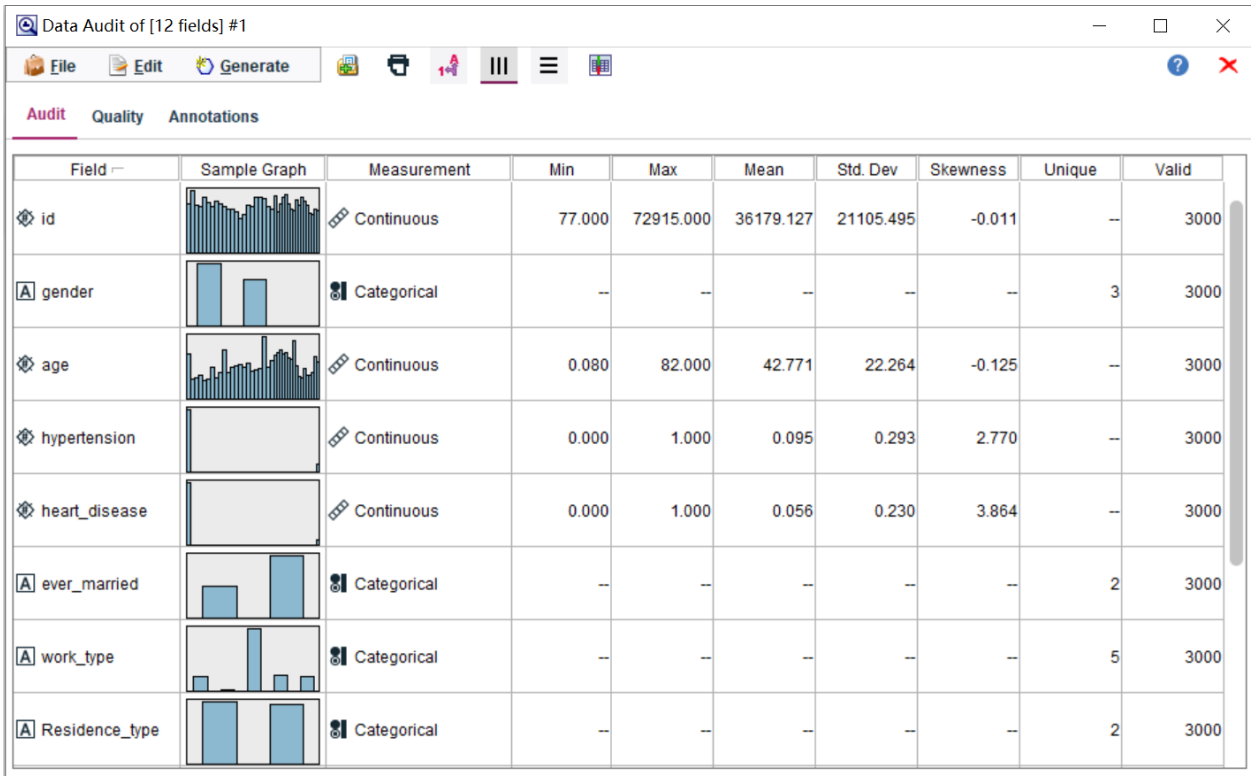
Figure 21. Dataset after cleaning extreme values

The method used to remove the outliers is to use extreme/ outlier node to calculate the deviation of the value with normal distribution, which selects all values in the rage of normal distribution and filters out those outlier values in BMI and average glucose(rows). There are 2,007 records left, and 2 records deleted.

From Figure 22 and Figure 23, we can see the difference of before cleaning dataset and after cleaning dataset.

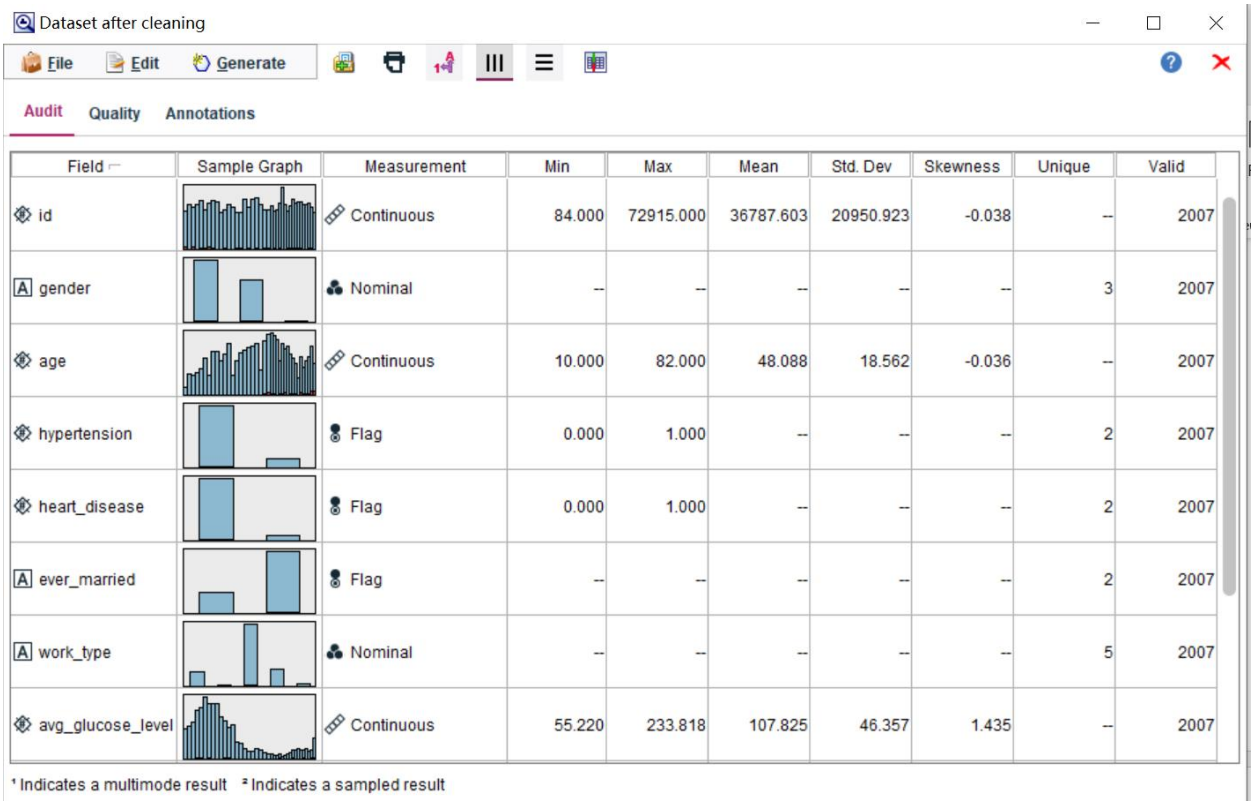
Figure 22 is showing the data audit before data cleaning.

Figure 23 is showing the data audit after data cleaning.



Field	Sample Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
id		Continuous	77.000	72915.000	36179.127	21105.495	-0.011	--	3000
gender		Categorical	--	--	--	--	--	3	3000
age		Continuous	0.080	82.000	42.771	22.264	-0.125	--	3000
hypertension		Continuous	0.000	1.000	0.095	0.293	2.770	--	3000
heart_disease		Continuous	0.000	1.000	0.056	0.230	3.864	--	3000
ever_married		Categorical	--	--	--	--	--	2	3000
work_type		Categorical	--	--	--	--	--	5	3000
Residence_type		Categorical	--	--	--	--	--	2	3000

Figure 22. Dataset before cleaning



Field	Sample Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
id		Continuous	84.000	72915.000	36787.603	20950.923	-0.038	--	2007
gender		Nominal	--	--	--	--	--	3	2007
age		Continuous	10.000	82.000	48.088	18.562	-0.036	--	2007
hypertension		Flag	0.000	1.000	--	--	--	2	2007
heart_disease		Flag	0.000	1.000	--	--	--	2	2007
ever_married		Flag	--	--	--	--	--	2	2007
work_type		Nominal	--	--	--	--	--	5	2007
avg_glucose_level		Continuous	55.220	233.818	107.825	46.357	1.435	--	2007

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

Figure 23. Dataset after cleaning

3.3 Constructing New Data

To construct new data by creating new features, here three new features are created, Smoking habit, BMI, and Work category. Smoking habit is set as a flag type from the nominal type of smoking status, the BMI is re-grouped by a certain range of values and the work category is re-classified by general knowledge base respectively.

1. **The smoking status**
   The initial smoking status includes formerly smoked, smoking, and never smoking. Studies show whether smoke or not is a key predictor of our target attribute stroke. Therefore, smoking status is flagged as smoking and never smoking, as Figure 24 and Figure 25 shows.

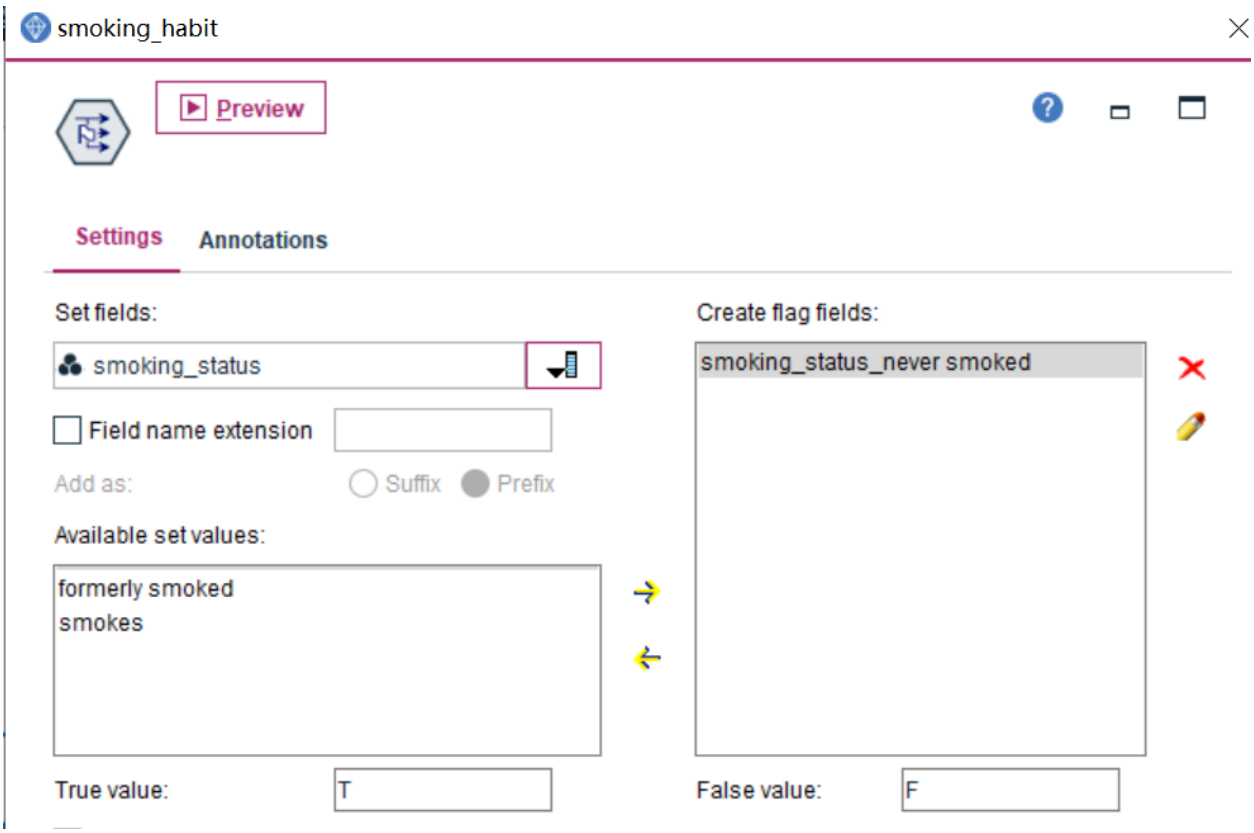


Figure 24. Smoking habit

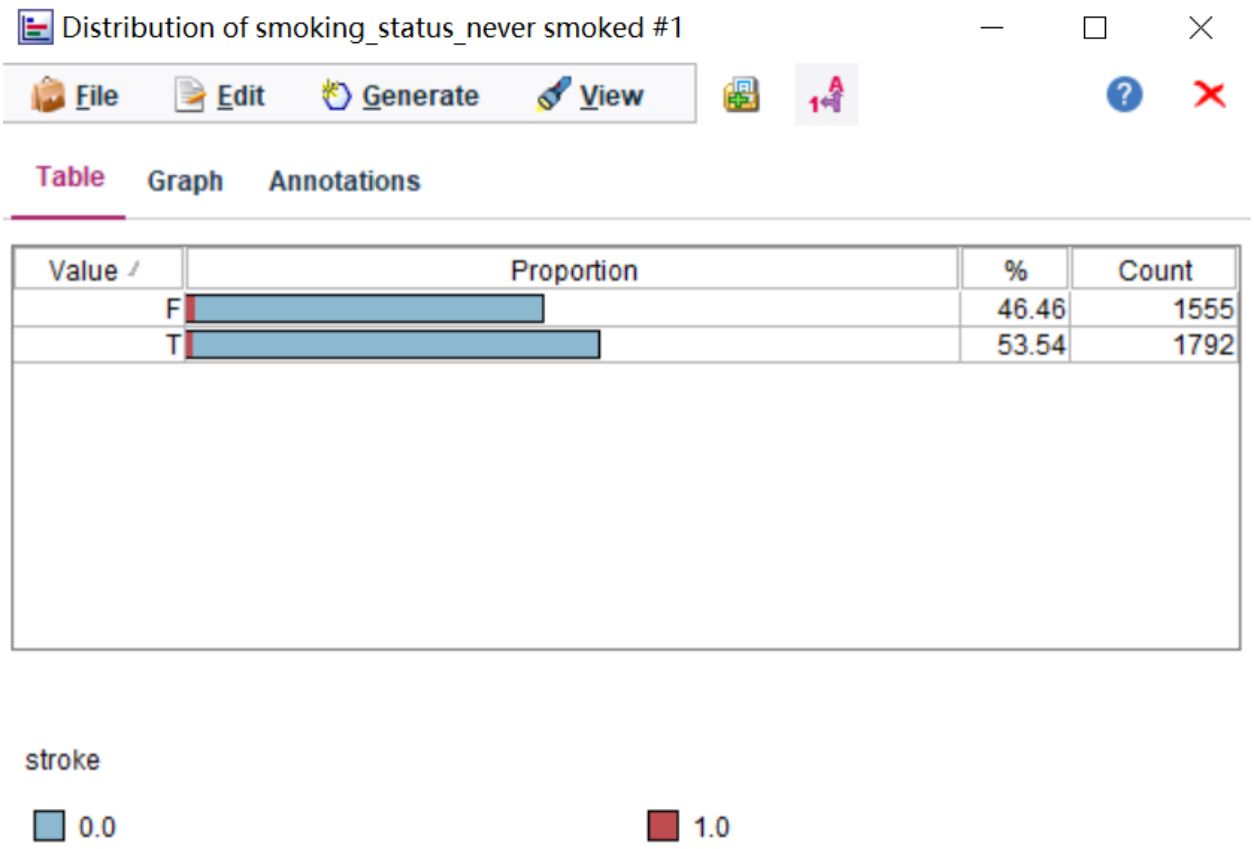


Figure 25. New Smoking status after set a flag

2. **BMI groups:**
   BMI stands for body mass index and is a simple calculation using a person's height and weight. The formula is BMI = kg/m2, where kg is a person's weight in kilograms and m2 is the square of their height in meters. A BMI of 25.0 or higher is considered overweight, while a healthy range is 18.5 to 24.9. And one of the risk factors for stroke, the target attribute of this study, is BMI. for these reasons, I created the characteristics of the weight categories based on the BMI results.

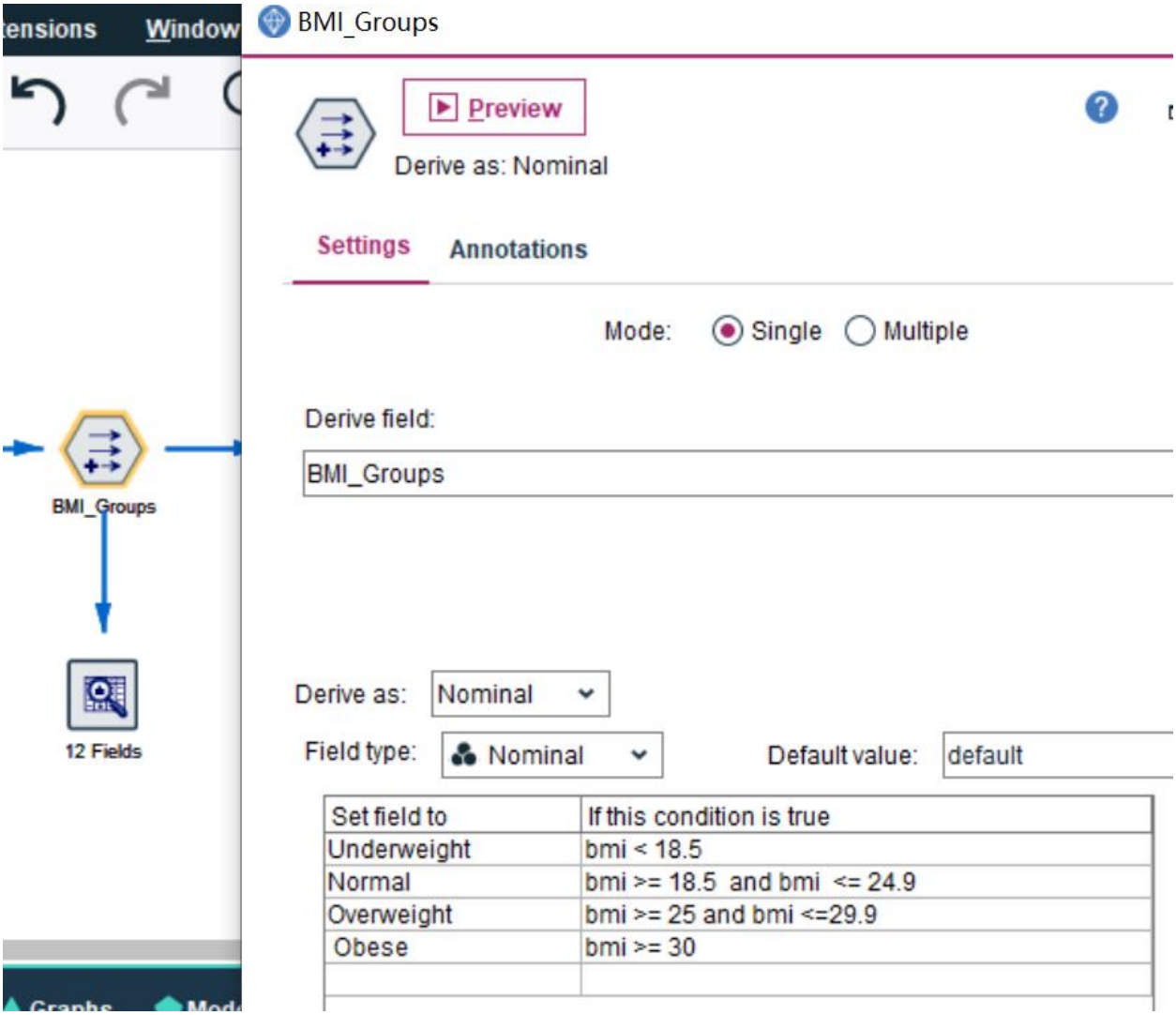

ensions
Window
BMI_Groups
Preview
Derive as: Nominal
Settings
Annotations
Mode:
Single
Multiple
Derive field:
BMI_Groups
BMI_Groups
12 Fields
Derive as:
Nominal
Field type:
Nominal
Default value:
default
Set field to
If this condition is true
Underweight
bmi < 18.5
Normal
bmi >= 18.5 and bmi <= 24.9
Overweight
bmi >= 25 and bmi <=29.9
Obese
bmi >= 30

Figure 26. Set BMI groups

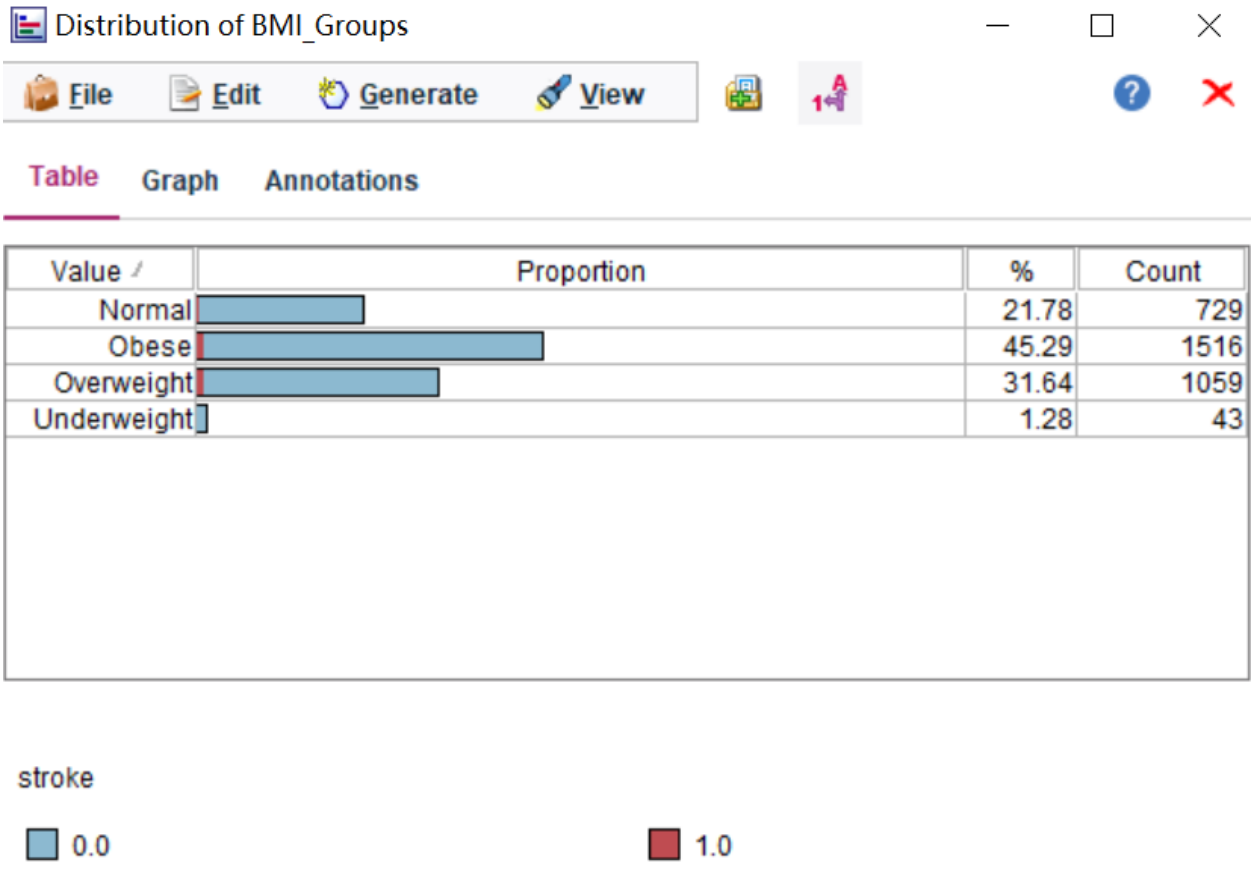


Figure 27. New BMI groups

3. **Work categories:**
   The work categories initially include five categories: children, never worked, self-employed, government jobs, and private. Some of these types should be grouped together because these job categories have a high degree of similarity to each other. Also, by merging these similar job types, the complexity of the module is reduced at the same time. It was decided to combine the categories Child and Never Worked into Never Worked and Self-Employed and Private Company Work into self-employed, as shown in Figure 28 and Figure 29.

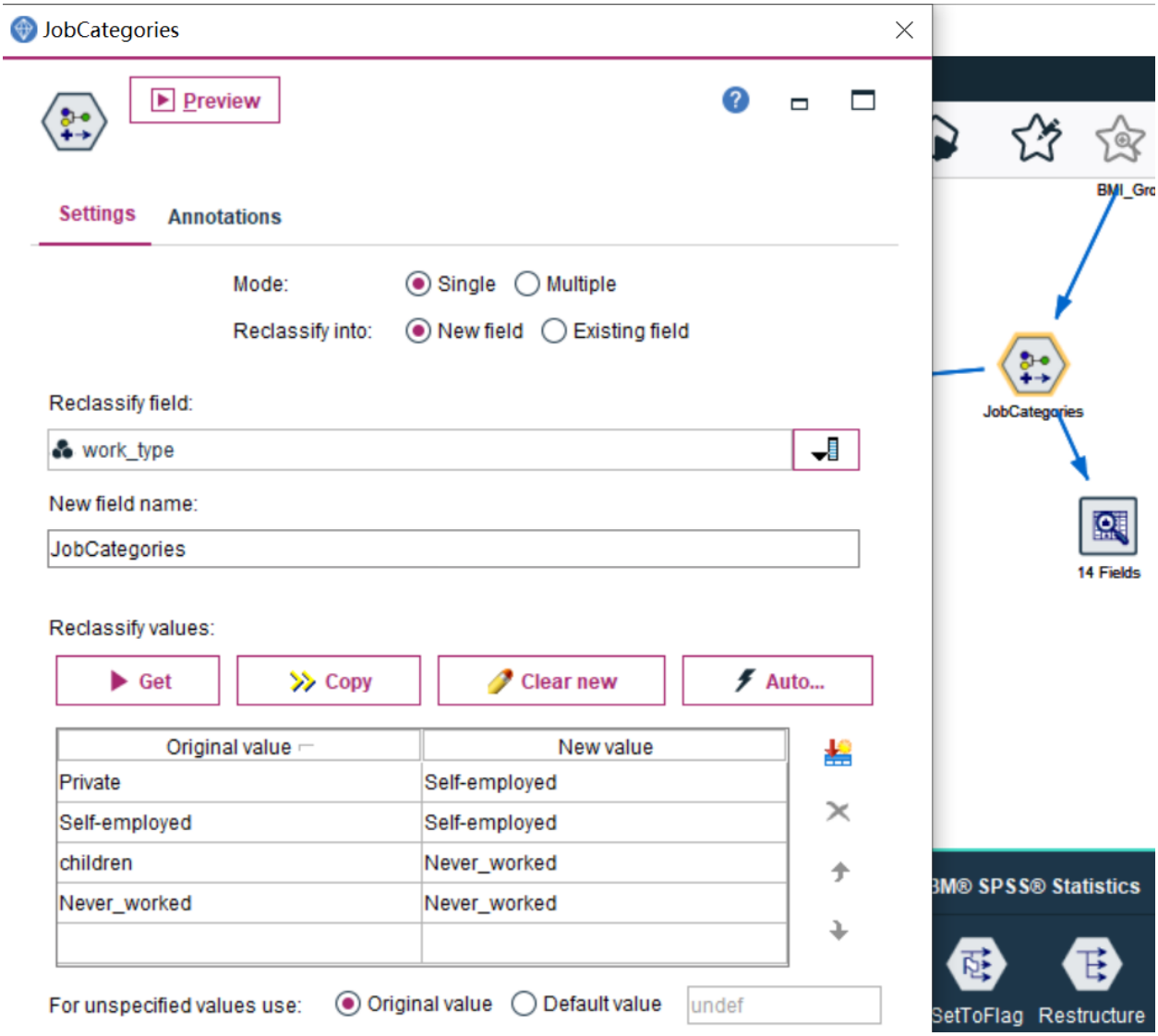

JobCategories
Preview
Settings
Annotations
Mode:
Single
Multiple
Reclassify into:
New field
Existing field
Reclassify field:
work_type
New field name:
JobCategories
Reclassify values:
Get
Copy
Clear new
Auto...
Original value
New value
Private
Self-employed
Self-employed
Self-employed
children
Never_worked
Never_worked
Never_worked
For unspecified values use:
Original value
Default value
undef
BMI_Gro
JobCategories
14 Fields
BM® SPSS® Statistics
SetToFlag
Restructure

Figure 28. Classify Work categories

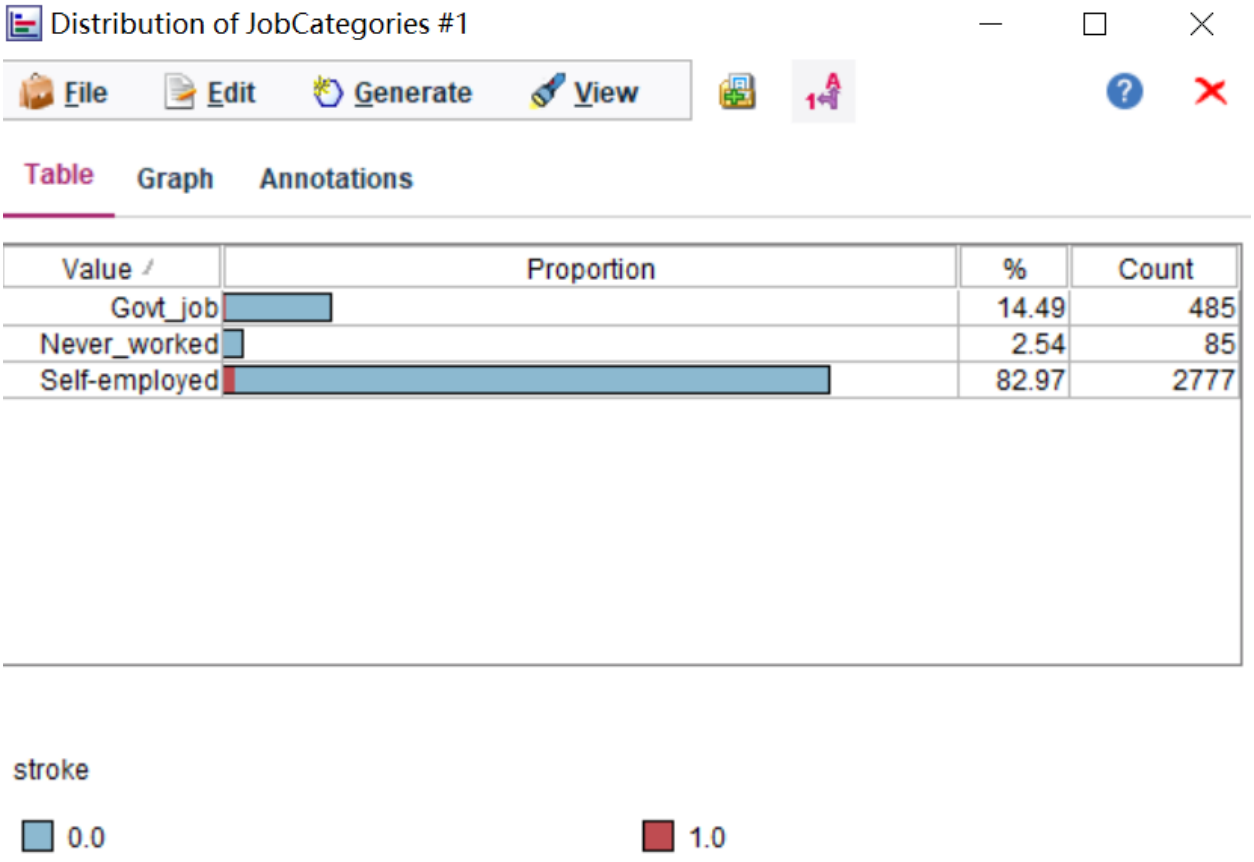


Figure 29. New Work categories

3.4 Integrating Data

Following the data cleaning process and constructing new features, the data cleaning process succeeded in removing the existing extreme and missing values. However, it reduced the overall quality of the data as most of this information was generated algorithmically rather than the actual values. Therefore, adding a further 2,000 instances would provide more accuracy to the projected values and ultimately result in better data quality. These 2,000 instances have the same attributes and the same sequence, as these 2,000 instances are from raw dataset as well. A merge path is created successfully as shown in Figure 30.

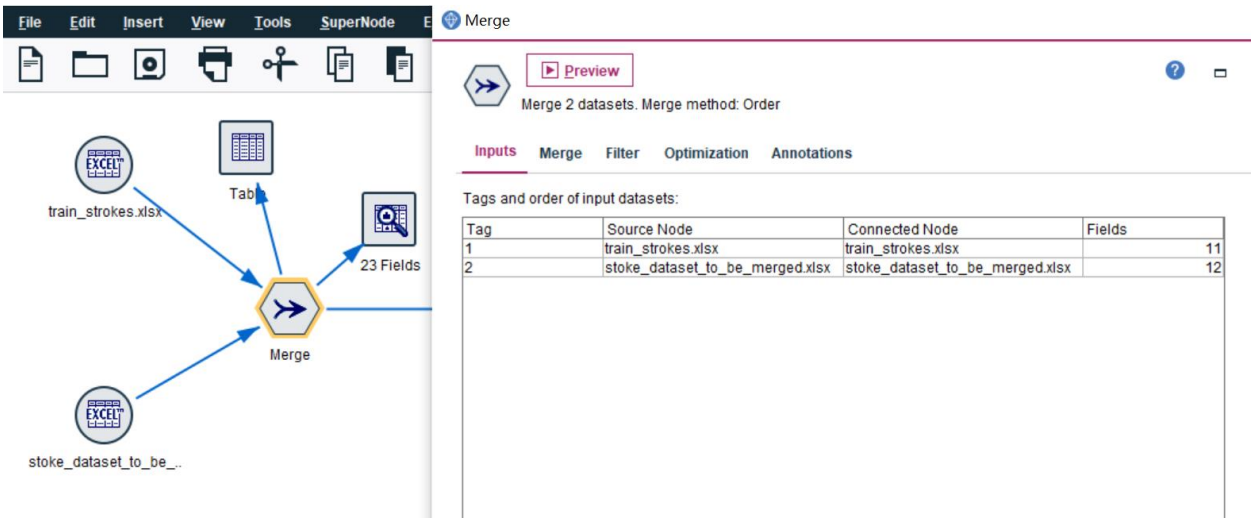


Figure 30. Merge two datasets

After merging two datasets successfully, we also delete the attribute of residence type, as we did for the initial dataset.

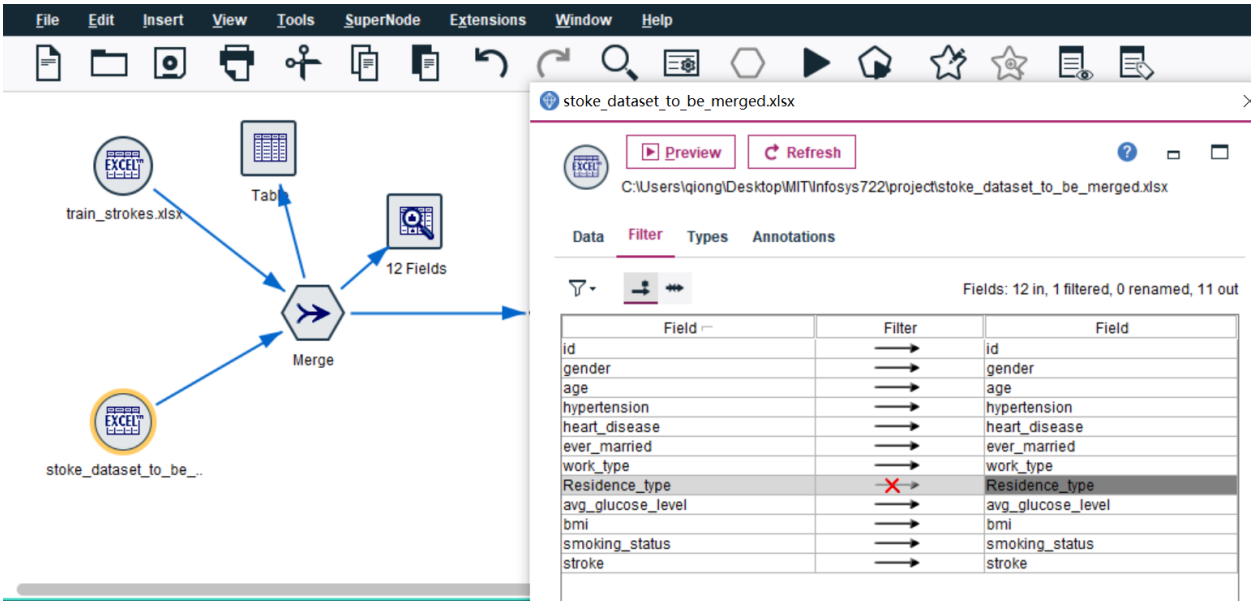


Figure 31. Attribute (column) selection for merged dataset

Then, we do the same step of data cleaning as we did for the initial dataset.

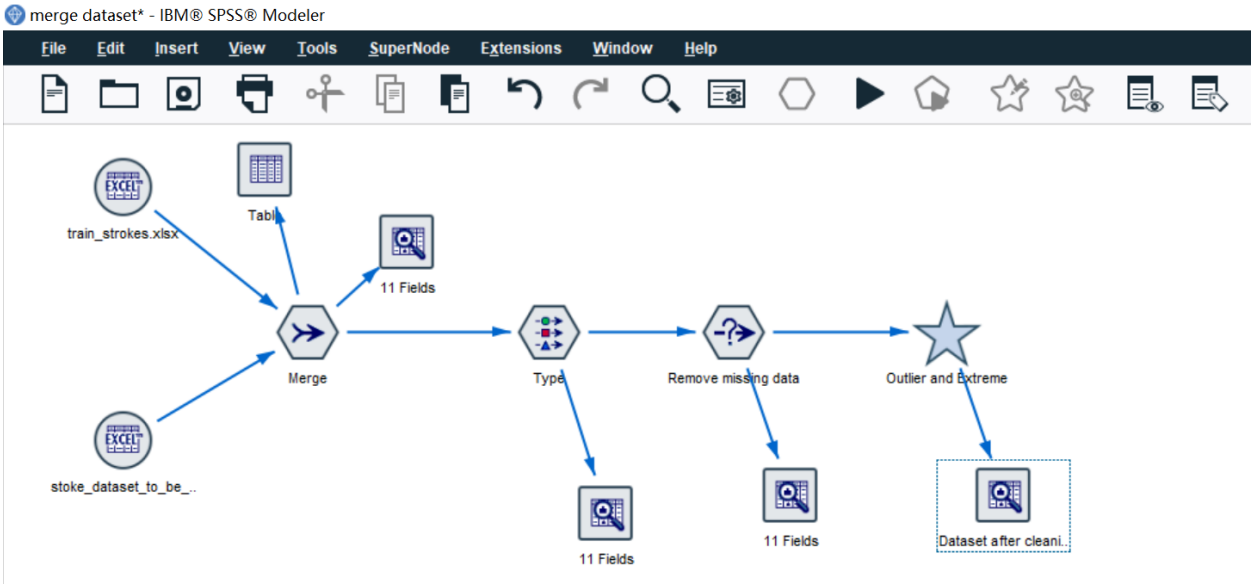


Figure 32. Completed stream output of merging data

3.5 Formatting Data

Data formatting, or primarily sorting the data, is the final step in data cleaning. The raw data may have an uneven distribution of our target attributes, especially with the most recently merged database. For the most recently joined database, there was a major problem with the ordering, where all stroke patients were listed in the first 249 positions, showing a value of 1 in the attribute, while records further back were for people who had a stroke.

Although this database does not sequence sensitive to the model used, we still need to format the implementation data to better distribute the data and allow for successful training and test splitting, otherwise, the model may be overfitted when training the first 249 data. In the case here, by looking at the dataset it was found that each individual's identification number (ID) was randomly assigned, and the sequence had no effect on the outcome attribute of interest or the rest of the records. Therefore, the individual's identification number will be used to sort the data and it will appear in ascending order.

Figure 33 is showing the sequence of data before sorting.

Figure 34 is showing the sequence of data after sorting.

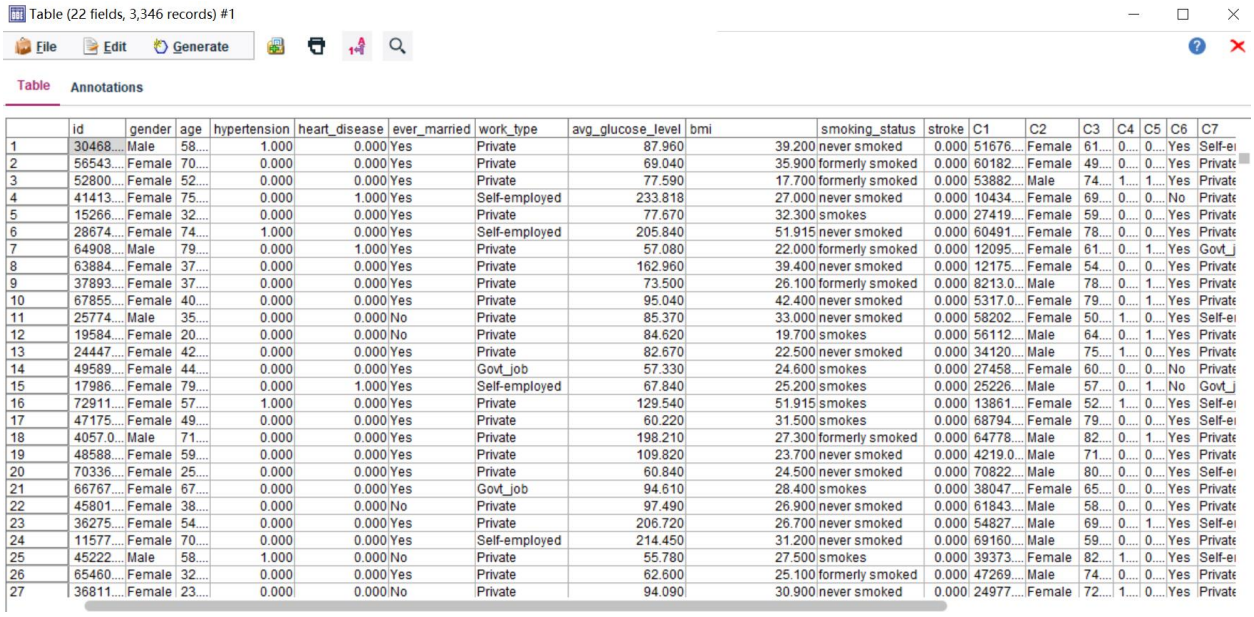


Table (22 fields, 3,346 records) #1

File Edit Generate

Table Annotations

	id	gender	age	hypertension	heart_disease	ever_married	work_type	avg_glucose_level	bmi	smoking_status	stroke	C1	C2	C3	C4	C5	C6	C7
1	30468....	Male	58....	1.000	0.000	Yes	Private	87.960	39.200	never smoked	0.000	51676....	Female	61....	0....	0....	Yes	Self-e
2	56543....	Female	70....	0.000	0.000	Yes	Private	69.040	35.900	formerly smoked	0.000	60182....	Female	49....	0....	0....	Yes	Private
3	52800....	Female	52....	0.000	0.000	Yes	Private	77.590	17.700	formerly smoked	0.000	53882....	Male	74....	1....	1....	Yes	Private
4	41413....	Female	75....	0.000	1.000	Yes	Self-employed	233.818	27.000	never smoked	0.000	10434....	Female	69....	0....	0....	No	Private
5	15266....	Female	32....	0.000	0.000	Yes	Private	77.670	32.300	smokes	0.000	27419....	Female	59....	0....	0....	Yes	Private
6	28674....	Female	74....	1.000	0.000	Yes	Self-employed	205.840	51.915	never smoked	0.000	60491....	Female	78....	0....	0....	Yes	Private
7	64908....	Male	79....	0.000	1.000	Yes	Private	57.080	22.000	formerly smoked	0.000	12095....	Female	61....	0....	1....	Yes	Govt_j
8	63884....	Female	37....	0.000	0.000	Yes	Private	162.960	39.400	never smoked	0.000	12175....	Female	54....	0....	0....	Yes	Private
9	37893....	Female	37....	0.000	0.000	Yes	Private	73.500	26.100	formerly smoked	0.000	8213.0....	Male	78....	0....	1....	Yes	Private
10	67855....	Female	40....	0.000	0.000	Yes	Private	95.040	42.400	never smoked	0.000	5317.0....	Female	79....	0....	1....	Yes	Private
11	25774....	Male	35....	0.000	0.000	No	Private	85.370	33.000	never smoked	0.000	58202....	Female	50....	1....	0....	Yes	Self-e
12	19584....	Female	20....	0.000	0.000	No	Private	84.620	19.700	smokes	0.000	56112....	Male	64....	0....	1....	Yes	Private
13	24447....	Female	42....	0.000	0.000	Yes	Private	82.670	22.500	never smoked	0.000	34120....	Male	75....	1....	0....	Yes	Private
14	49589....	Female	44....	0.000	0.000	Yes	Govt_job	57.330	24.600	smokes	0.000	27458....	Female	60....	0....	0....	No	Private
15	17986....	Female	79....	0.000	1.000	Yes	Self-employed	67.840	25.200	smokes	0.000	25226....	Male	57....	0....	1....	No	Govt_j
16	72911....	Female	57....	1.000	0.000	Yes	Private	129.540	51.915	smokes	0.000	13861....	Female	52....	1....	0....	Yes	Self-e
17	47175....	Female	49....	0.000	0.000	Yes	Private	60.220	31.500	smokes	0.000	68794....	Female	79....	0....	0....	Yes	Self-e
18	4057.0....	Male	71....	0.000	0.000	Yes	Private	198.210	27.300	formerly smoked	0.000	64778....	Male	82....	0....	1....	Yes	Private
19	48588....	Female	59....	0.000	0.000	Yes	Private	109.820	23.700	never smoked	0.000	4219.0....	Male	71....	0....	0....	Yes	Private
20	70336....	Female	25....	0.000	0.000	Yes	Private	60.840	24.500	never smoked	0.000	70822....	Male	80....	0....	0....	Yes	Self-e
21	66767....	Female	67....	0.000	0.000	Yes	Govt_job	94.610	28.400	smokes	0.000	38047....	Female	65....	0....	0....	Yes	Private
22	45801....	Female	38....	0.000	0.000	No	Private	97.490	26.900	never smoked	0.000	61843....	Male	58....	0....	0....	Yes	Private
23	36275....	Female	54....	0.000	0.000	Yes	Private	206.720	26.700	never smoked	0.000	54827....	Male	69....	0....	1....	Yes	Self-e
24	11577....	Female	70....	0.000	0.000	Yes	Self-employed	214.450	31.200	never smoked	0.000	69160....	Male	59....	0....	0....	Yes	Private
25	45222....	Male	58....	1.000	0.000	No	Private	55.780	27.500	smokes	0.000	39373....	Female	82....	1....	0....	Yes	Self-e
26	65460....	Female	32....	0.000	0.000	Yes	Private	62.600	25.100	formerly smoked	0.000	47269....	Male	74....	0....	0....	Yes	Private
27	36811....	Female	23....	0.000	0.000	No	Private	94.090	30.900	never smoked	0.000	24977....	Female	72....	1....	0....	Yes	Private

Figure 33. the sequence of data before sorting

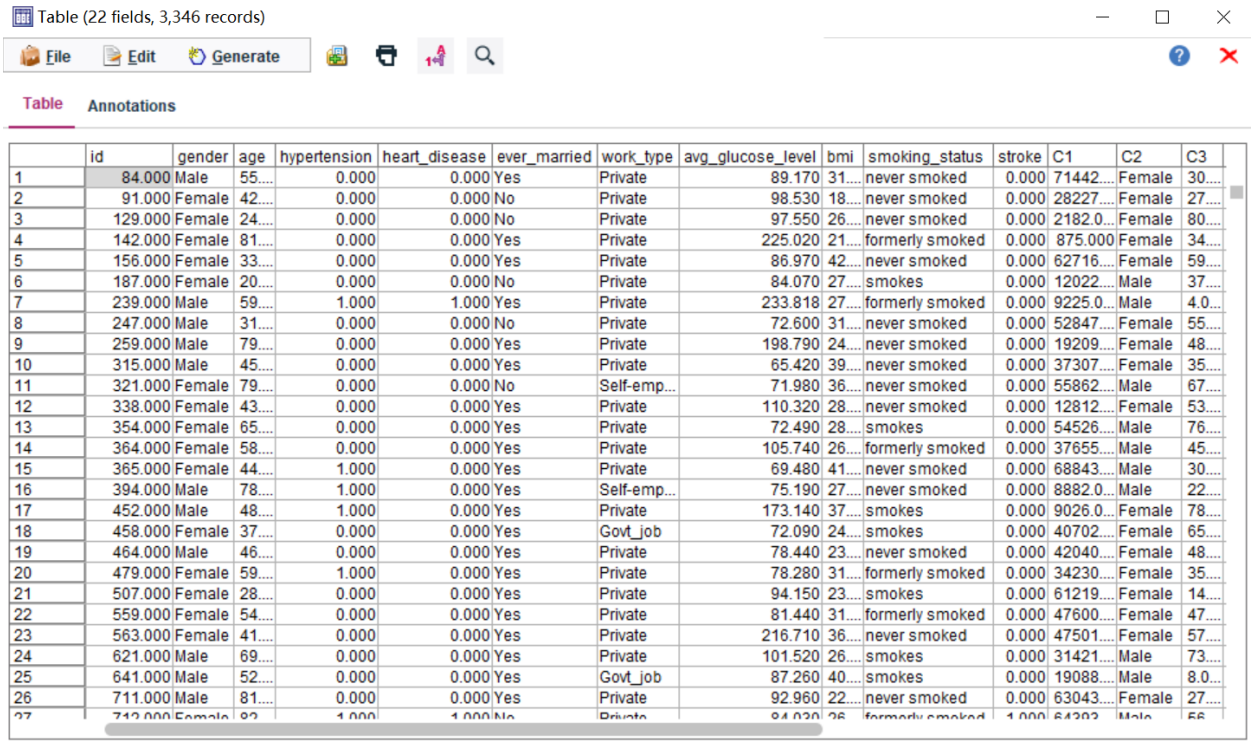


Table (22 fields, 3,346 records)

File Edit Generate

Table Annotations

	id	gender	age	hypertension	heart_disease	ever_married	work_type	avg_glucose_level	bmi	smoking_status	stroke	C1	C2	C3
1	84.000	Male	55....	0.000	0.000	Yes	Private	89.170	31....	never smoked	0.000	71442....	Female	30....
2	91.000	Female	42....	0.000	0.000	No	Private	98.530	18....	never smoked	0.000	28227....	Female	27....
3	129.000	Female	24....	0.000	0.000	No	Private	97.550	26....	never smoked	0.000	2182.0....	Female	80....
4	142.000	Female	81....	0.000	0.000	Yes	Private	225.020	21....	formerly smoked	0.000	875.000	Female	34....
5	156.000	Female	33....	0.000	0.000	Yes	Private	86.970	42....	never smoked	0.000	62716....	Female	59....
6	187.000	Female	20....	0.000	0.000	No	Private	84.070	27....	smokes	0.000	12022....	Male	37....
7	239.000	Male	59....	1.000	1.000	Yes	Private	233.818	27....	formerly smoked	0.000	9225.0....	Male	4.0....
8	247.000	Male	31....	0.000	0.000	No	Private	72.600	31....	never smoked	0.000	52847....	Female	55....
9	259.000	Male	79....	0.000	0.000	Yes	Private	198.790	24....	never smoked	0.000	19209....	Female	48....
10	315.000	Male	45....	0.000	0.000	Yes	Private	65.420	39....	never smoked	0.000	37307....	Female	35....
11	321.000	Female	79....	0.000	0.000	No	Self-emp...	71.980	36....	never smoked	0.000	55862....	Male	67....
12	338.000	Female	43....	0.000	0.000	Yes	Private	110.320	28....	never smoked	0.000	12812....	Female	53....
13	354.000	Female	65....	0.000	0.000	Yes	Private	72.490	28....	smokes	0.000	54526....	Male	76....
14	364.000	Female	58....	0.000	0.000	Yes	Private	105.740	26....	formerly smoked	0.000	37655....	Male	45....
15	365.000	Female	44....	1.000	0.000	Yes	Private	69.480	41....	never smoked	0.000	68843....	Male	30....
16	394.000	Male	78....	1.000	0.000	Yes	Self-emp...	75.190	27....	never smoked	0.000	8882.0....	Male	22....
17	452.000	Male	48....	1.000	0.000	Yes	Private	173.140	37....	smokes	0.000	9026.0....	Female	78....
18	458.000	Female	37....	0.000	0.000	Yes	Govt_job	72.090	24....	smokes	0.000	40702....	Female	65....
19	464.000	Male	46....	0.000	0.000	Yes	Private	78.440	23....	never smoked	0.000	42040....	Female	48....
20	479.000	Female	59....	1.000	0.000	Yes	Private	78.280	31....	formerly smoked	0.000	34230....	Female	35....
21	507.000	Female	28....	0.000	0.000	Yes	Private	94.150	23....	smokes	0.000	61219....	Female	14....
22	559.000	Female	54....	0.000	0.000	Yes	Private	81.440	31....	formerly smoked	0.000	47600....	Female	47....
23	563.000	Female	41....	0.000	0.000	Yes	Private	216.710	36....	never smoked	0.000	47501....	Female	57....
24	621.000	Male	69....	0.000	0.000	Yes	Private	101.520	26....	smokes	0.000	31421....	Male	73....
25	641.000	Male	52....	0.000	0.000	Yes	Govt_job	87.260	40....	smokes	0.000	19088....	Male	8.0....
26	711.000	Male	81....	0.000	0.000	Yes	Private	92.960	22....	never smoked	0.000	63043....	Female	27....
27	712.000	Female	82....	1.000	1.000	No	Private	84.030	26....	formerly smoked	1.000	64293....	Male	56....

Figure 34. the sequence of data after sorting

4. Data transformation

4.1 Data Reduction

Before building a model, we need to make choices about the attributes to use and think about which attributes can be used to create a model with high predictive accuracy. Identify attributes that are highly relevant and significantly affect the target attributes and discard other attributes that may affect and mislead the algorithm or model off the right track. Thus, to reduce the impact of irrelevant attributes on the overall model, we use SPSS feature selection to select the important features and minimize the insignificant ones, while reducing processing costs. All attributes except the identifying attribute ID of each record are used as input values for this feature, and the impact of every single attribute on the final target attribute is evaluated in this method. Because these variables are mentioned as risk factors for stroke in medical clinical reports and studies, we needed to select features using the SPSS feature selection tool and incorporate the study focus. When using the SPSS feature selection tool, we set stroke as the target, indicating whether the patient had previously had a stroke.

Figure 35 shows the results of the feature selection function execution and it can be seen that four attributes were automatically selected as significant and the remaining eight attributes were marked as insignificant for various reasons.

Analyzing these selected unimportant attributes, we need to perform a further manual selection. Heart disease was discarded due to many single categories (a few people suffer from this disease, while most do not experience it) and gender is about to be discarded due to its low sensitivity to the final result. In contrast, smoking status, work category, and weight type (BMI groups) needed to be additionally selected, as a large number of studies have shown that these attributes are highly correlated with stroke. Therefore, we ended up with a total of 7 attributes selected.

This function of performing data selection reduces the computational cost of constructing further modules to unnecessarily process the data and reduces the chance of misleading the algorithm.

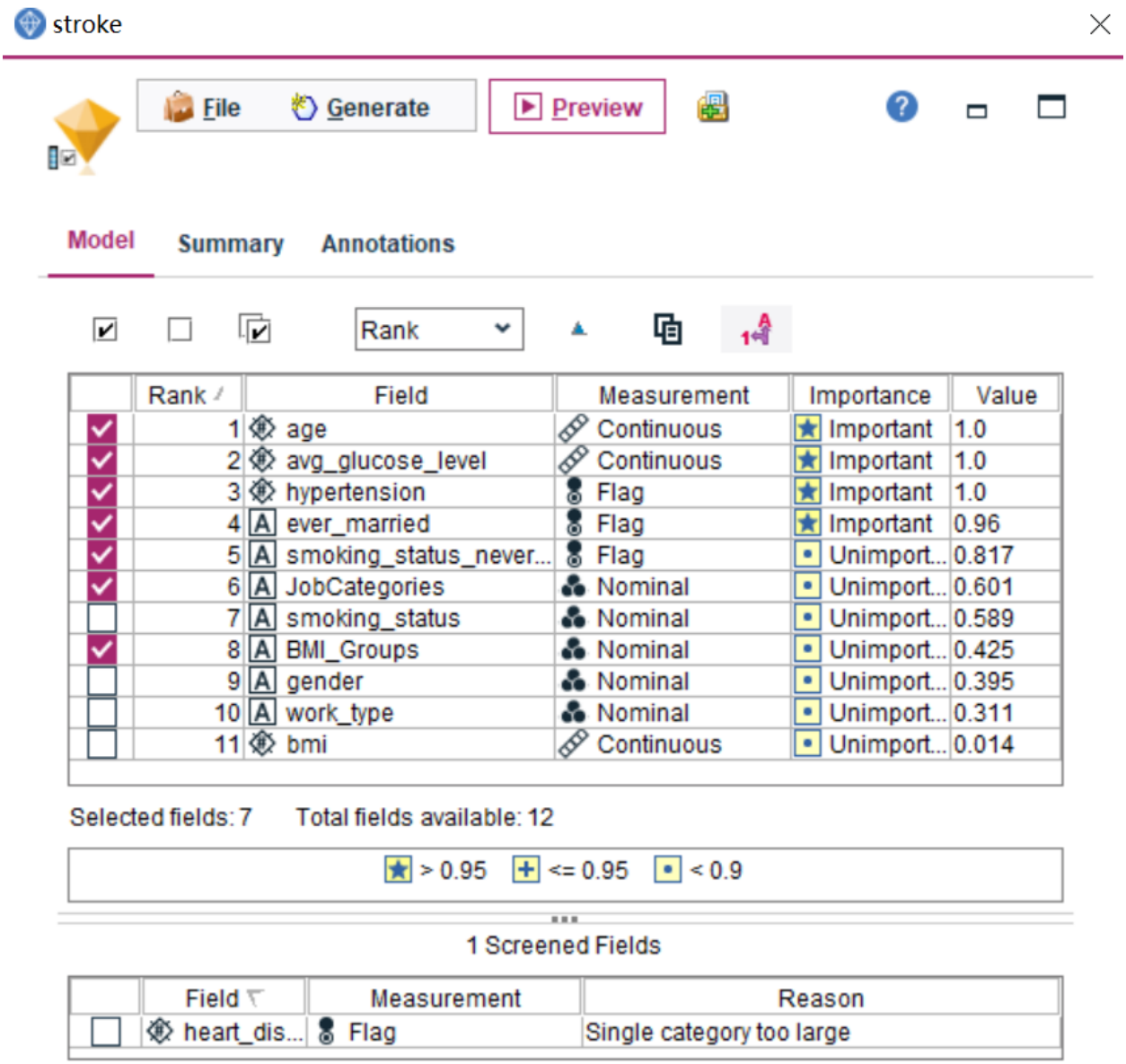


Figure 35. Execution of feature selection

4.2 Data Projection

By observing the distribution of the values of the target attributes as shown in Figure 26, we found that the 2 outcomes 0 indicates individuals who have never had a stroke, while 1 indicates patients who have had a stroke, and the distribution of the values of stroke and no stroke differ greatly, with only 66 patients who had a stroke accounting for 98.03% of the total records, while 3281 patients who did not have a stroke accounted for 1.97% of the total records. And the extreme imbalance between these two results can affect the algorithm in building the model with reduced prediction accuracy and also become insensitive or overfitting in predicting unknown records. So before building the module, we need to perform data projection to create records with similar values.

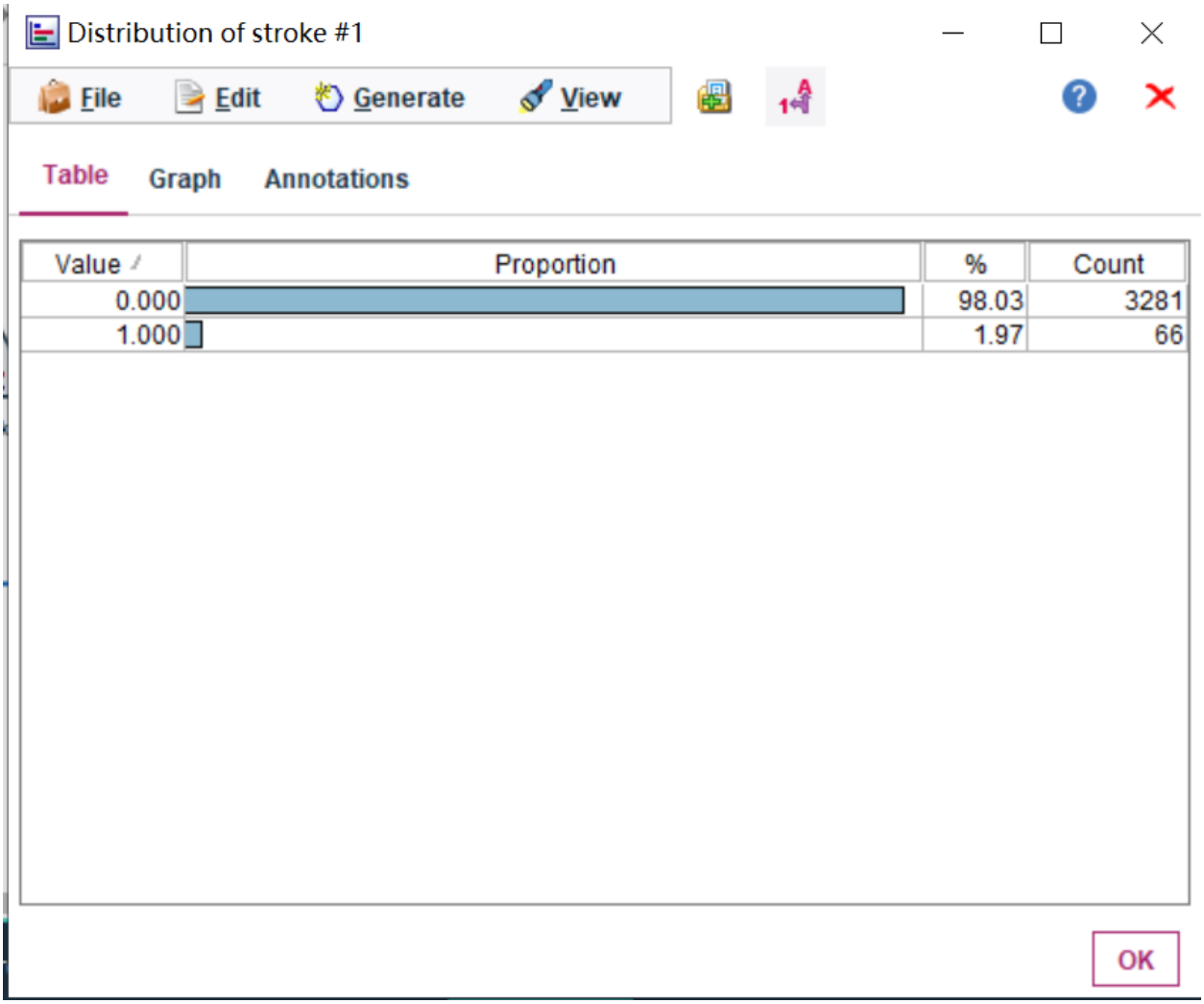


Figure 36. Initial distribution of target attribute stroke

We are able to use balance nodes to balance the data distribution, and when the factors used in balancing cause observations from less frequent categories to be repeated, it is called boosting. The balancing node works by copying or discarding records in the dataset according to the balancing instructions specified in the node. A factor greater than 1 causes duplication of records in the dataset, while a factor less than 1 causes records to be discarded. We use factor 48 for stroke value equals 1.

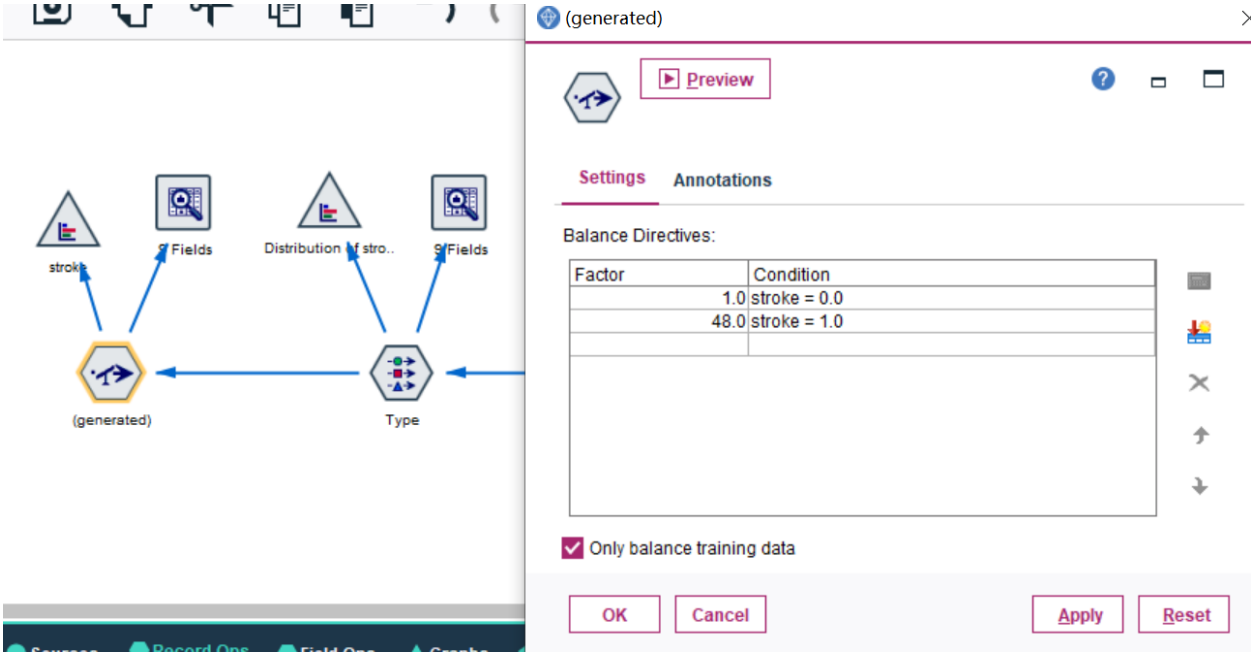


Figure 37. Use the balance node to balance data in the goal attribute

Figure 38 below shows the results of the balanced data projection. We can see that it adds 3,102 cases to the existing dataset. With the data projection implementation, the potential problems of insensitivity to unknown data and model overfitting mentioned earlier will be solved at the expense of accuracy.

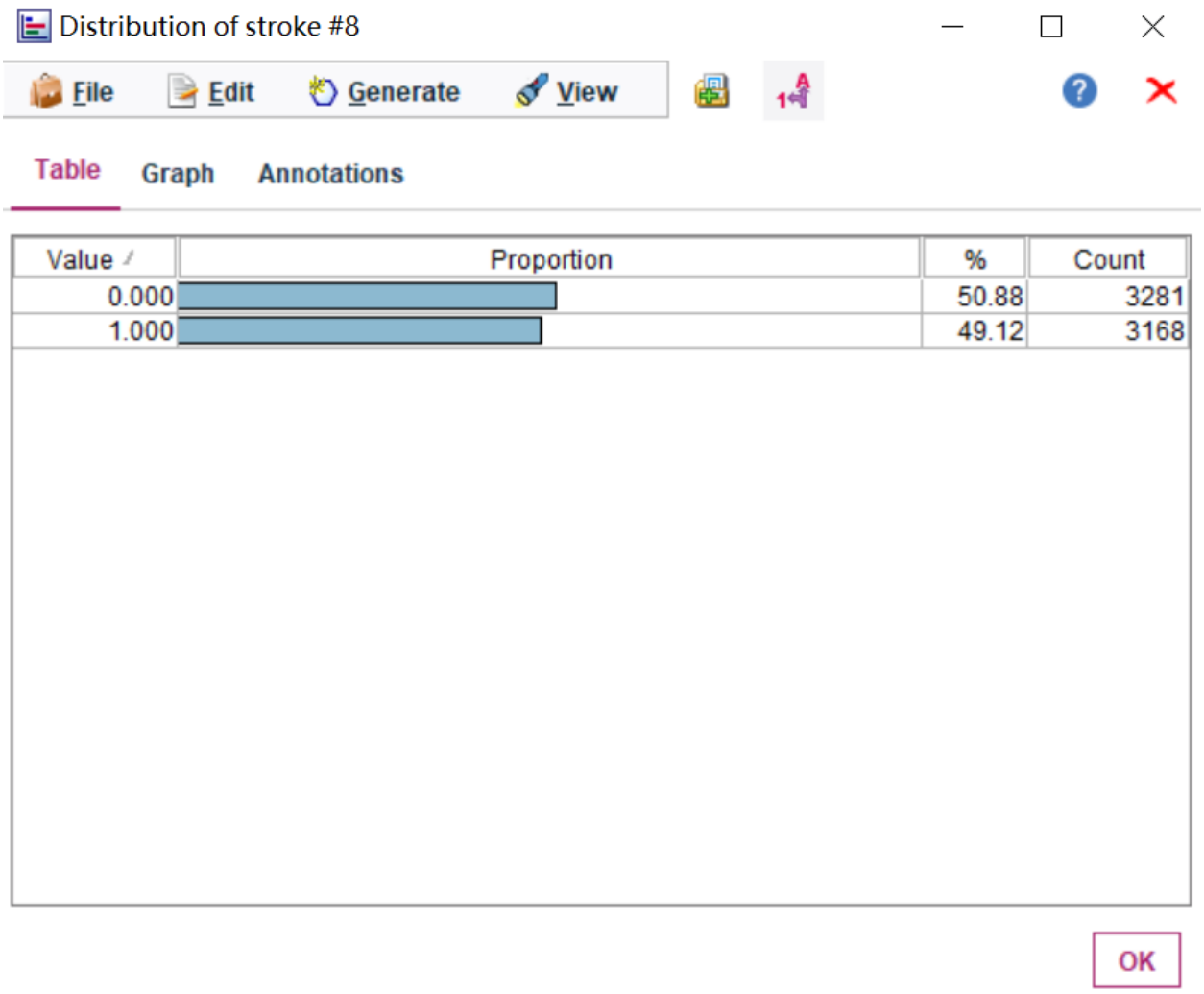


Figure 38. distribution of target attribute stroke after the balance

5. Data-Mining Method(s) Selection

5.1. Discussion of Data Mining Methods in Context of Data Mining Objectives

As mentioned earlier in the business objectives, we want to predict, prevent and reduce the occurrence of stroke disease based on the current status of the patient. Combined with the initial data mining objectives, we want to clarify the relationship between the patient's current health index (current disease and BMI), age, and current life status (smoking status, marital status, type of work, living environment, etc.) and stroke. And by analyzing these relationships, the data is used to train and fit models that are applicable to the data predictions, and by providing these predictions to analyze the causes of stroke in current patients and to reduce the chances of stroke in potential patients at an early stage.

Before building the module, it is essential to review our initial intention of doing this research and choose to narrow or keep this goal for the module building. The predictive system is based on a model that follows a number of patterns, which are the relationships between the target and all relevant information relationships.

Modeling assumptions:

In the preparation phase, a total of 7 attributes were selected: age, average blood glucose level, hypertension, whether married, whether smoking, type of work, and weight type (BMI groups), excluding the record identification number and target attributes. We assume that these attributes are highly correlated with the target attributes, while the deleted attributes are not regrouped with smoking status, job type, and BMI, gender, heart disease, etc. for various reasons that would affect the accuracy of the model's prediction of the target attributes.

Make predictions:

The main feature of this prediction system is to find the cause of stroke or to predict the level of stroke risk for an individual. In other words, it needs to collect information from individuals and use existing training models to evaluate the level of risk based on these collected data for consumption. The seven attributes that are to be predicted to be collected as candidates or potential influences on stroke. The total number of instances prepared for the data mining process was 6,449.

And before we build a model for the prediction system, we need to create the test design for the validation of the prediction model. When creating a test design, the entire data will be used to divide the training and test sets into a 7:3 ratio.

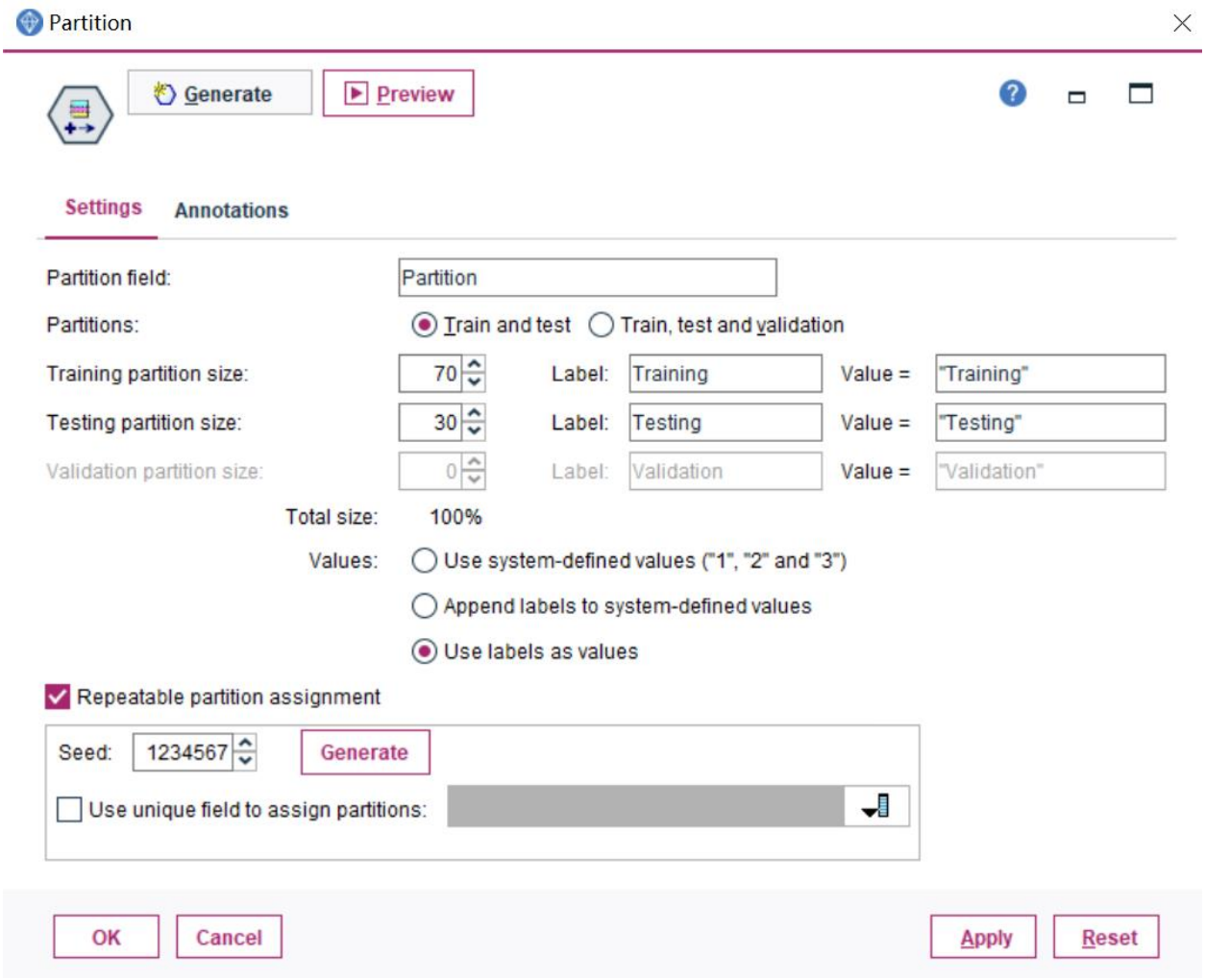

Partition
Generate
Preview
Settings
Annotations
Partition field:
Partition
Partitions:
Train and test
Train, test and validation
Training partition size:
70
Label:
Training
Value =
"Training"
Testing partition size:
30
Label:
Testing
Value =
"Testing"
Validation partition size:
0
Label:
Validation
Value =
"Validation"
Total size:
100%
Values:
Use system-defined values ("1", "2" and "3")
Append labels to system-defined values
Use labels as values
Repeatable partition assignment
Seed:
1234567
Generate
Use unique field to assign partitions:
OK
Cancel
Apply
Reset

Figure 39. data split to training and test set

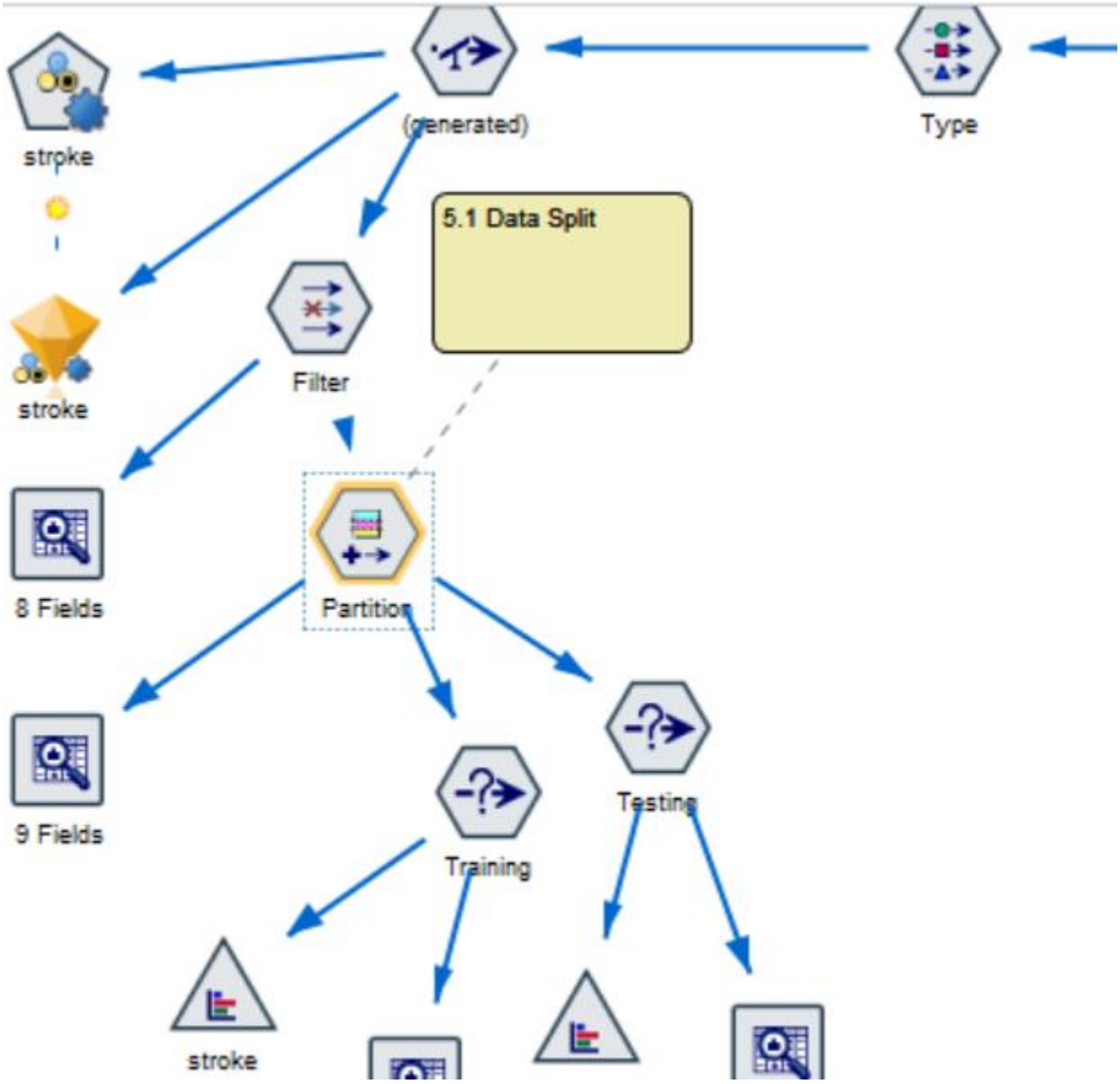


Figure 40. After data split

Choosing the Right Modeling Techniques:

1. Association
   It is used to find correlations between two or more items by identifying hidden patterns in a dataset, hence the term relationship analysis. This method is used in shopping cart analysis to predict customer behavior. For example, the classic case, purchase (x, "beer") -> purchase (x, "diaper") [support = 1%, confidence = 50%], where x represents a customer who buys beer and diapers together. Confidence

indicates that if a customer buys beer, there is a 50% chance that he/she will also buy diapers as well. Support implies that 1% of all transactions analyzed indicate that beer and diapers are purchased together.
There are two types of association rules.
Unidimensional association rules: These rules contain repeated individual attributes.
Multidimensional association rules: These rules contain multiple attributes that are repeated.

2. Clustering
   A cluster is a collection of data objects; these objects are similar in the same cluster. This means that objects in the same group are similar to each other, and they are quite different, or they are different or unrelated to objects in other groups or other clusters. Cluster analysis is the process of discovering groups and clusters in the data, making the degree of association between two objects the highest if they belong to the same group and the lowest between them otherwise. The results of this analysis can be used to create customer analysis.

3. Classification
   This data mining method is used to distinguish items in a dataset into classes or groups. It helps to accurately predict the behavior of entities within the group. It is a two-step process.
   Learning step (training phase): Here, the classification algorithm constructs the classifier by analyzing the training set.
   Classification step: Test data is used to estimate the accuracy or precision of the classification rules.
   Conclusions are drawn based on the relationship between each field value and the final target. For example, a medical researcher analyzes cancer data to predict which drug to prescribe to a patient, groups patients according to their lives, and finds relationships or patterns between attributes.

5.2. Selecting the appropriate Data-Mining method(s)

According to the objectives of data mining mentioned in section 1.4, the three data mining methods mentioned in section 5.1, clustering, association and classification methods. Considering the suitability of this project, the classification method was selected as the data mining method.

Both classification methods are more suitable for this project than the other two methods, clustering and association. In terms of prediction of the target attributes, the association

method analyses the relationship between two or more attributes and draws conclusions about association rules to predict the probability of one attribute occurring when another occurs. When there are multiple attributes, association methods are less sensitive to the data than classification methods. Classification methods, on the other hand, combine all the important attributes to make predictions about the future target data.

In terms of data grouping, association methods only absorb continuous values and make predictions. Classification methods also provide the ability to use non-continuous values as predictors. And a larger range of predictors can produce more accurate answers to this. However, cluster analysis, the process of discovering clusters and clustering in the data, is not compatible with our project objectives and is therefore not chosen.

In conclusion, the method of classification is the most suitable one for this case.

6. Data-Mining Algorithm(s) Selection

6.1 Algorithms analysis

According to the algorithm of Classification discussed in section 5.2, includes supervised learning and unsupervised learning algorithms. The difference between these two types is the target label. The information from supervised learning will be given a result label and it is mainly used in terms of pattern search. Unsupervised learning will not be given any outcome labels and it focuses on the relationships in the dataset. It is mainly used in clustering methods (Mohamed, Yap, & Berry, 2019). In this data mining process, algorithms categorized as supervised learning will be used depending on the data mining objectives. Based on the limitations mentioned above, three data mining algorithms are being selected and discussed based on the data mining objectives.

The algorithms C5.0 1, Tree-AS 1, and CHAID are selected and discussed.

1. C5.0.
   C5.0 algorithm is one of the most well-known decision tree implementations. Because it performs well for the majority of issue types right out of the box, the C5.0 algorithm has taken the lead in setting the standard for creating decision trees in the industry. The decision trees used in the C5.0. algorithm typically perform nearly as well as more complex and sophisticated machine learning methods (such as Support Vector Machines and Neural Networks), but they are considerably simpler to use and comprehend. By reducing the estimated entropy value, this approach employs an information entropy computation to find the best rule that divides the data at that node into purer classes.

As a result, each subset of the data split by the rule will initially include only one class and ultimately contain less diversity of classes as each node divides the data depending on the rule at that node. Because it is easy to compute, C5.0. completes the operation rapidly. C5.0 is resilient. It can be used with category or numerical data. Additionally, it can accept missing data values. A decision tree or a set of rules can be produced by the R implementation. New unclassified data can be given a class using the output model.

2. Tree-AS
   Data in a distributed context can be utilized with the Tree-AS node. You have the option to use the CHAID or Exhaustive CHAID model when creating decision trees with this node. Exhaustive CHAID is a CHAID variant that examines all potential splits for each predictor more thoroughly but takes longer to compute.

3. CHAID
   CHAID decision tree stands for a Chi-square automatic interaction detection decision tree technique. This algorithm based on the adjusted significance testing. For each category predictor, CHAID produces every conceivable cross tabulation up until the best result is reached and no further splitting is feasible. We can visually observe the connections between the split variables and the corresponding component inside the tree using the CHAID approach. The process of creating a decision tree begins with determining the dependent or target variable, which is the tree's root. CHAID analysis divides the target into two or more categories known as the initial, or parent, nodes. The nodes are subsequently divided into child nodes using statistical techniques. The CHAID method has the advantage of not requiring that the data be regularly distributed.

All of those mentioned algorithms will be run in SPSS and evaluated in the following chapters.

6.2 SPSS Algorithms analysis

1. Data-Mining Objective: Decision Tree C5.0 Algorithm

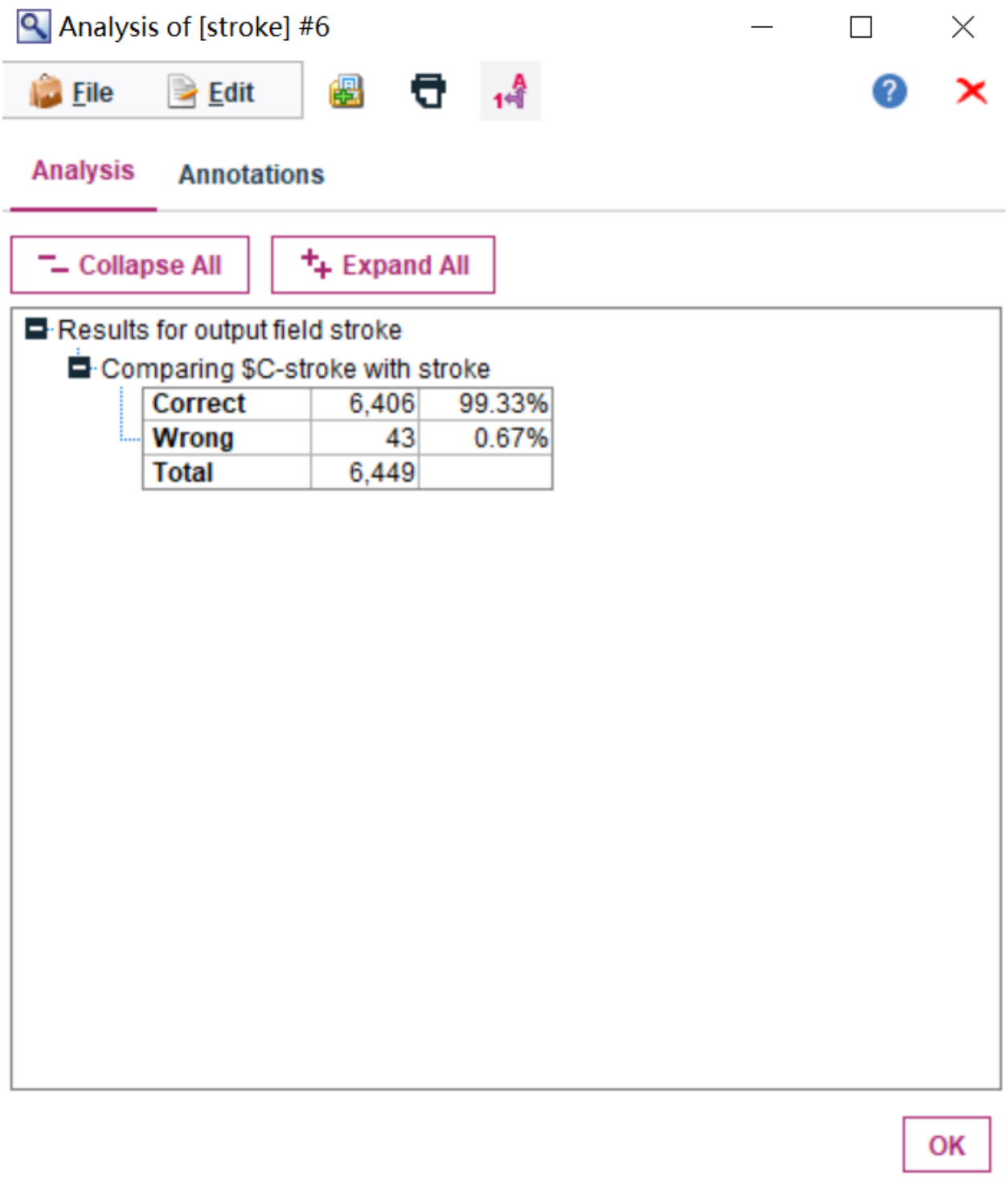
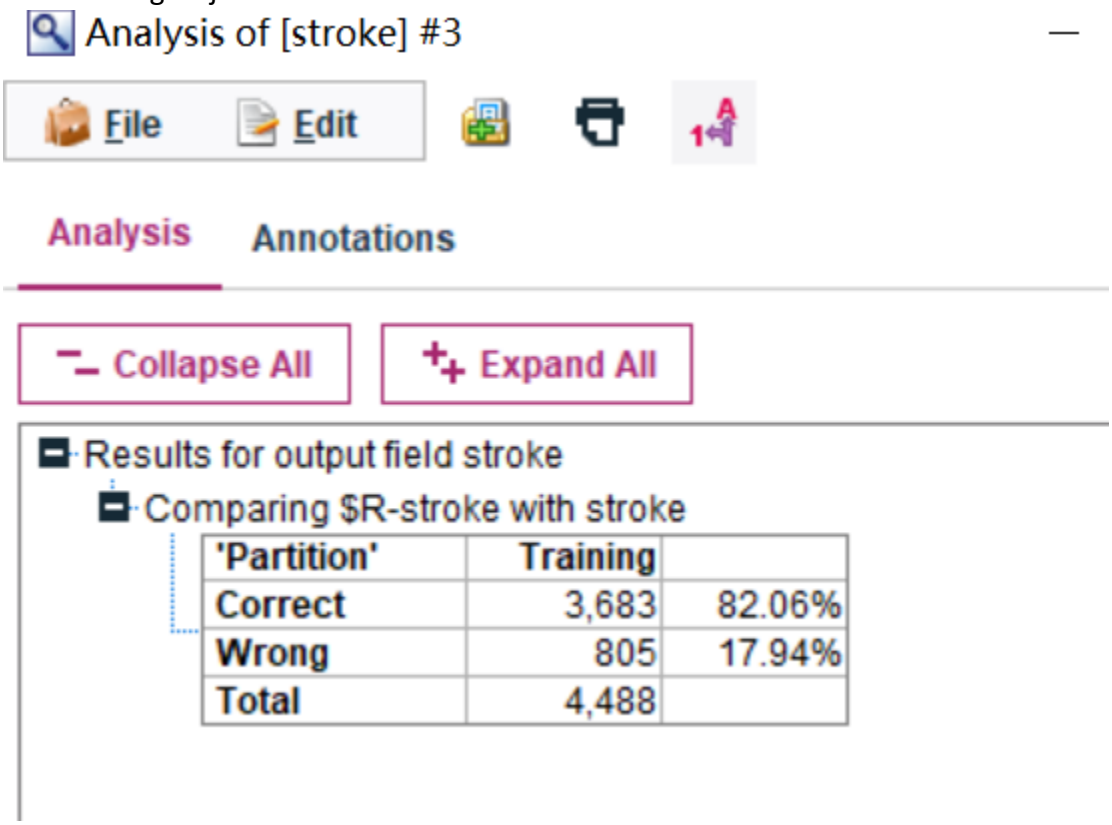


Figure 41. The accuracy of stroke prediction using C5.0.

In the implementation of SPSS algorithm, this node constructs a rule set or a decision tree using the C5.0 algorithm. A C5.0 model divides the sample according to the field that yields the most information gain. The process is repeated until the subsamples cannot be further divided, generally based on a different

field, for each sub-sample determined by the initial split. The lowest-level splits are then reviewed again, and any that do not significantly add to the model's value are either eliminated or pruned. C5.0 can create two different types of models. The divides that the algorithm discovered are simply described by a decision tree. Each instance in the training data corresponds to exactly one terminal (or "leaf") node in the tree, and each terminal node defines a specific subset of the training data. In other words, a decision tree can only make exactly one prediction for each given data record.

2. Data-Mining Objective: Tree-AS



'Partition'	Training	
Correct	3,683	82.06%
Wrong	805	17.94%
Total	4,488	

Figure 42. The accuracy of stroke prediction using tree-AS algorithm

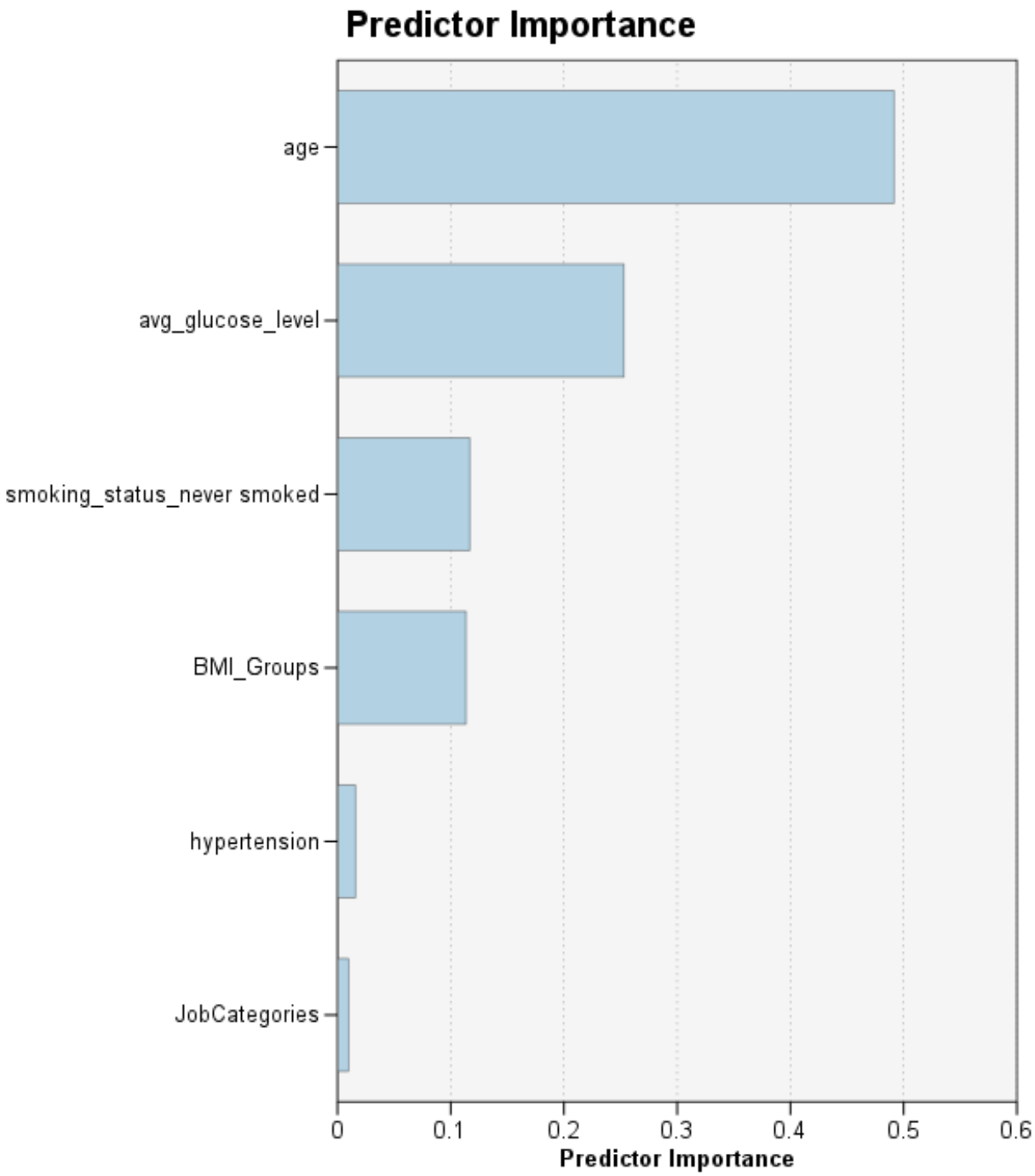
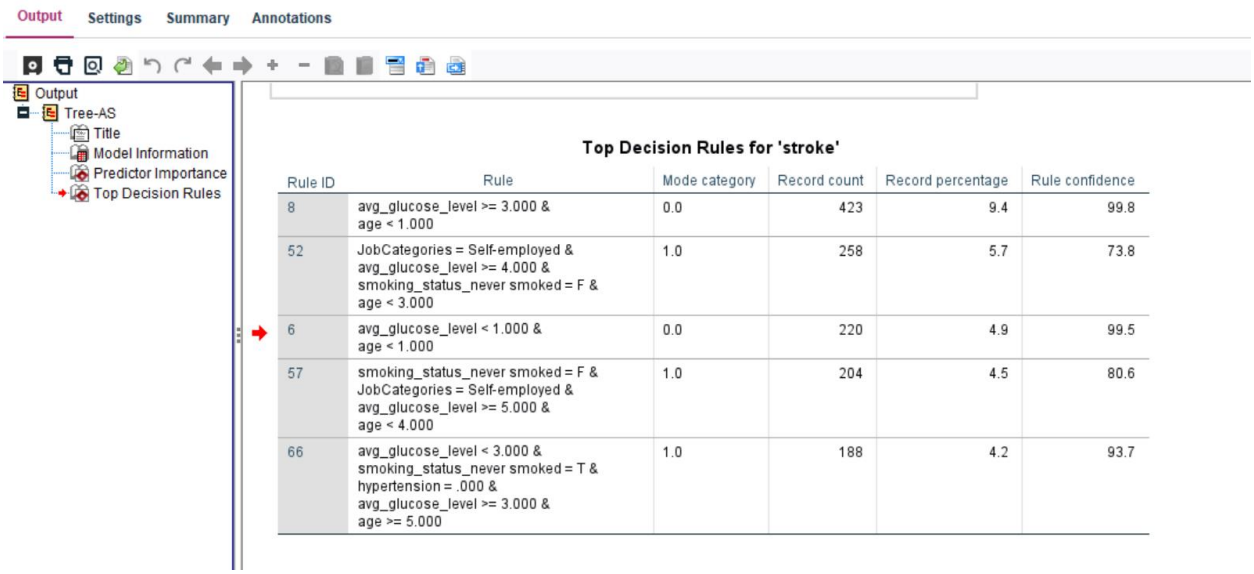


Figure 43. Results of Tree-As algorithm feature importance analysis



Top Decision Rules for 'stroke'

Rule ID	Rule	Mode category	Record count	Record percentage	Rule confidence
8	avg_glucose_level >= 3.000 & age < 1.000	0.0	423	9.4	99.8
52	JobCategories = Self-employed & avg_glucose_level >= 4.000 & smoking_status_never smoked = F & age < 3.000	1.0	258	5.7	73.8
6	avg_glucose_level < 1.000 & age < 1.000	0.0	220	4.9	99.5
57	smoking_status_never smoked = F & JobCategories = Self-employed & avg_glucose_level >= 5.000 & age < 4.000	1.0	204	4.5	80.6
66	avg_glucose_level < 3.000 & smoking_status_never smoked = T & hypertension = .000 & avg_glucose_level >= 3.000 & age >= 5.000	1.0	188	4.2	93.7

Figure 44. Tree-AS top decision rules

For the Tree-AS modeling implementation, the algorithm automatically chooses to build a decision tree with CHAID or an exhaustive CHAID model. Exhaustive CHAID is a CHAID variant that examines all potential splits for each predictor more thoroughly but takes longer to compute. Nodes can be divided into two or more subgroups at each level; target and input fields can be continuous or categorical. In the model, all ordinal fields must support numeric storage (not string). Use the Reclassify node to transform them if necessary.

3. Data-Mining Objective: CHAID

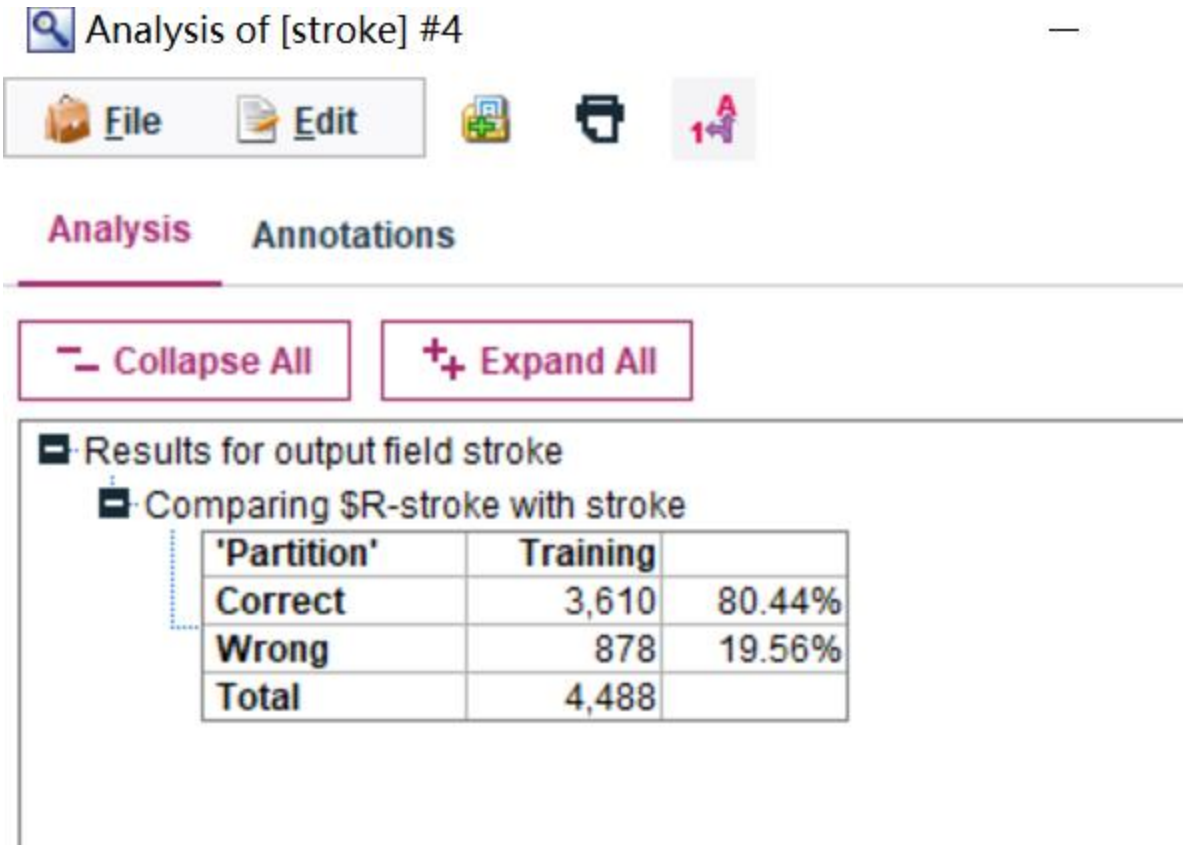


Figure 45. The accuracy of stroke prediction using CHAID

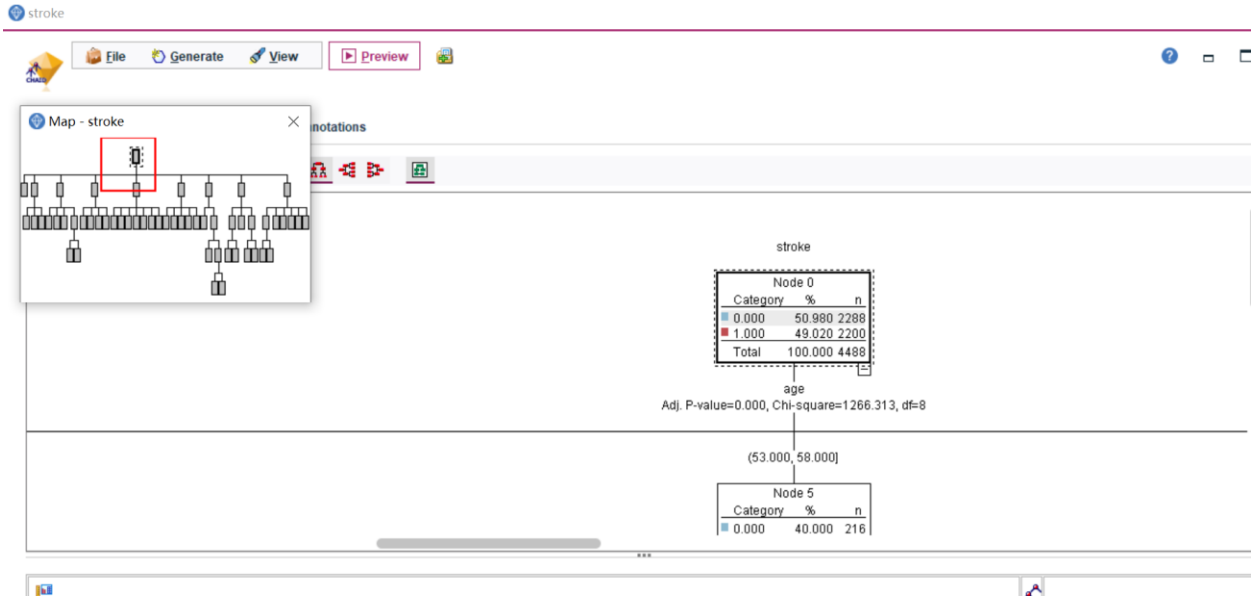


Figure 46. CHAID tree map

In the implementation of CHAID mode of decision tree, CHAID first looks at the crosstabulations between each of the input fields and the result, then does a chi-square independence test to see whether any differences are significant. CHAID will choose the most significant input field if more than one of these relations is statistically significant (smallest p value). When there are more than two categories in an input, the categories are compared, and those that provide the same results are compacted as a single group. This is accomplished by sequentially connecting the categories with the smallest difference. When all remaining categories diverge at the designated testing level, the category-merging procedure comes to an end. Any categories may be combined for nominal input fields, but only contiguous categories may be merged for an ordinal collection.

Figure 41, 42, 45 is showing the accuracy of stroke predictions using three algorithms (C5.0., Tree-AS and CHAID).

From the accuracy perspective, the C5.0. algorithm has the highest accuracy, which reaches 99.33%, then it followed by the Tree-AS (82.06%) and CHAID (80.44%). Based on the accuracy result, the C5.0 algorithm is worth being picked as the final prediction models. However, with this high accuracy of prediction, this also needs to be considered if the model is over-fitting, as the accuracy is the same for each run. The dataset needs to be further randomly divided into training and test sets to ensure that the model is not over-fitted. According to the computation time, C5.0 algorithm take less time than Tree-AS and CHAID algorithms.

Based on the above experiment, the model C5.0. algorithm is being picked as a prediction model.

6.3 SPSS Analysis discussion

We will choose Decision Tree C5.0 Algorithm based on our data mining objective, as it has the highest accuracy (99.33%) compare with Tree-As algorithm (82.06%) and CHAID (80.44%).

From these 3 models, we can see that age is the most influential attribute towards the target attribute as these 3 algorithms are shown above. From the feature importance generated by the decision tree, the top three factors that matter the decision are age, glucose level and smoking status. While the job type of patients has the least impact on the decision. The order of importance showed an agreement of the clinical findings.

Furthermore, reasons we choose C5.0. model as our model:

According to our data mining objective, using C5.0. algorithms can easily achieve these goals. In the analysis result of C5.0. Algorithm, we can see the most influential attributes, also the relationship between these attributes and stroke. As shown in Figure 47, the preceding

branching conditions (age and average glucose level) are the more important factors influencing stroke.

In C5.0. algorithm, each instance in the training data corresponds to exactly one terminal (or "leaf") node in the tree, and each terminal node defines a specific subset of the training data. In other words, a decision tree can only make exactly one prediction for each given data record. However, it can still give high accuracy predictions, which means that C5.0. model is well placed to combine various influencing factors to make more accurate stroke predictions.

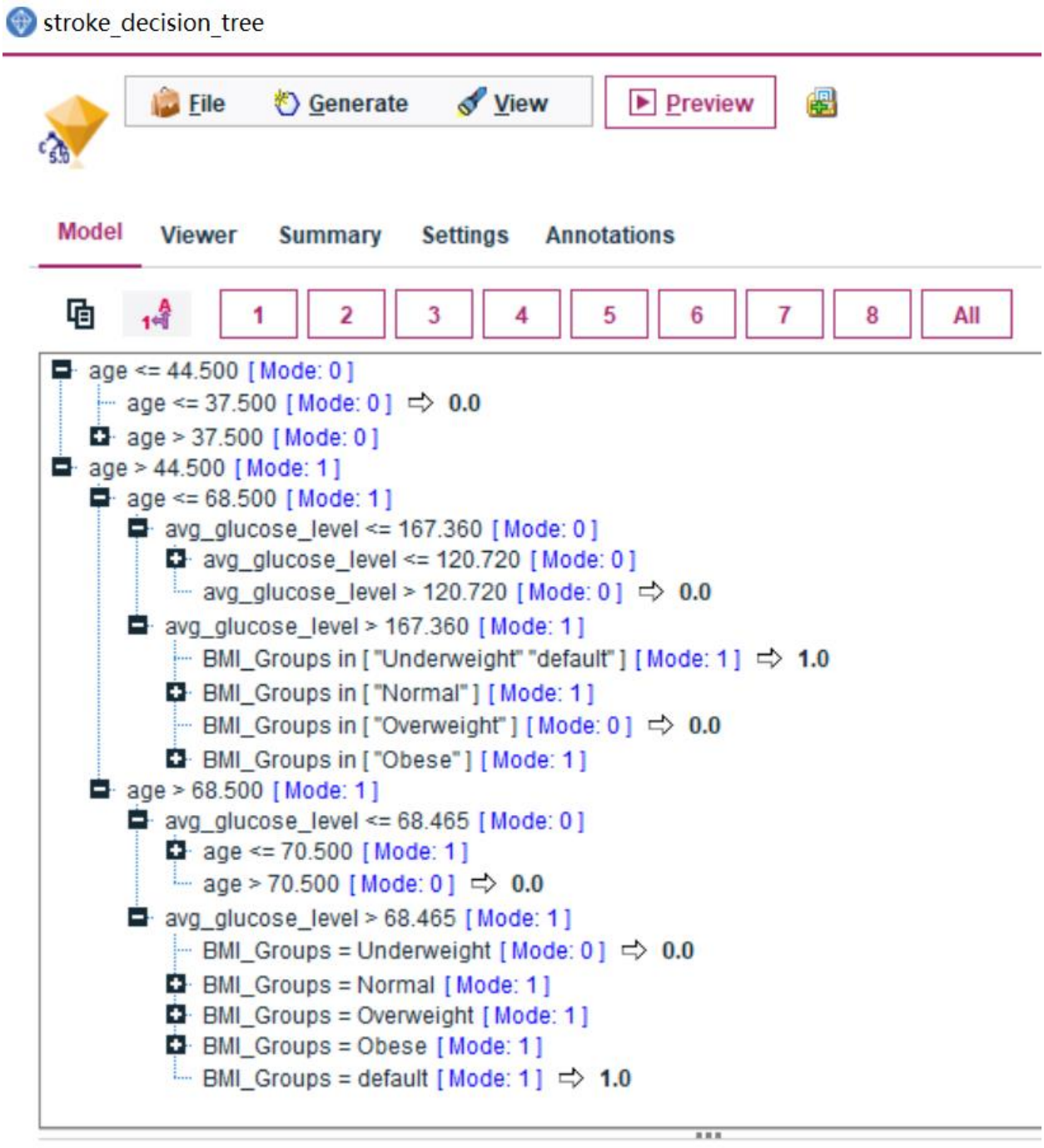


Figure 47. C5.0. decision tree

7. Data Mining

7.1 Logical test designs

Before the datamining was conducted, we first spited the dataset into a training and testing dataset as shown in section 5.1. A method for assessing a machine learning algorithm's performance is the train-test split. It may be applied to issues involving classification or regression as well as any supervised learning technique. The process entails splitting the dataset into two subgroups. The training dataset is the initial subset, which is used to fit the model. The model is not trained using the second subset; rather, it is given the input element of the dataset, and its predictions are then produced and contrasted with the expected values. The test dataset is the second dataset in question. For fitting the machine learning model, use the train dataset. Test dataset used to assess how well a machine learning model fits the data.

The dataset is split into two parts, the training and test datasets. This operation simulates the actual environment faced by the model. Furthermore, by using the training/testing set, it is possible to assess the sensitivity of the model in the face of unseen data. The content inside the training dataset is the knowledge possessed before being labelled, while the content inside the test set is all the inputs, then the model needs to predict the answer based on the given information. Finally, the performance of this model can be evaluated based on various methods.

The split data ratio was set at 7:3, with the training data accounting for 70% of the total records and 30% reserved for the test data set. the 7:3 ratio allows the model to be adequately trained as it contains multiple cases where patterns can be found, and the 30% reset is large enough to cover sufficient instances (cases) for potential input and evaluation. Importantly, the data structure of training and test datasets should remain the same, especially for target attribute as shown in Figure 48.

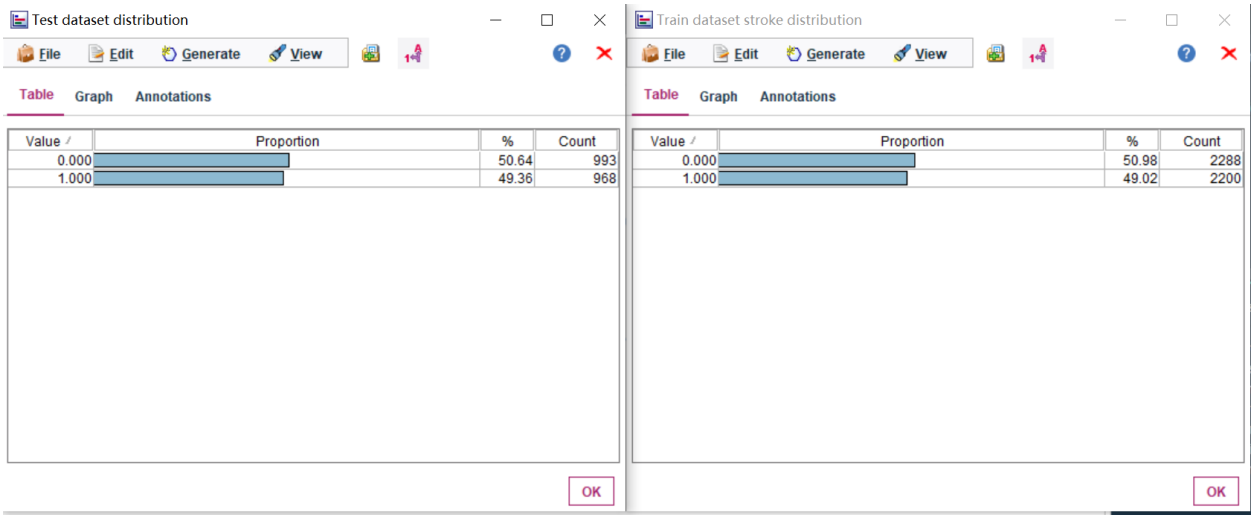
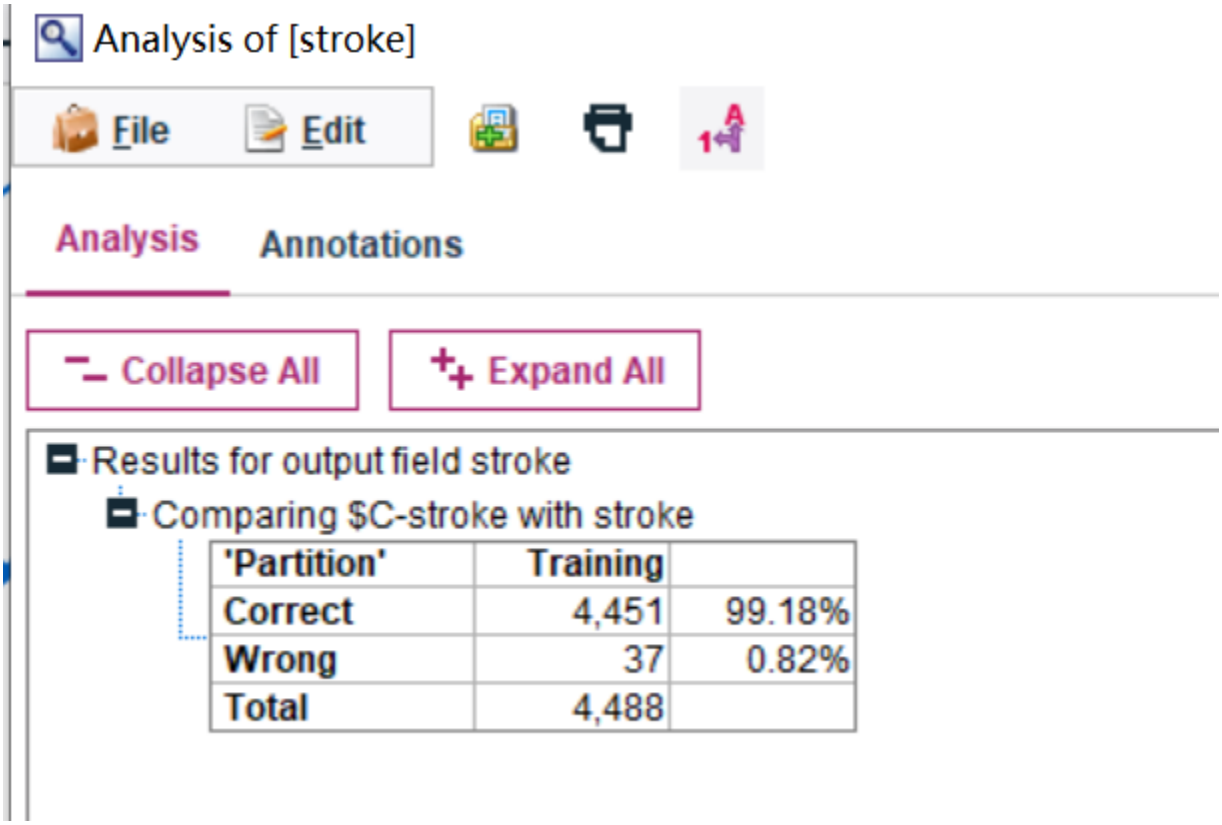


Figure 48. The data structure of stroke for the train/test set

The key to the success of this model is that the model is able to successfully find patterns of relationships and is successful in predicting outcomes with a high degree of accuracy.

7.2 Conduct Data mining

We run the partition that we have split in section 5.1, which split the train set and test set with a 7:3 ratio. Then, we do data mining with selected Decision Tree C5.0 Algorithm.



'Partition'	Training	
Correct	4,451	99.18%
Wrong	37	0.82%
Total	4,488	

Figure 49. Results of training accuracy.

From the result, the training accuracy was 99.18%, which indicated a very good performance of the used C5.0 algorithm for decision tree.

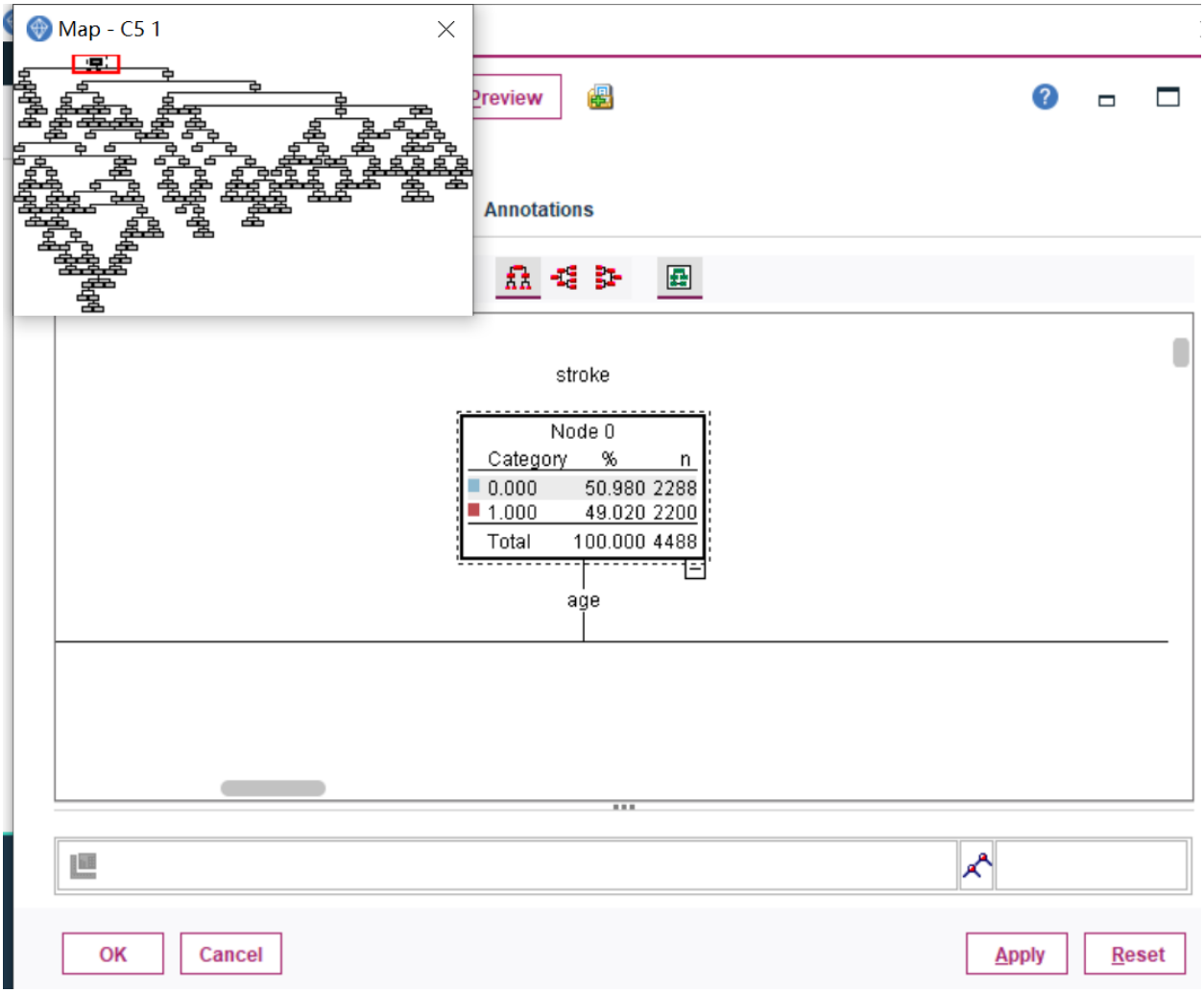


Figure 50. Analysis of the first node in decision tree

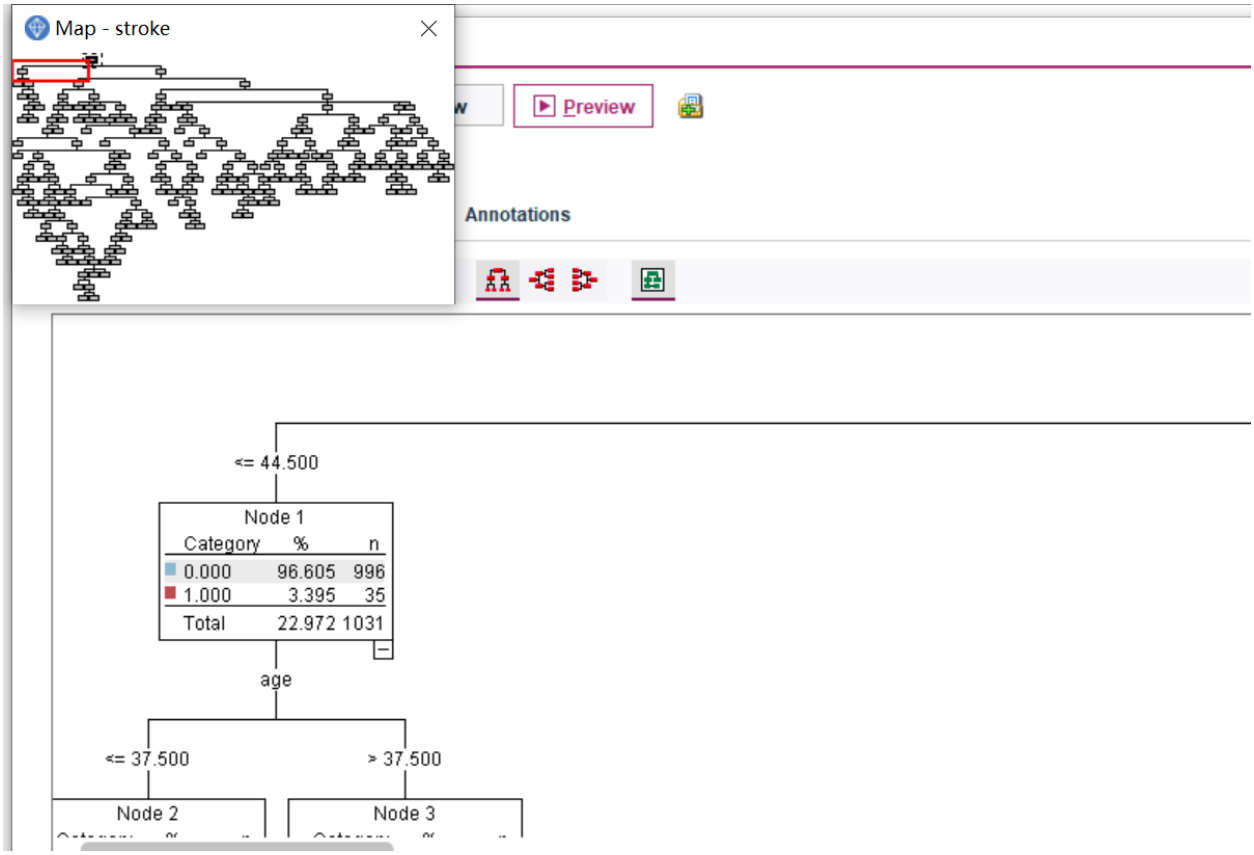


Figure 51. C5.0. model first split (left node)

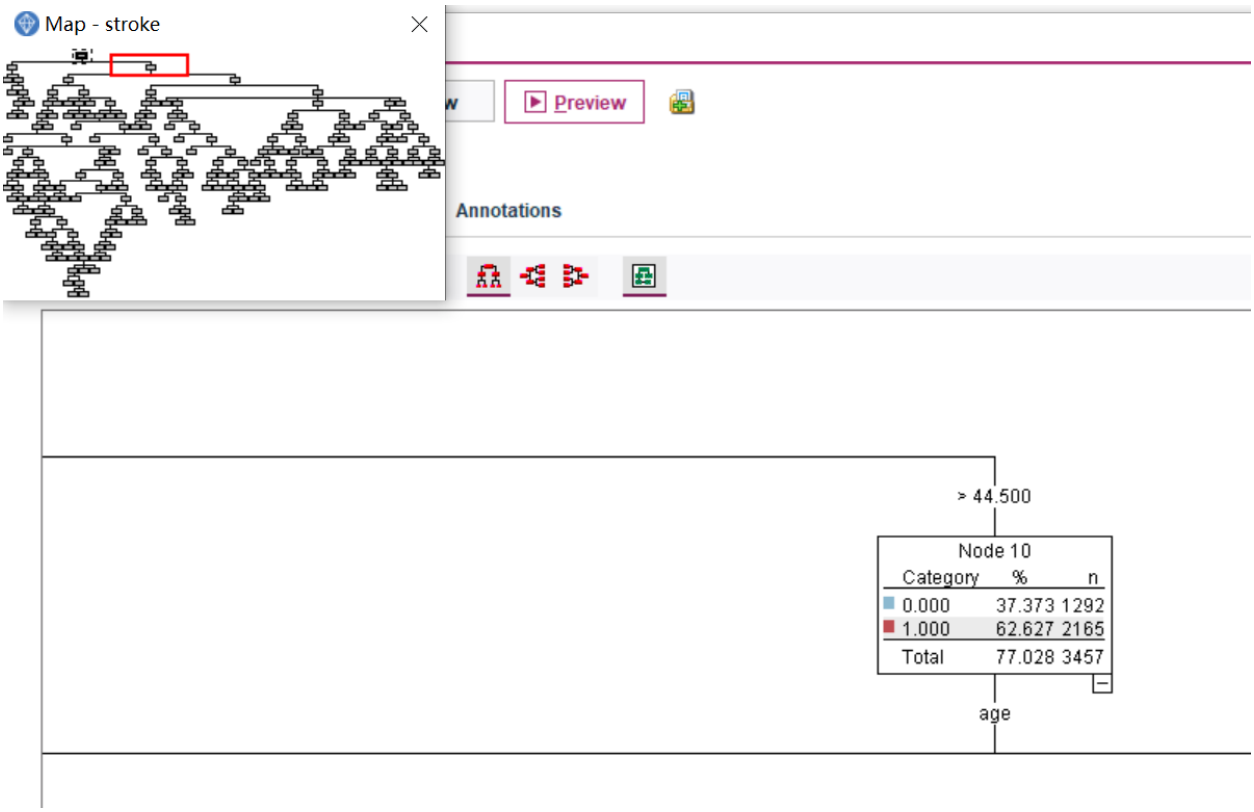


Figure 52. C5.0. model first split (right node)

We can see that the decision tree split age into 2 groups (<=44.5 and >44.5) at the beginning, we analyzed the classification information of the first node ('Node 0') of the decision tree. The result indicates at the very beginner, the decision node barely has the ability to decide the patient with stroke or not. This is quite straightforward, since only through a more hierarchical feature combination, the performance of decision tree can be improved.

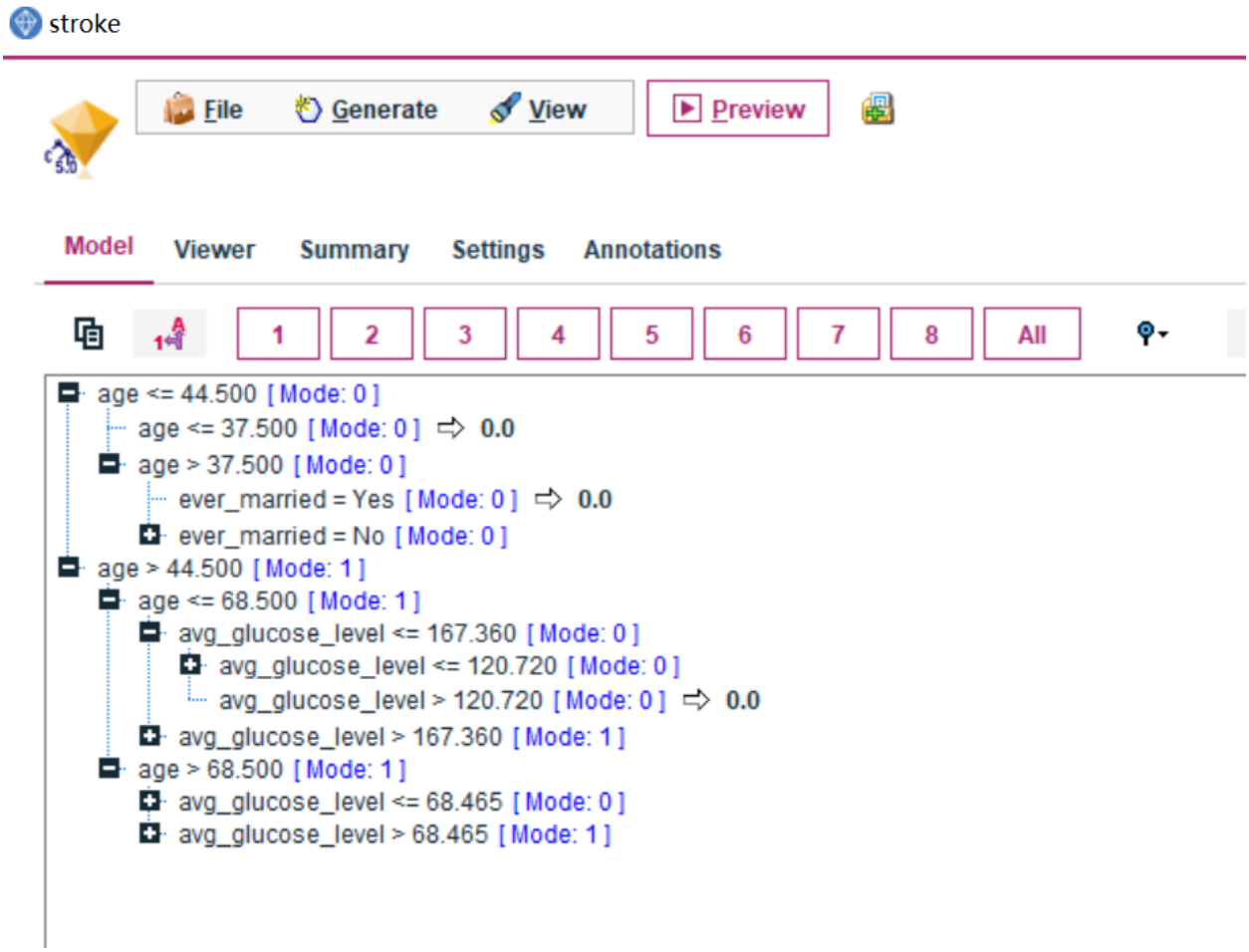


Figure 53. C5.0. model rule set

From Figure 36. C5.0 rule sets view, we can easily see that the most important two predictors of stoke are age and average glucose, as the condition of the rule sets is based on these splits. Therefore, we detailed see how these two important predictors influence the target attribute stroke, and their relationship with the stroke respectively as below.

The relationship between age and stroke:

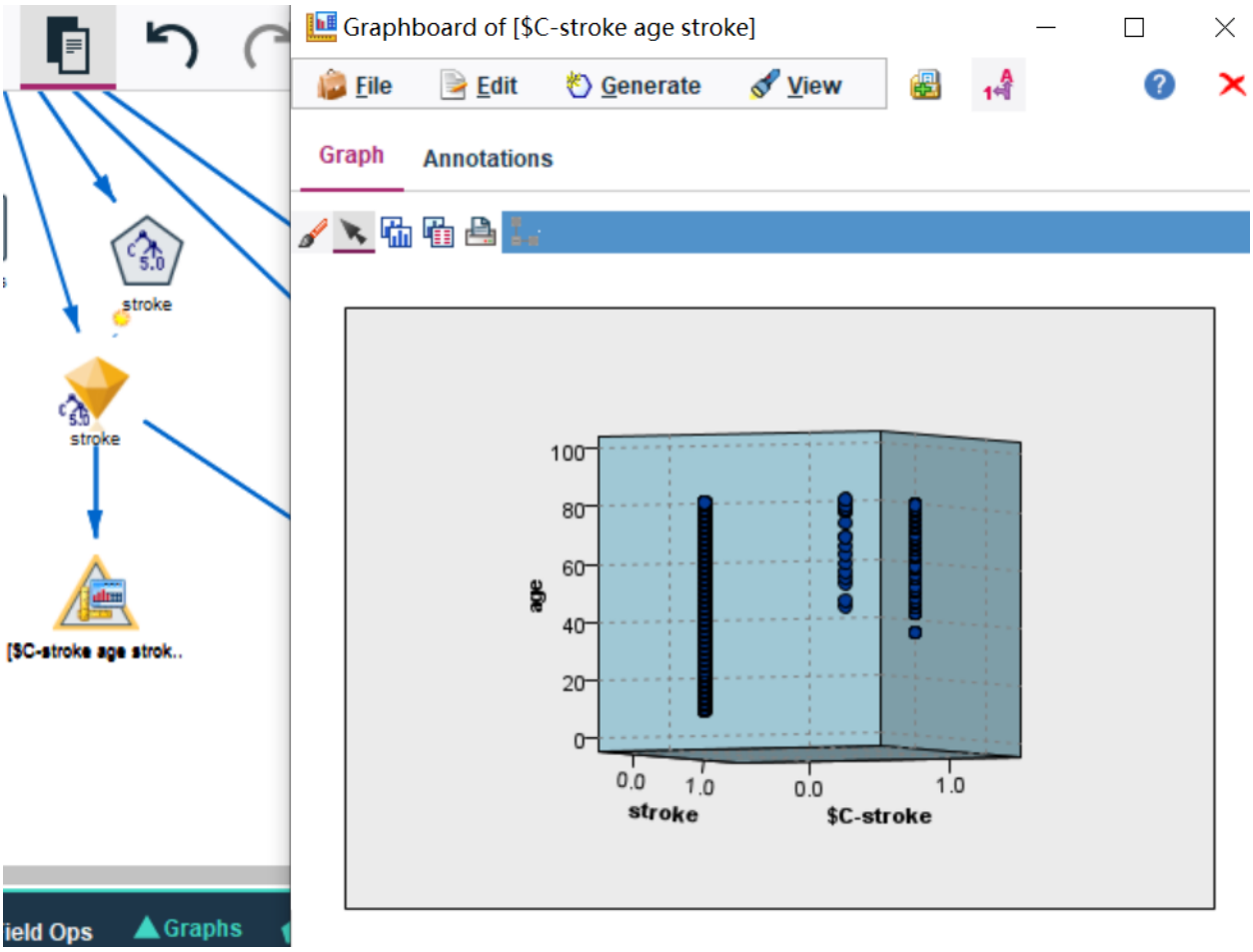


Figure 54. Results of training

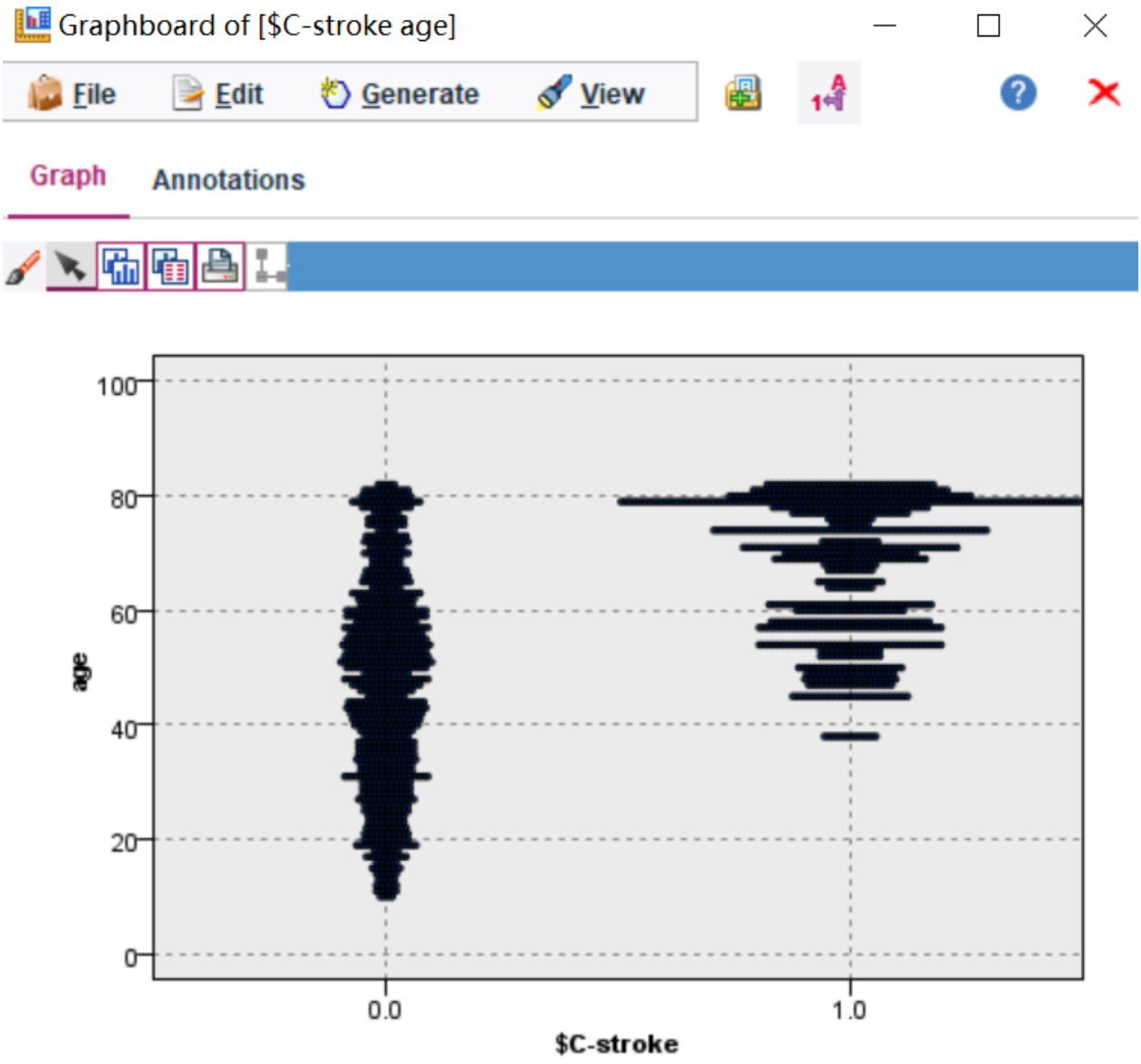


Figure 55. Relation between age and stroke prediction.

From above Figures, we can clearly see the relation between age and stroke prediction, the age of patients above 38 years old has higher chance to get stroke, especially for the patients whose age is about 80 years old.

The relationship between average glucose level with stroke:

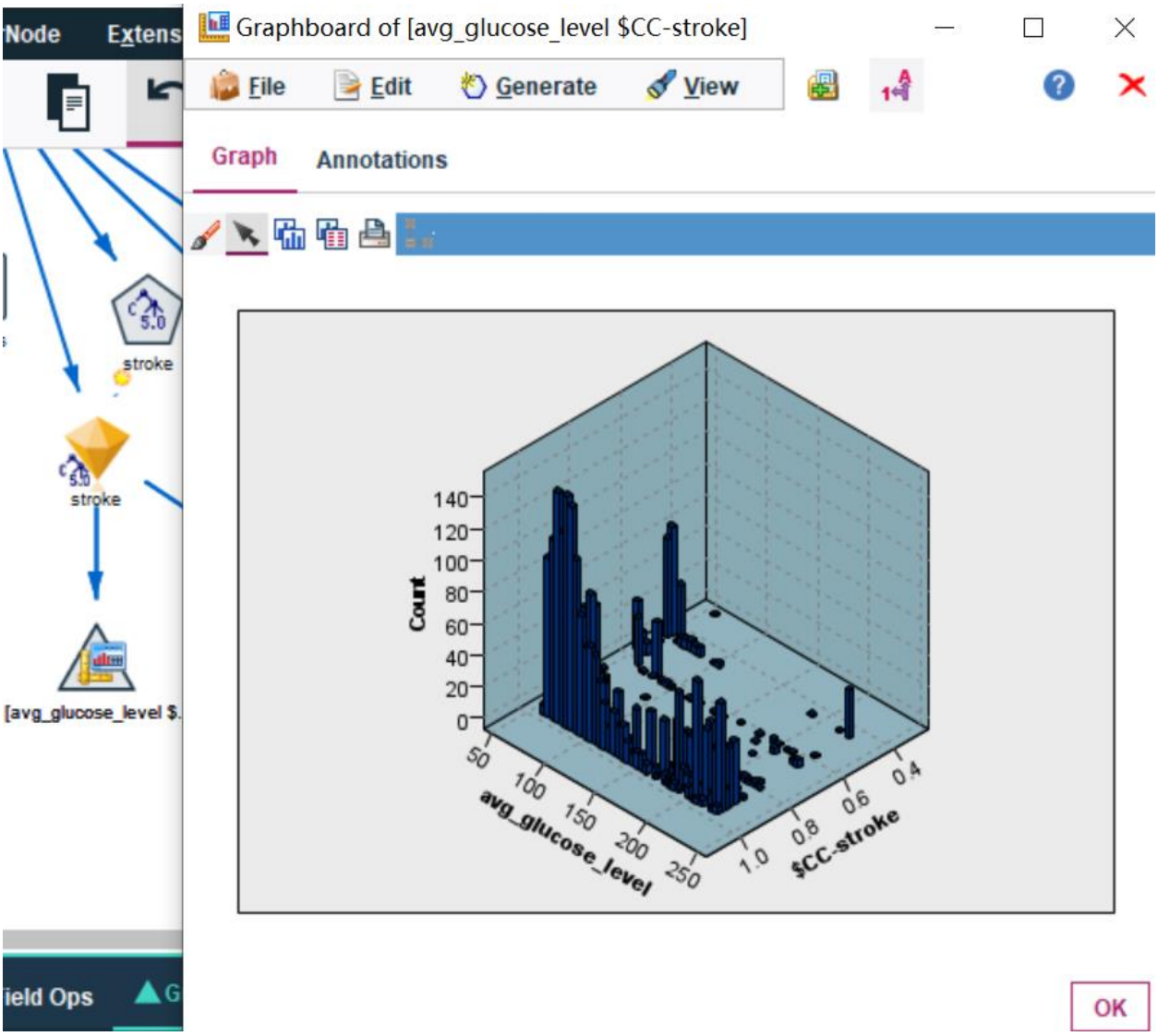


Figure 56. Relation between average glucose level and stroke prediction

The relationship between the data on the graph shows that patients with blood glucose in the range of 50-70 mg/dL and those with blood glucose in the range of 200-250 mg/dL have a higher chance of having a stroke. We know that the normal blood glucose range is 70 to 99* mg/dL (Cleveland Clinic medical professional, 2018), and that blood glucose fluctuates before and after eating. From this we know that people with both low and high blood pressure are more likely to have a stroke. Patients with low blood pressure are more likely to have a stroke than those with high blood pressure.

7.3 The output of models (Search for patterns)

From the result, the algorithm accuracy was 99.18%, which indicated a very good performance of the used C5.0 algorithm for decision tree.

From Figure 50, we analyzed the classification information of the first node ('Node 0') of the decision tree. The result indicates at the very beginner, the decision node barely has the ability to decide the patient with stroke or not. This is quite straightforward, since only through a more hierarchical feature combination, the performance of decision tree can be improved.

The pattern discovery is from the model running result in section 7.2. From Figure 53, we can see that age and average glucose level is the most important predictor of stroke, as these 2 conditions are as the first 2 split conditions when building the tree. This result proved the reliability of feature importance analysis for the decision tree predictor.

Pattern 1:

It is obvious that age is the most important predictor of stroke, as the tree split from start by condition age. The person who has stroke are mainly likely to have happened in their elder age (>44.5 years old). As shown in Figure 51, Figure 52. We can see that the patients under or equal to 44.5 years old, only 3.395% patients got stroke. However, the patients above 44.5 years old, the number of patients who had a stroke climbed to 62.627%.

From the Figure 55 result, the result potentially indicates that the probability of stroke happening is related with the age group. Patients with an age from 44.5 - 80 are more likely to be predicted as stroke happened.

Pattern 2:

Figure 53 can only provide limited information about the importance of average glucose level predictor. Further research is conducted using the plot below be found in Figure 57.

Figure 57 is showing the relationship between the stroke and the average glucose level. The person who has had the stroke is more likely to have a more significant glucose level.

As a result, a higher glucose level will indirectly affect stroke conditions.

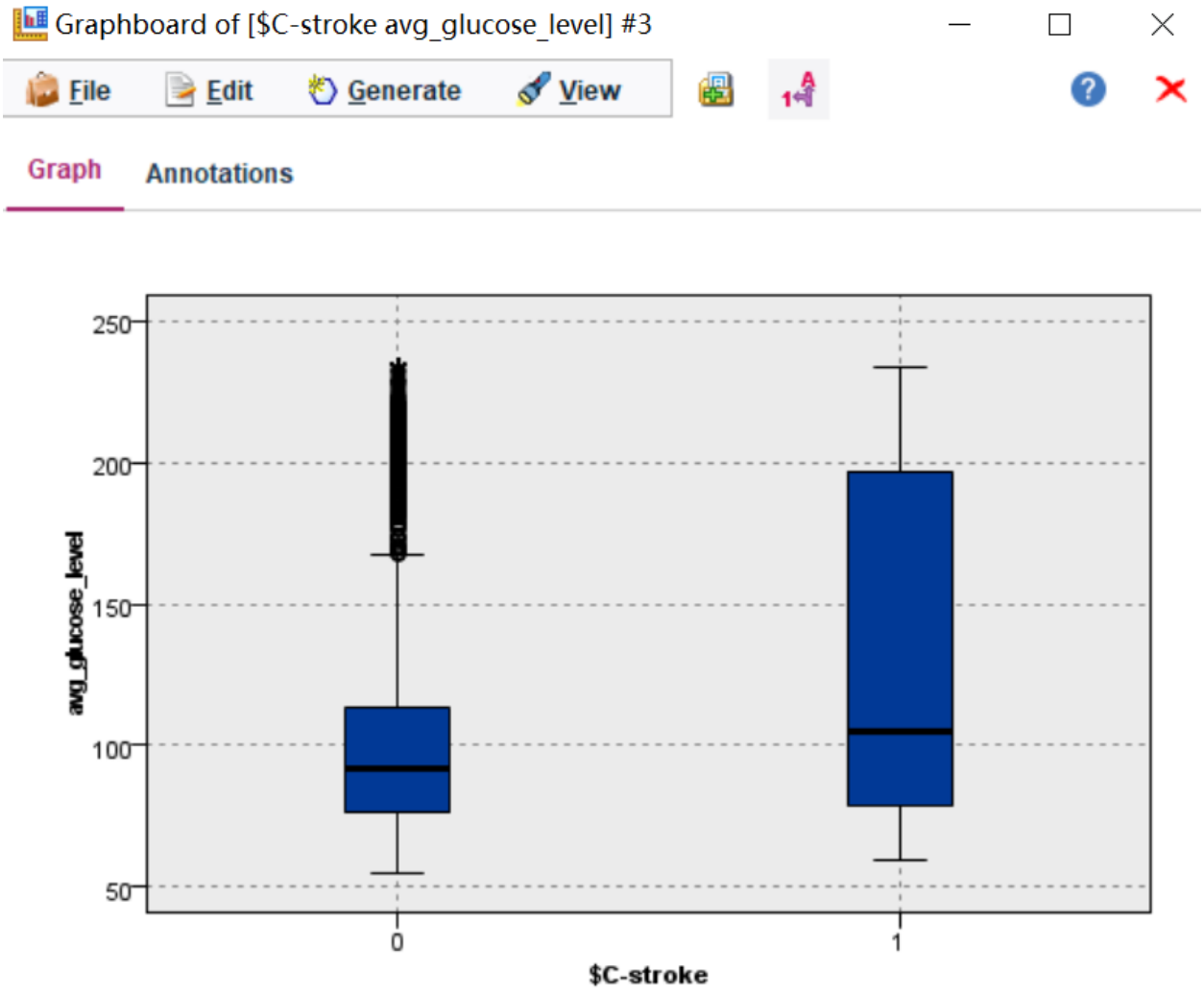


Figure 57. Relationship between average glucose level and stroke prediction

In addition, a combination of several factors will have an impact on the final travel conditions. As the patients who has higher/lower glucose level than normal range and elder age would have more chance to get stroke. Also, there are more combination conditions to get stroke is drawn by C5.0 algorithm in the tree.

8. Interpretation

8.1 Data Mining Pattern

Observing the tree C5.0. model is generated, we choose the deepest path, and trying to visualize each split conditions until the deepest leaf. We want to discover patterns from this deepest path as shown the way in Figure 58.

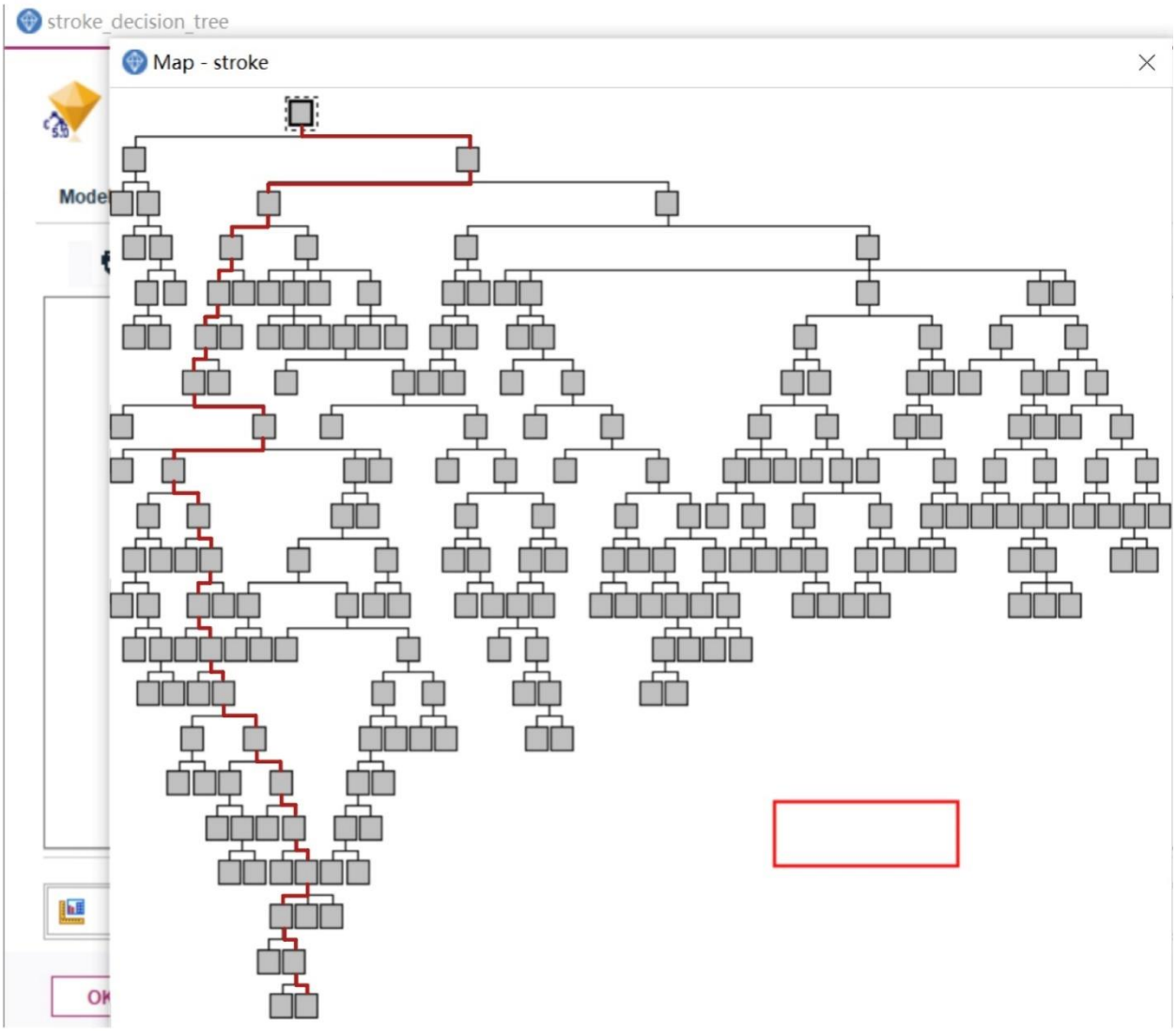

stroke_decision_tree
Map - stroke
Mode
OK

Figure 58. C5.0 Algorithm decision tree the deepest path

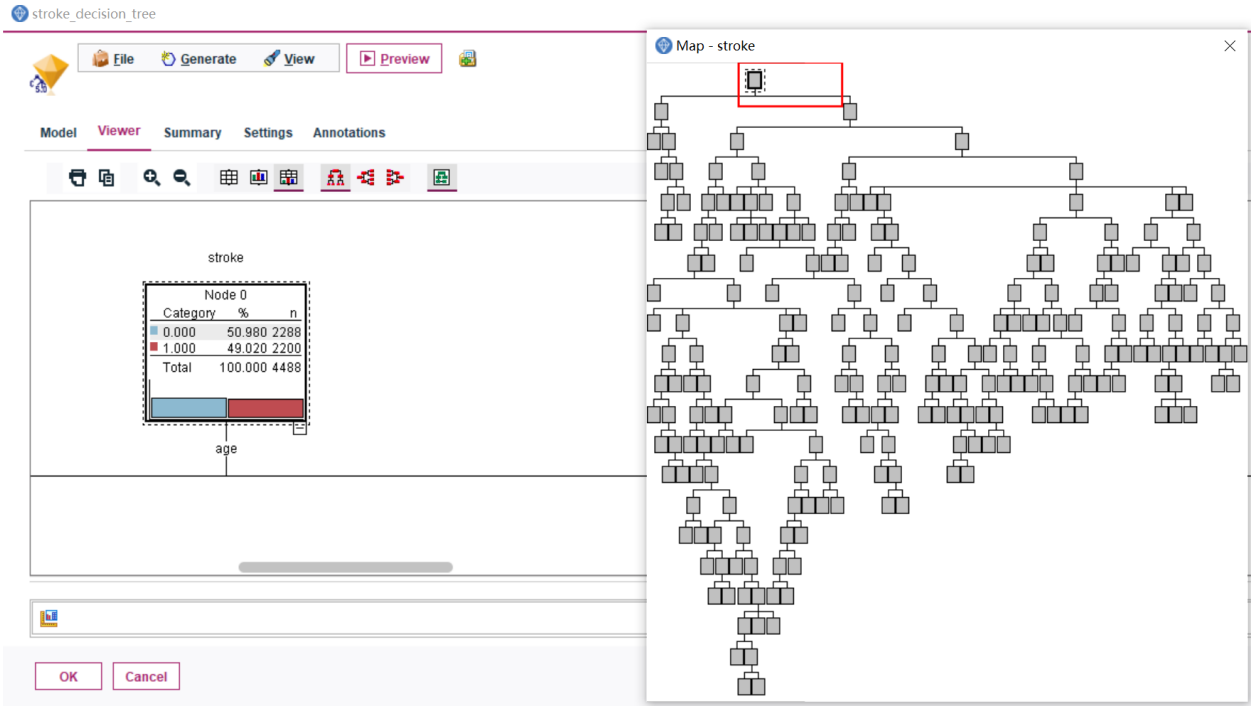


Figure 59. C5.0 Algorithm decision tree 1st branch

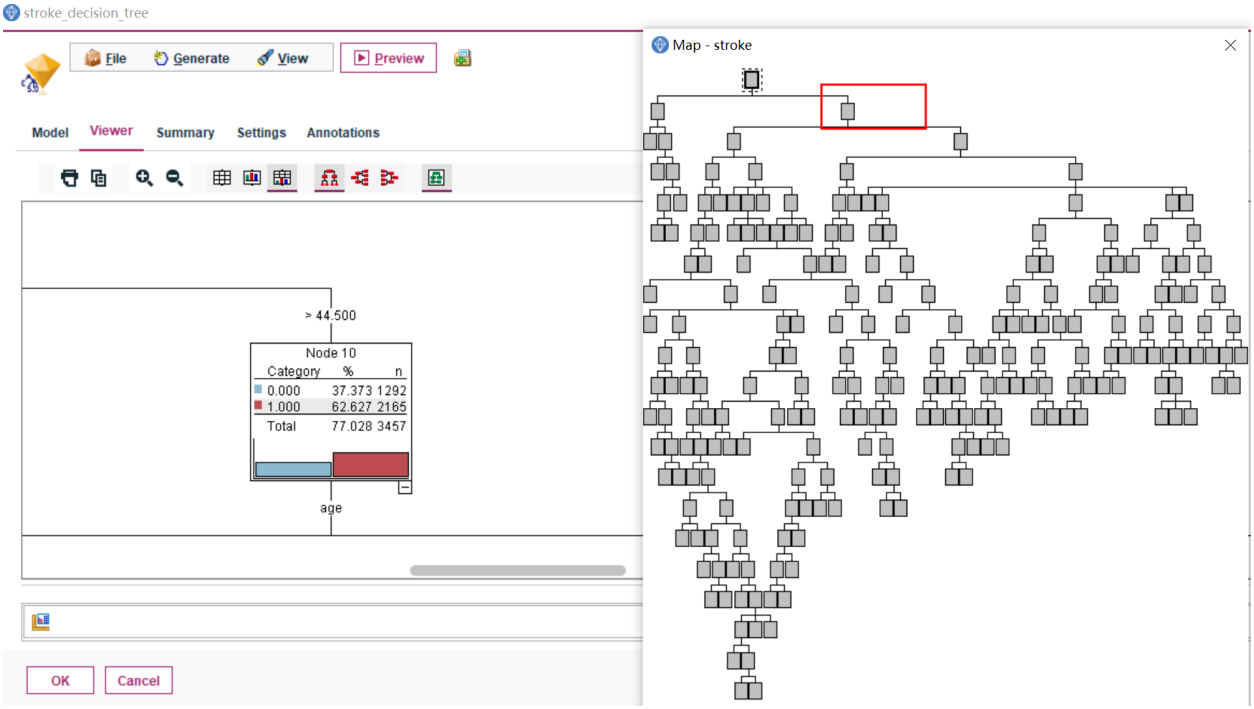


Figure 60. C5.0 Algorithm decision tree 2nd branch

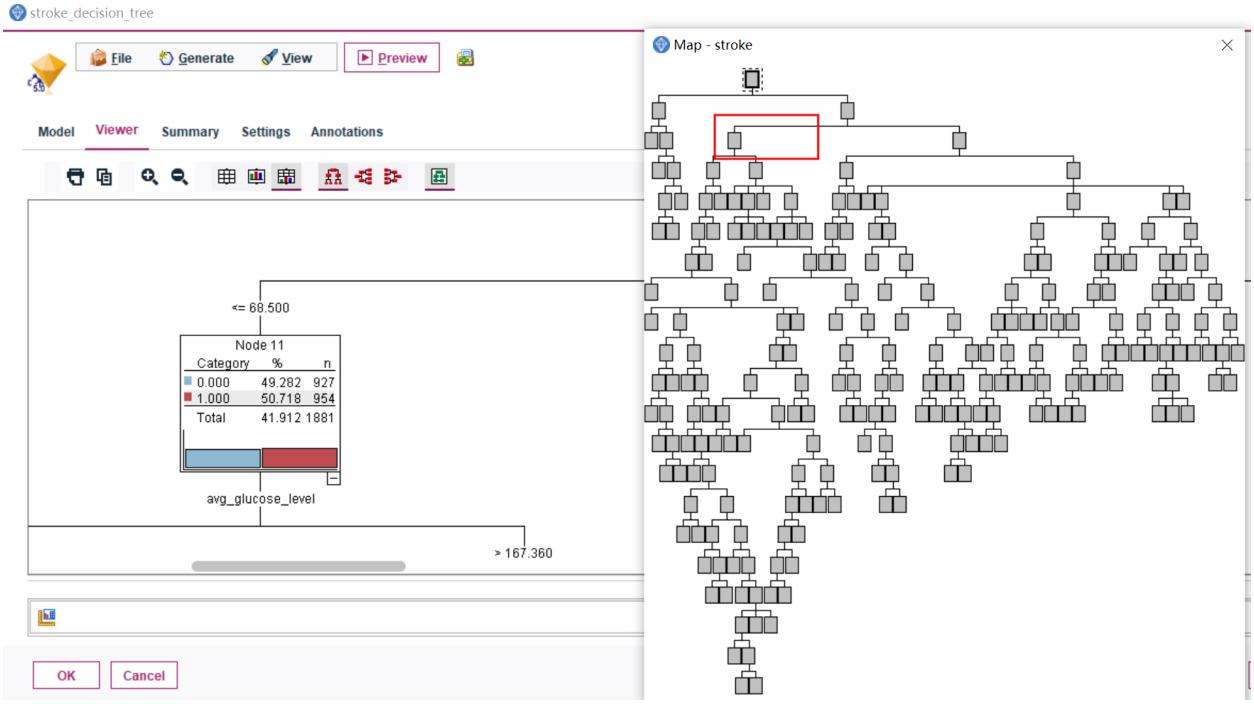


Figure 61. C5.0 Algorithm decision tree 3rd branch

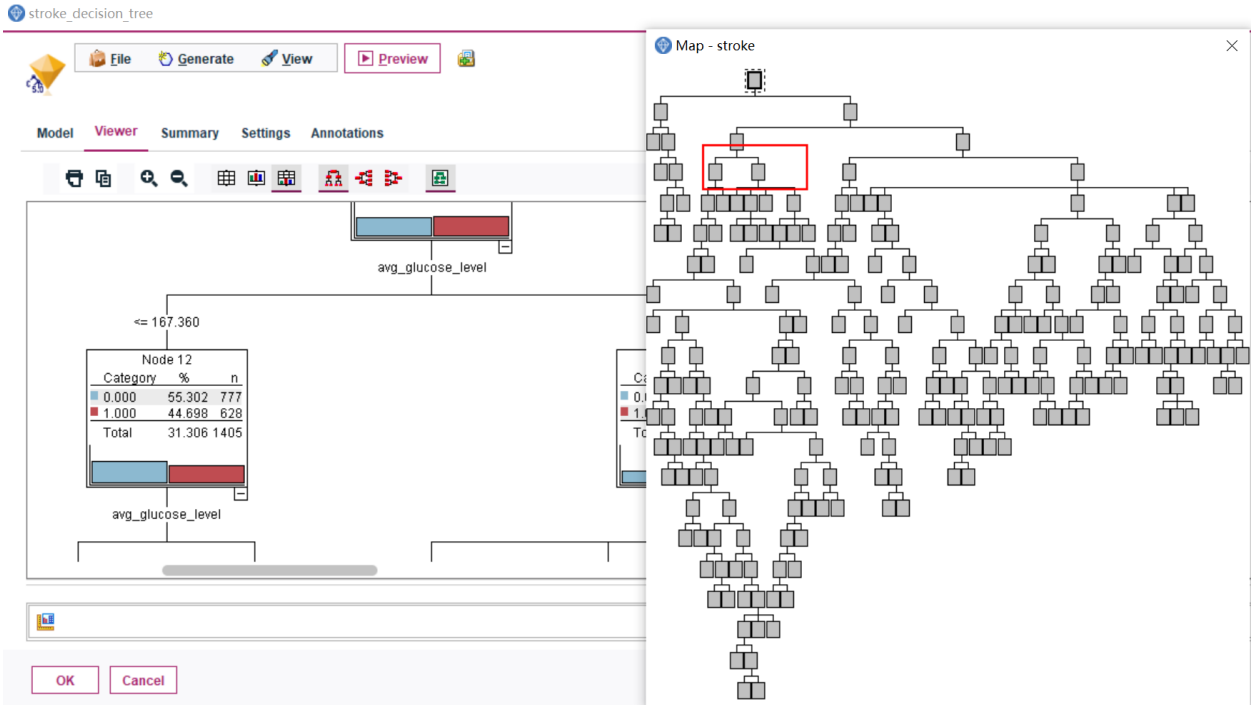


Figure 62. C5.0 Algorithm decision tree 4th branch

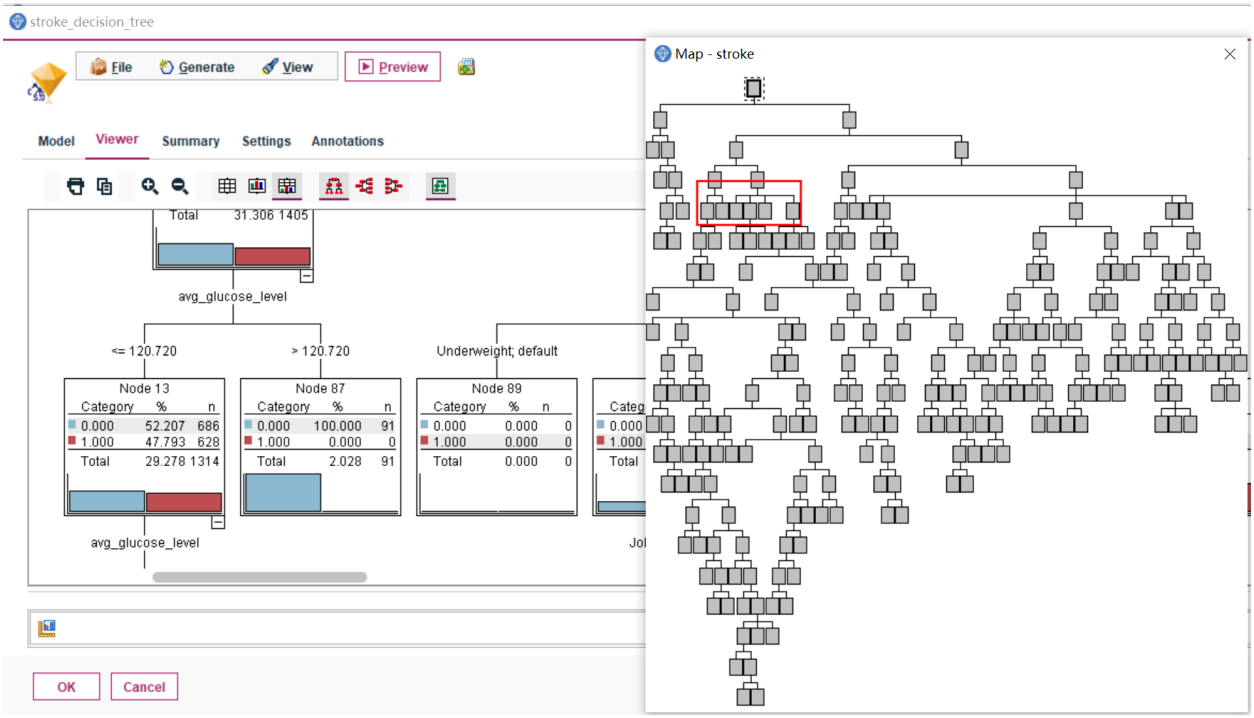


Figure 63. C5.0 Algorithm decision tree 5th branch

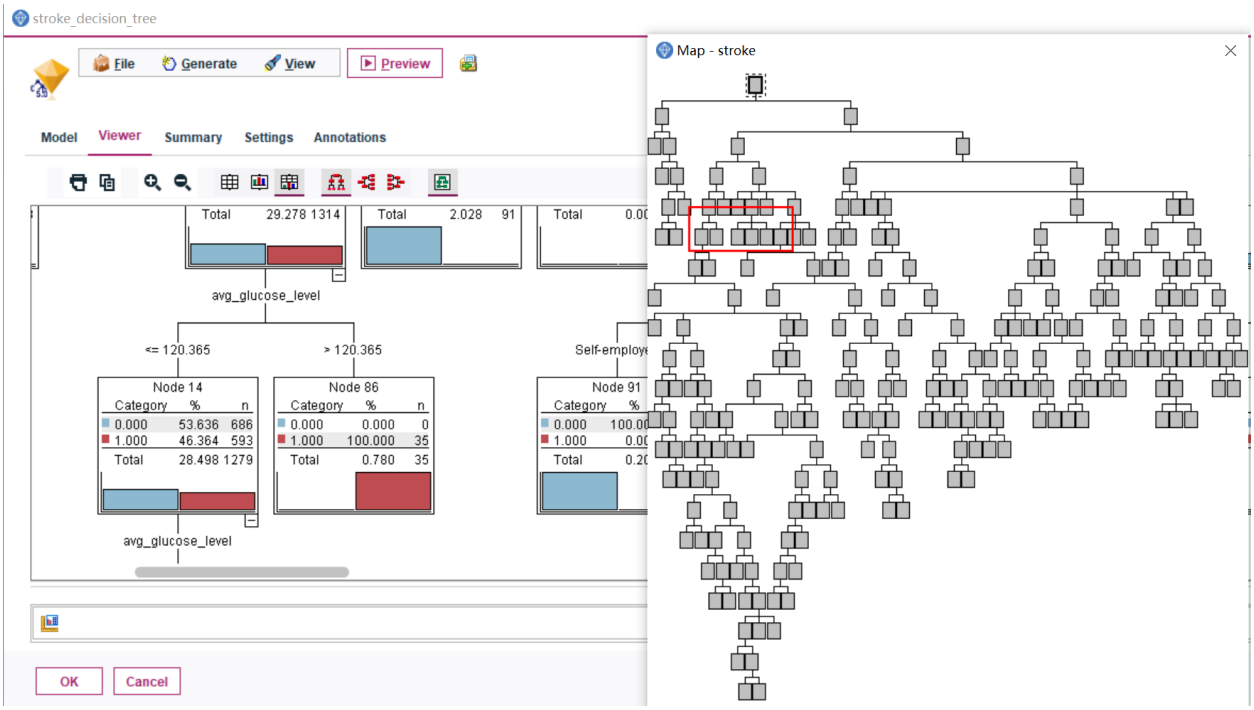


Figure 64. C5.0 Algorithm decision tree 6th branch

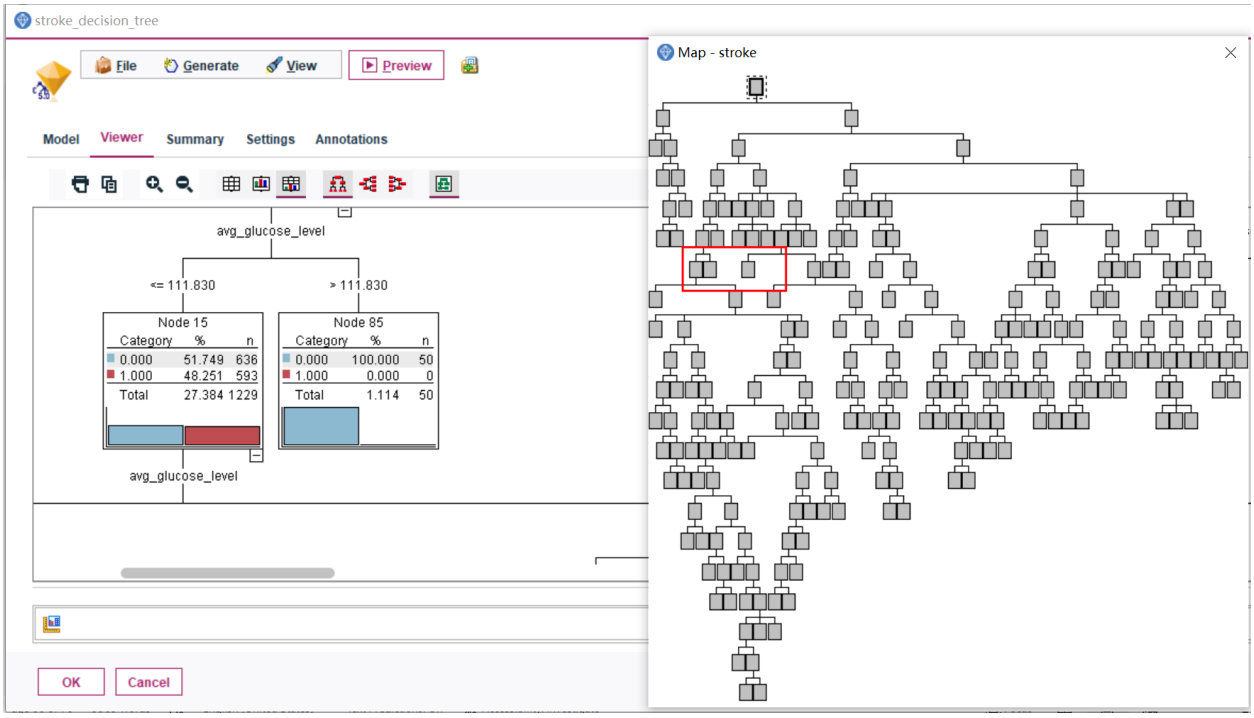


Figure 65. C5.0 Algorithm decision tree 7th branch

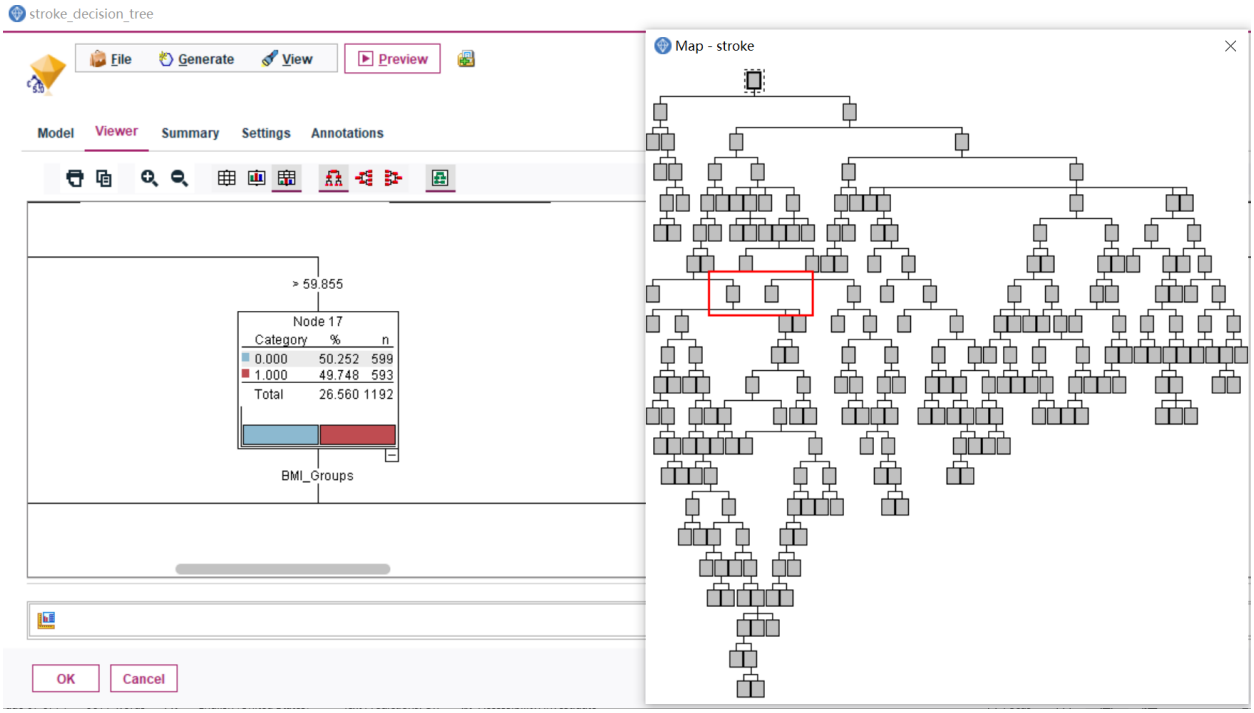


Figure 66. C5.0 Algorithm decision tree 8th branch

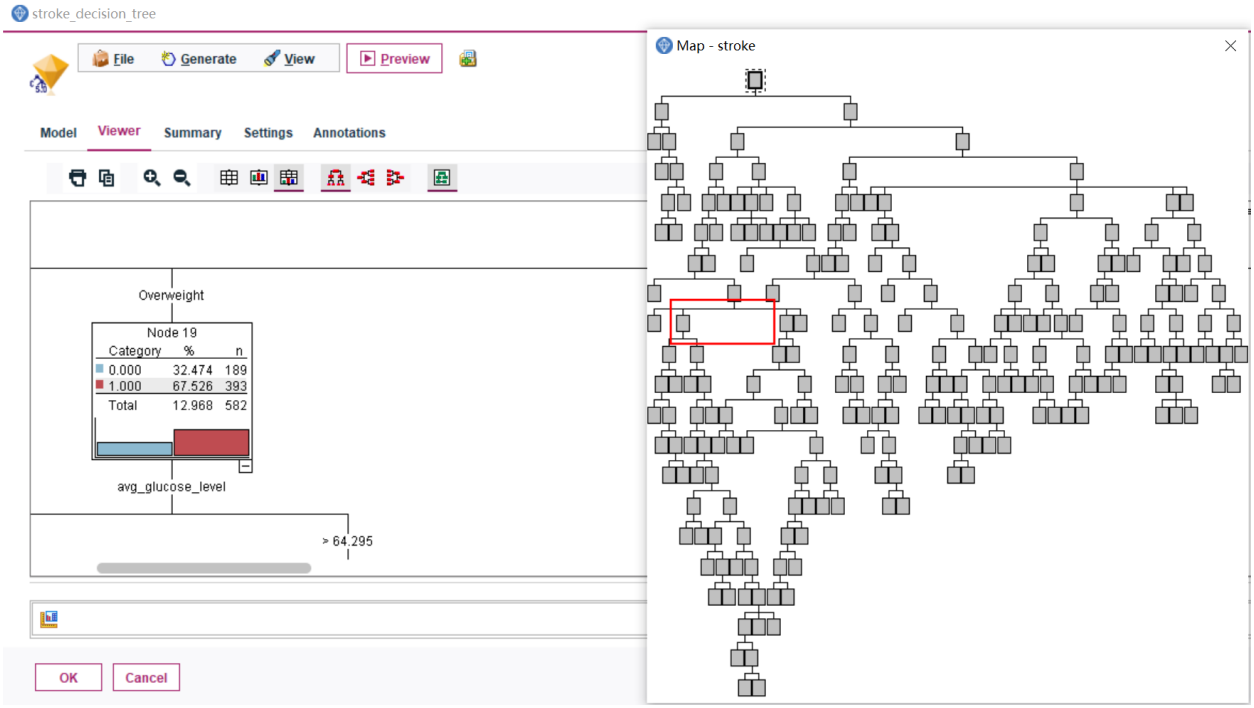


Figure 67. C5.0 Algorithm decision tree 9th branch

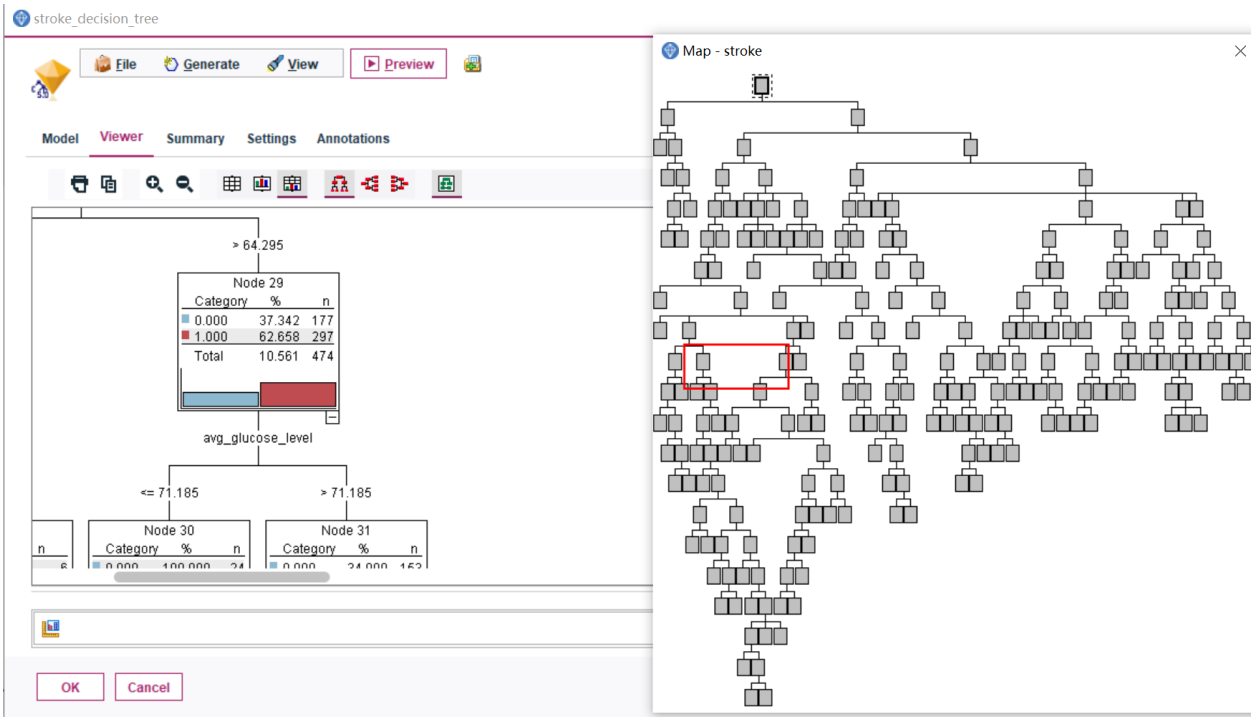


Figure 68. C5.0 Algorithm decision tree 10th branch

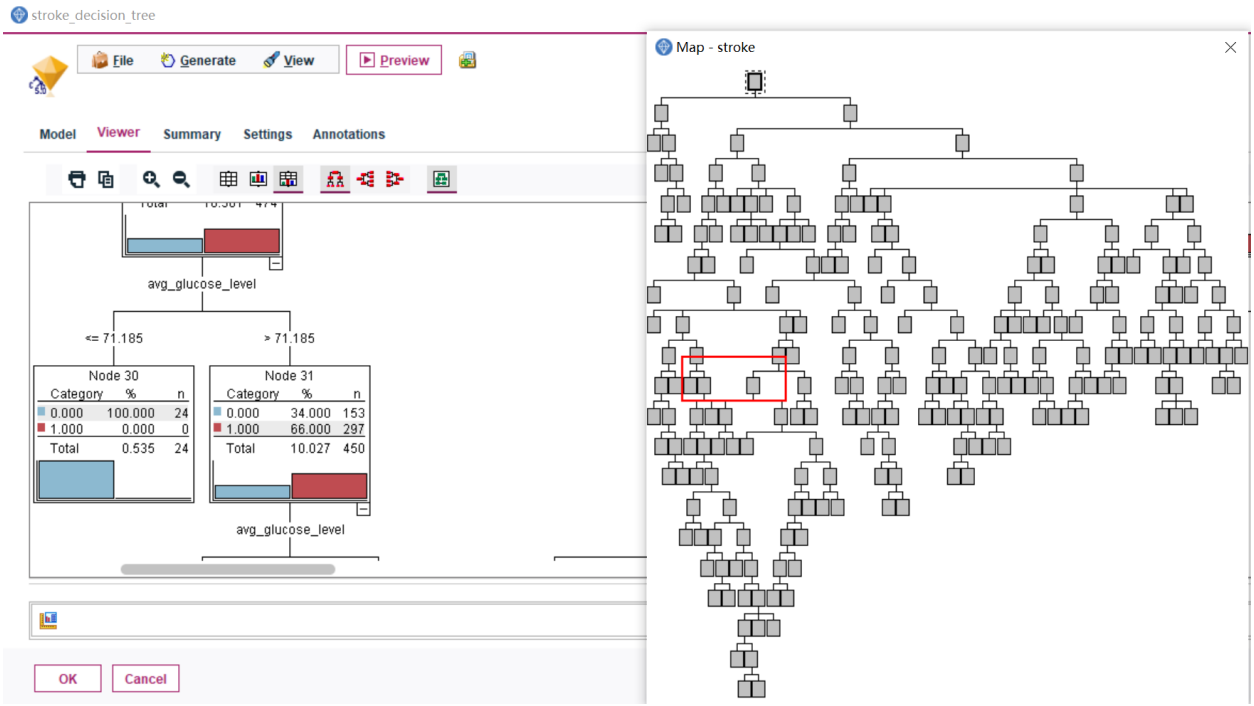


Figure 69. C5.0 Algorithm decision tree 11st branch

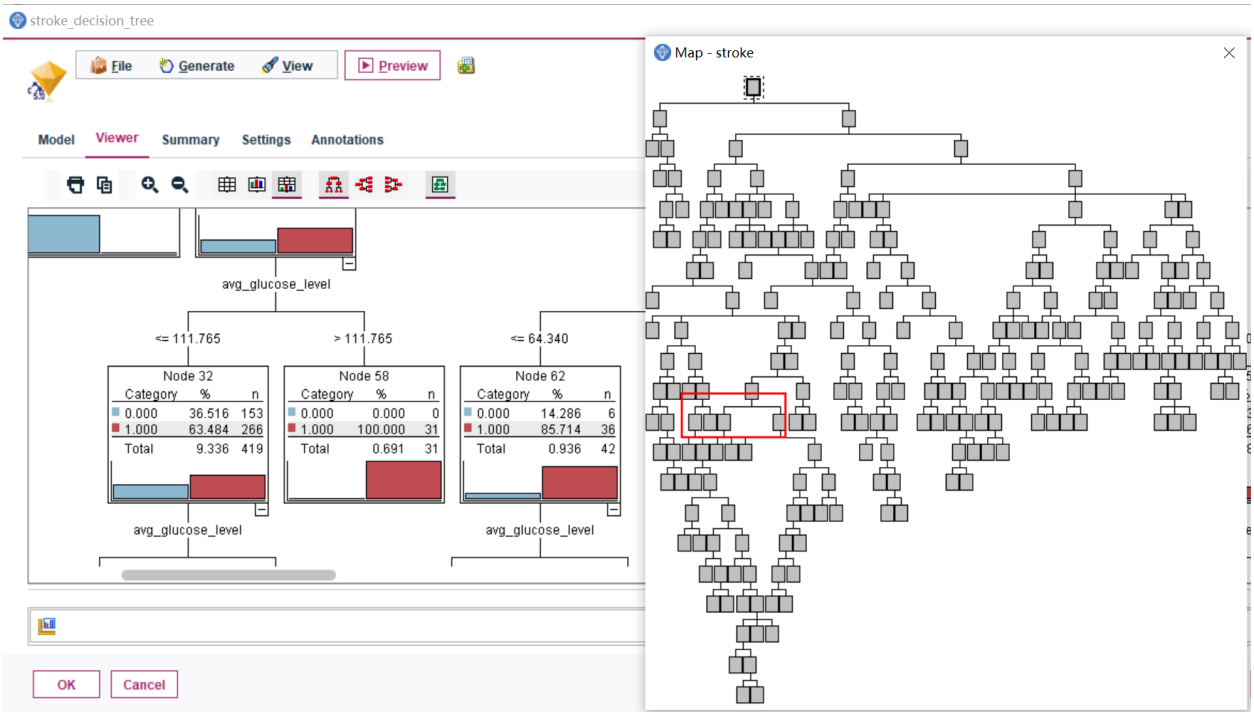


Figure 70. C5.0 Algorithm decision tree 12nd branch

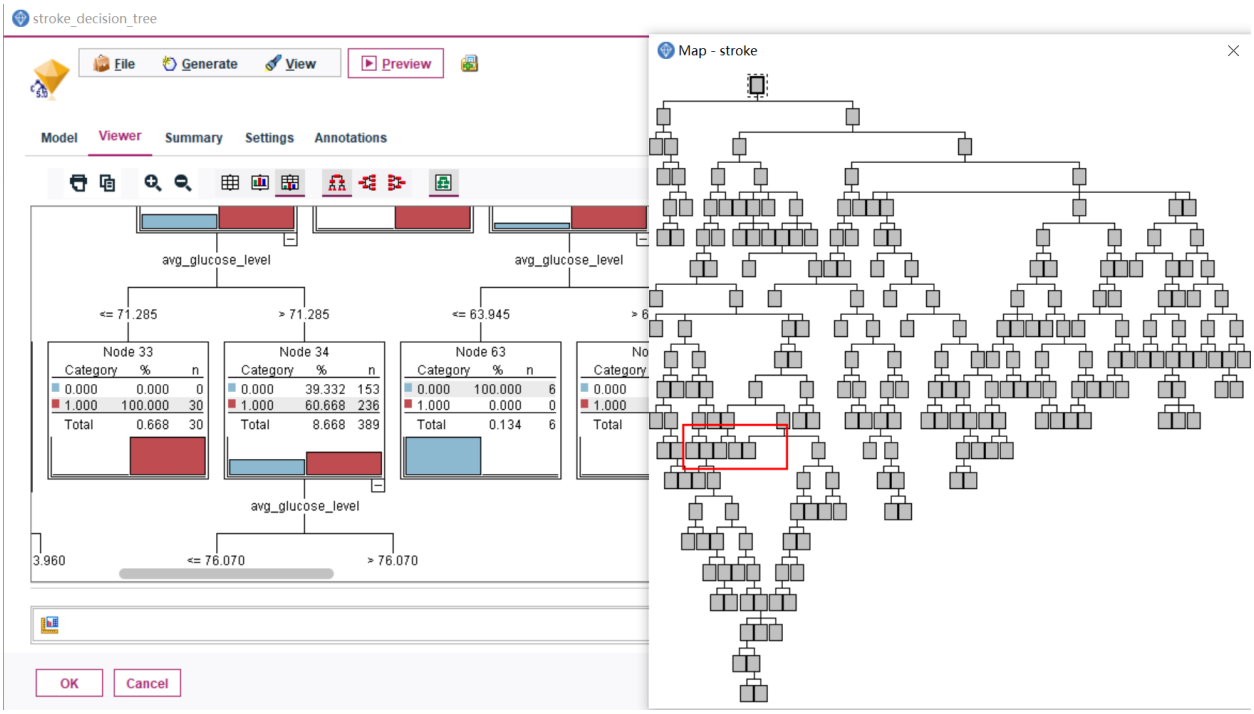


Figure 71. C5.0 Algorithm decision tree 13rd branch

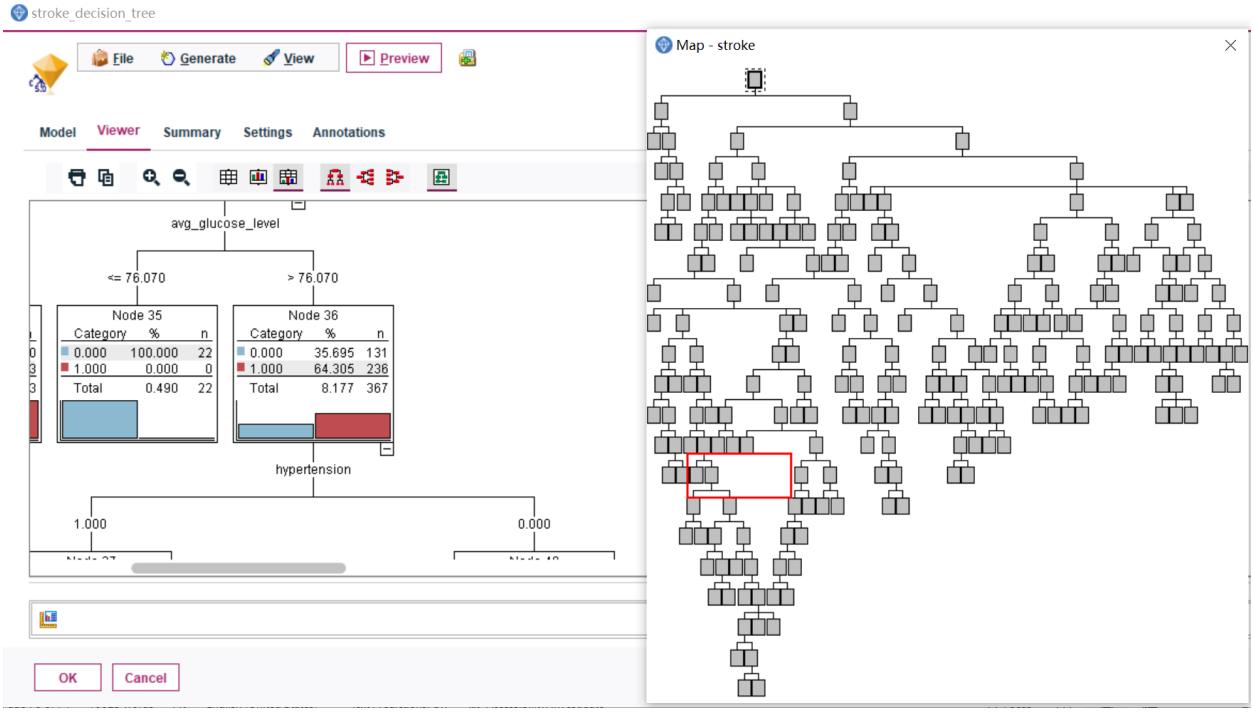


Figure 72. C5.0 Algorithm decision tree 14th branch

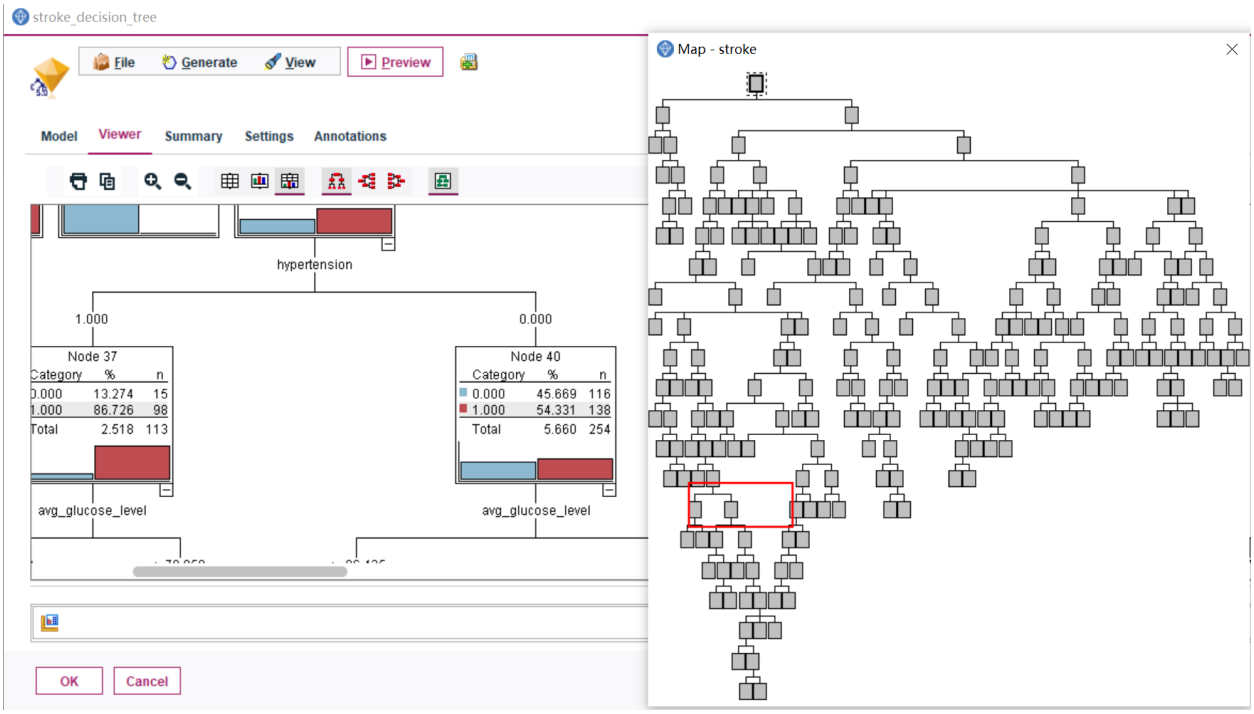


Figure 73. C5.0 Algorithm decision tree 15th branch

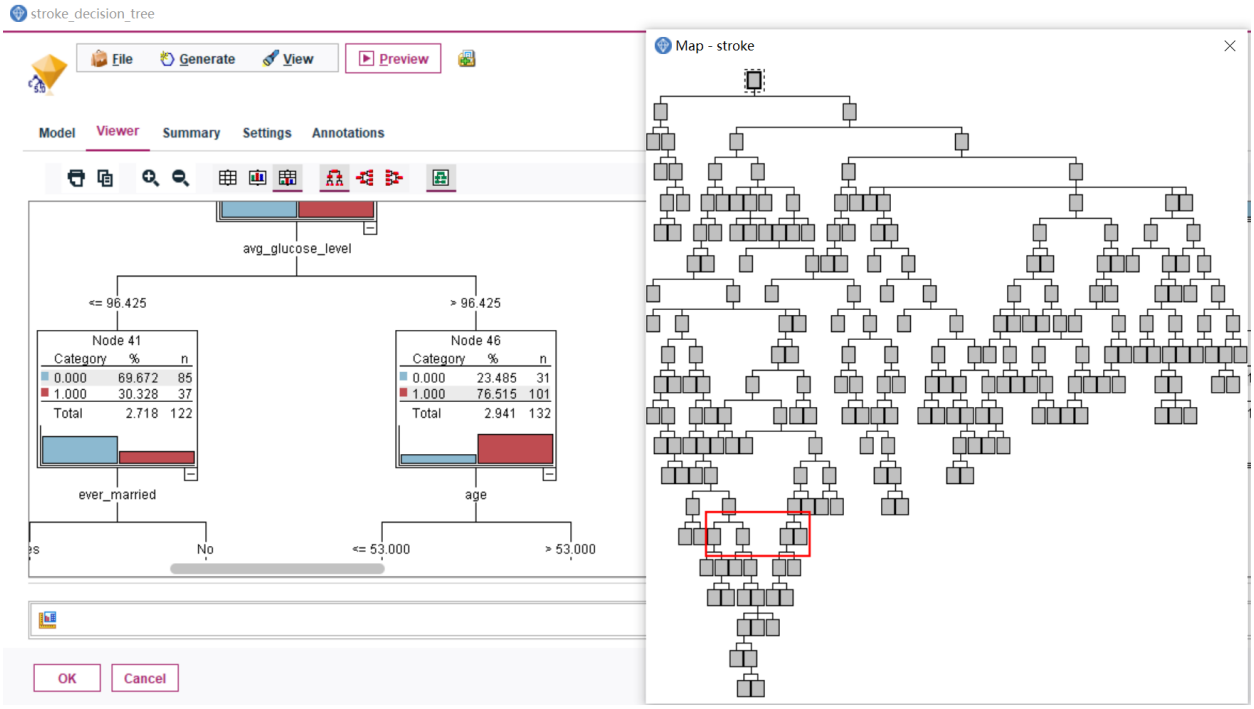


Figure 74. C5.0 Algorithm decision tree 16th branch

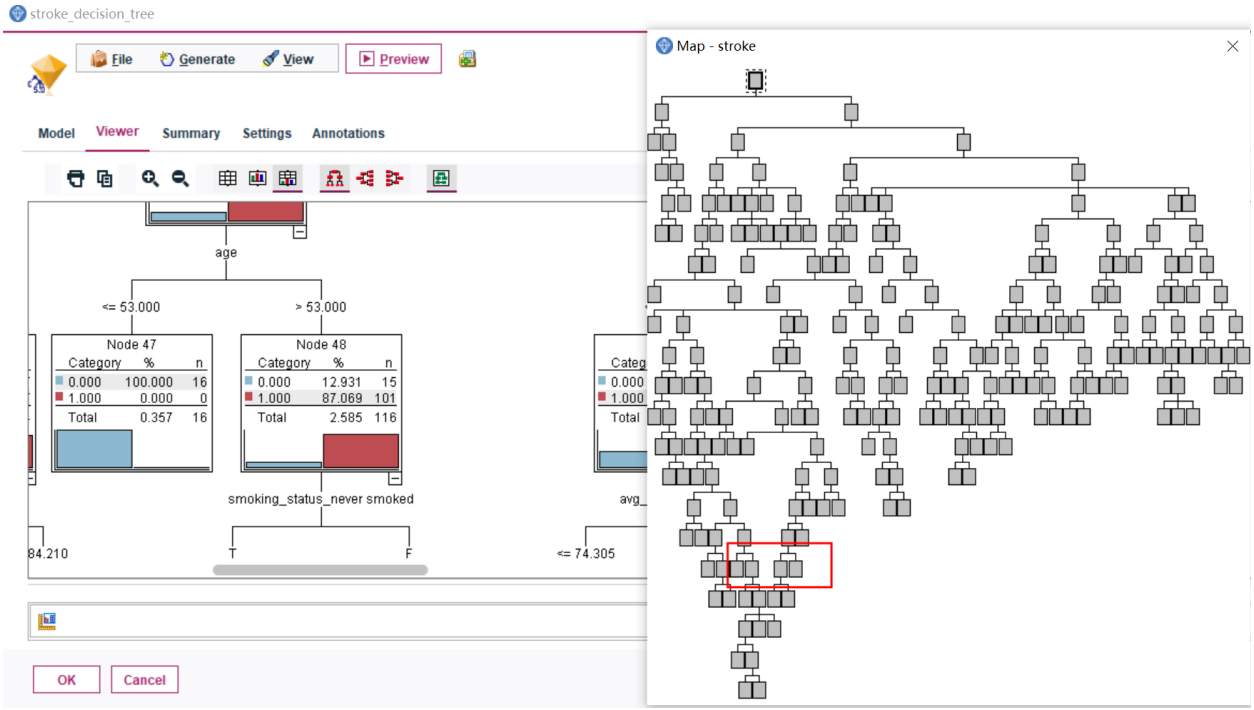


Figure 75. C5.0 Algorithm decision tree 17th branch

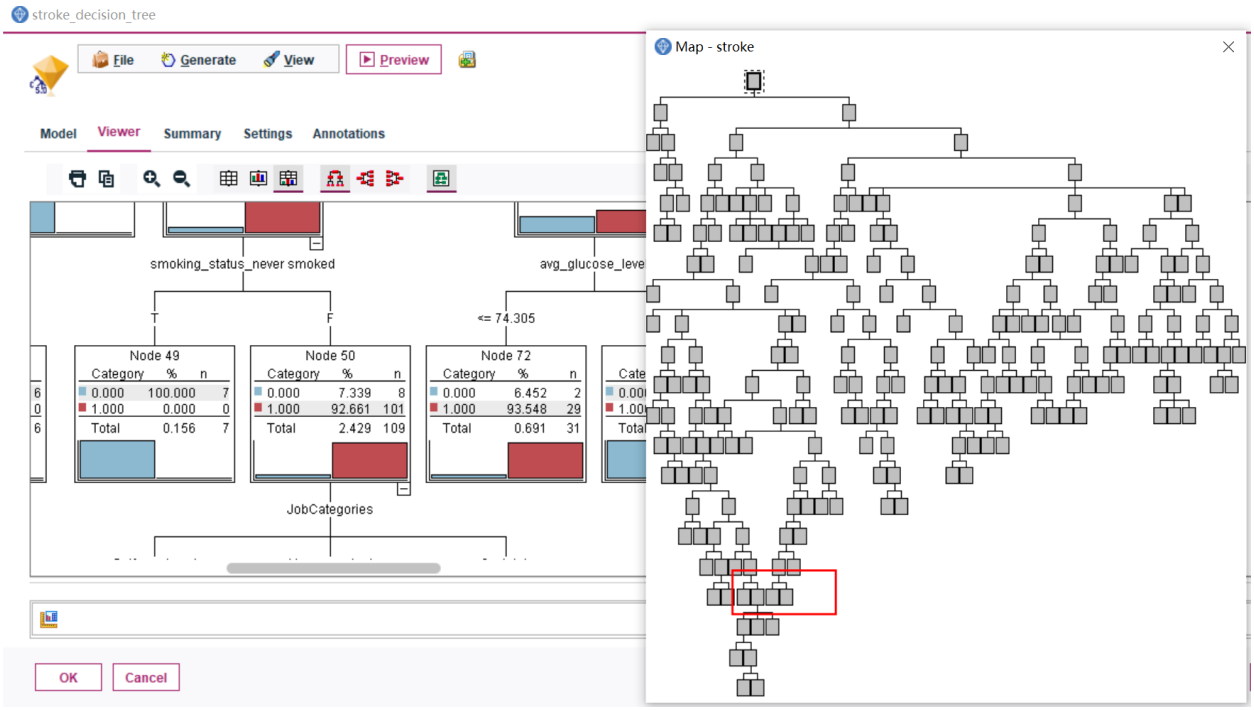


Figure 76. C5.0 Algorithm decision tree 18th branch

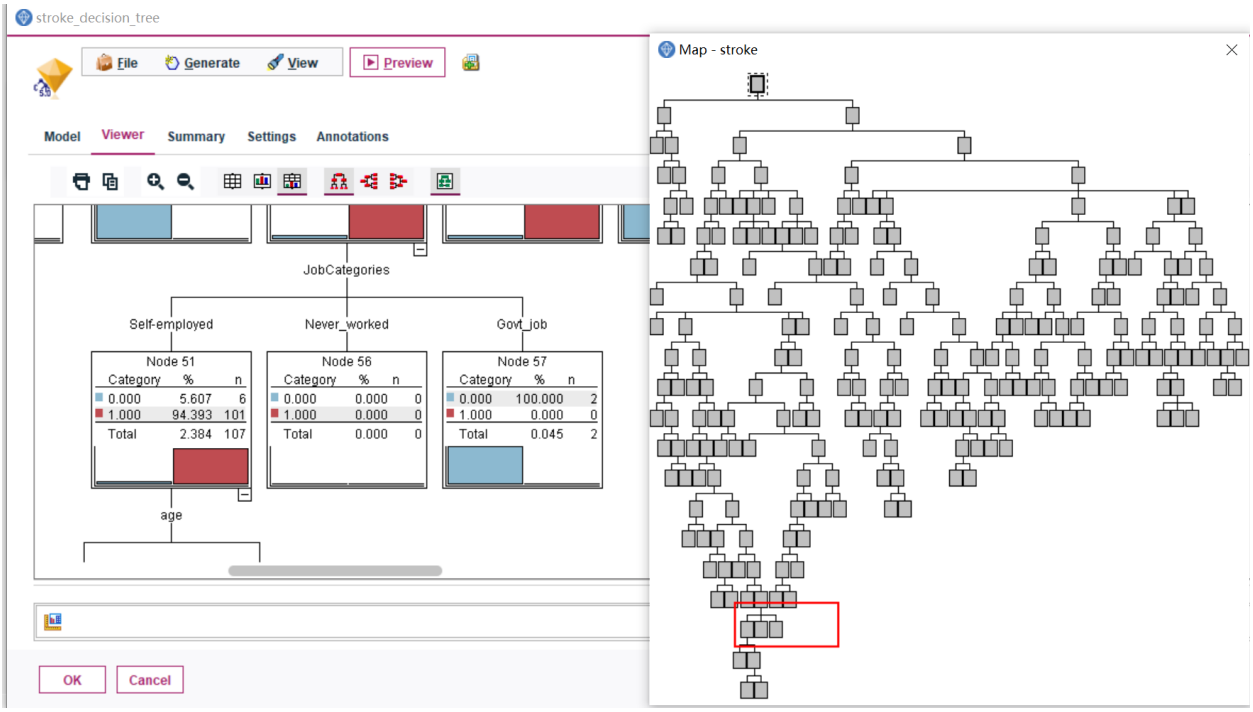


Figure 77. C5.0 Algorithm decision tree 19th branch

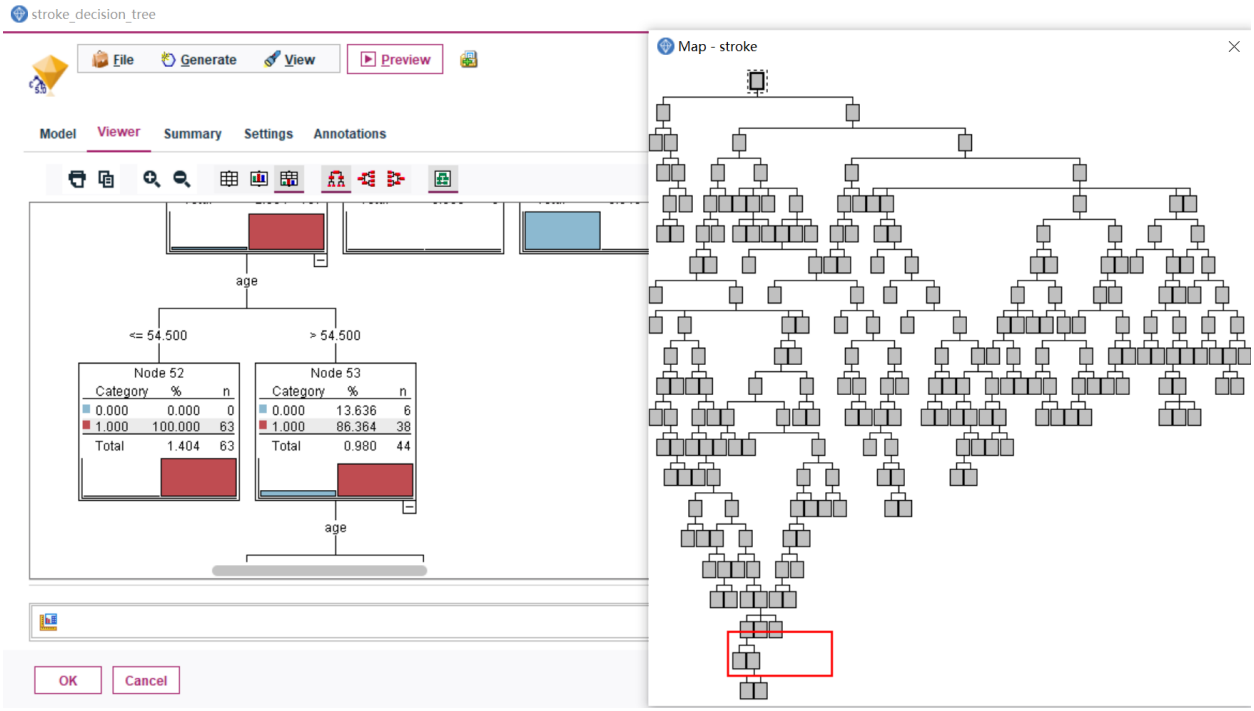


Figure 78. C5.0 Algorithm decision tree 20th branch

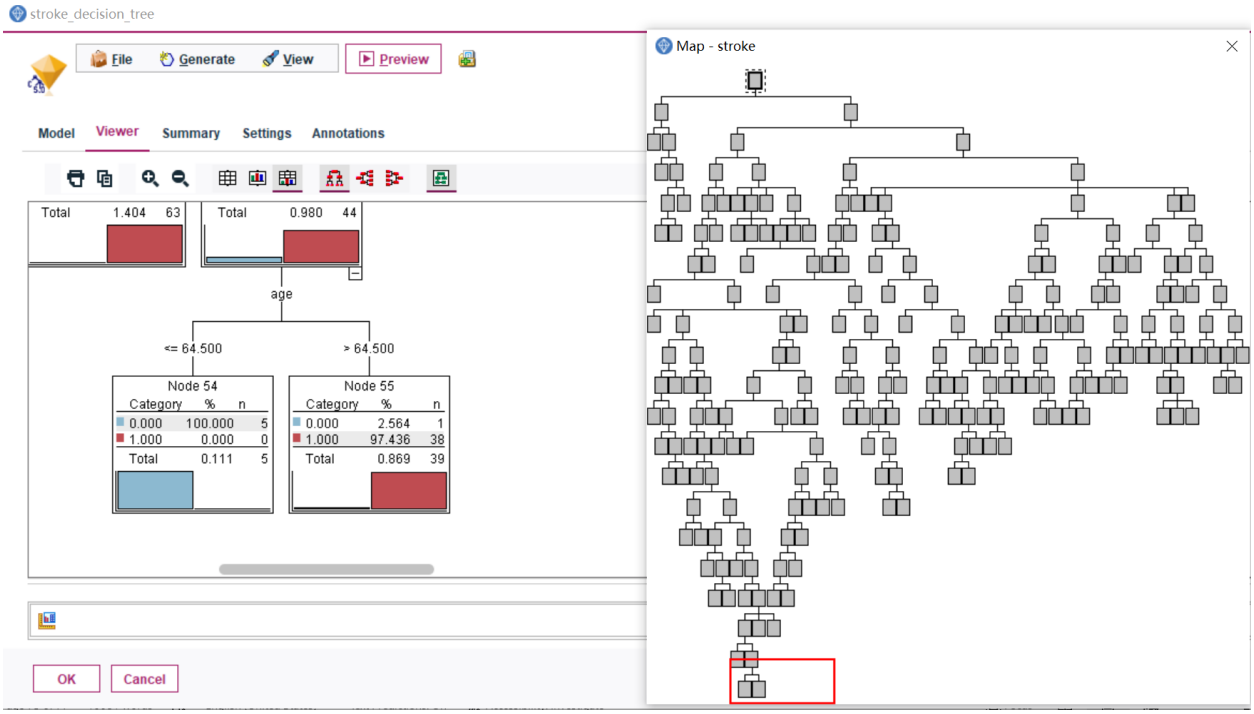


Figure 79. C5.0 Algorithm decision tree 21st branch

From above Figures, we can see this deepest path has 21 nodes, which means for every node, there is a condition to branch deeper, until there is no case to go further.

There are nearly half and half stroke and non-stroke cases in the dataset. Condition for the 1st node to branch is age <= 44.5 years old, age > 44.5 years old.

There are 62.627% patients get stroke and their age is over 44.5 years old. Condition for the 2nd node to branch is age as well but branch the age group to smaller groups.

Condition for the 3rd node to branch is age <= 68.5 years old and age > 68.5 years old.

Condition for the 4th node to branch is average glucose level <= 167.36 and >167.3.

Condition for the 5th node to branch is average glucose level <= 120.720 and >120.720.

Condition for the 6th node to branch is average glucose level <= 120.365 and >120.365.

Condition for the 7th node to branch is average glucose level <= 111.830 and >111.830. From these branches use average glucose level to branch, we can see that the patients have glucose level less or equal to 111.830 have around 50% chance to get stroke, but no patients who has glucose level higher than 111.830 got stroke.

Condition for the 8th node to branch is average glucose level <= 59.855 and >59.855. we can see the patients have glucose level higher than 59.855 have around 50% chance to get stroke.

Therefore, from this C5.0 Algorithm, the deepest path show that patients have glucose level from 59.855 to 111.830 have 50% chance to get stroke. This range includes normal glucose level, we need to do further discover to find whether patients with blood sugars slightly below and slightly above the normal range are more likely to have a stroke. Also, how much below or above the normal range is blood sugar more likely to cause a stroke.

Condition for the 9th node to branch is BMI group. There is no underweight and normal weight patients get stroke, overweight patients have 67.526% chance to get stroke, and obese patients have 39.716% chance to get stroke.

Condition for the 10th node to branch is average glucose level <=64.295 and > 64.295.

Condition for the 11st node to branch is average glucose level <=71.185 and > 71.185. There is no patient whose glucose level is under or equal to 71.185 get stroke.

Condition for the 12nd node to branch is average glucose level <=111.765 and > 111.765. There is no patient whose glucose level is higher than 111.765 get stroke.

Condition for the 13rd node to branch is average glucose level <=71.285 and > 71.285.

Condition for the 14th node to branch is average glucose level <=76.070 and > 76.070. There is no patient whose glucose level is under or equal to 76.070 get stroke. Therefore, from this C5.0 Algorithm, the deepest path show that patients have glucose level from 76.070 to 111.765 have 60% chance to get stroke.

Condition for the 15th node to branch is hypertension. Patients who have hypertension get 86.726% chance to get stroke, compare with patients who do not has hypertension get 54.331% chance to get stroke. We observe, patients who have hypertension of illness have more chance to get stroke.

Condition for the 16th node to branch is average glucose level <=96.425 and > 96.425. Patients have glucose level higher than 96.425 have 76.515% chance to get stroke. We observe, patients who have slightly higher glucose level are more likely to get stroke, as the normal range of glucose level is 70 – 99 mg/dl.

Condition for the 17th node to branch is age <=53 and > 53. There is no patient whose age is under or equal to 53. And patients whose age is higher than 53 have 87.069% chance to get stroke.

Condition for the 18th node to branch is smoking status. There is no patient who never smoke get stroke. And patients who smokes have 92.661% chance to get stroke.

Condition for the 19th node to branch is work type. There is no patient who works for government get stroke. And patients who are self-employed have 94.393% chance to get stroke. It suggests that stable jobs (such as government jobs) reduce the likelihood of people

having strokes, while self-employed people have a higher risk of stroke, perhaps because self-employed people are under more stress.

Condition for the 20th node to branch is age <=54.5 and > 54.5. There is no patient whose age is under or equal to 54.5. And patients whose age is higher than 54.5 have 86.364% chance to get stroke.

Condition for the 21st node to branch is age <=64.5 and > 64.5. There is no patient whose age is under or equal to 64.5. And patients whose age is higher than 64.5 have 97% chance to get stroke. Much higher chance to get stroke, which means when people getting older, they have higher chance to get stroke.

There are multiple predictors that can influence the final trip conditions, some of which can affect the results when combined. As we have analysis in the model above, people with high age, higher blood sugar, BMI (overweight), high blood pressure, smoking habits (smoking), and type of work (self-employed) are more likely to have a stroke.

In fact, the real factor should be age. As we age, glucose levels increase with age. The main cause of the increase in glucose levels with age is directly related to ageing, a process of degeneration which includes degenerative processes of energy and glucose metabolism, thus making age an important and independent risk factor for diabetes, and therefore the middle-aged and older people, more importantly, should pay more attention to the prevention of diabetes (Harris et al., 1987). From an age-related perspective, the data itself points to the fact that ageing increases the chance of stroke. The medical literature states that as one ages, the blood vessels within the brain may begin to suffer from microcirculation and marco circulation, which will increase the chances of acquiring a stroke (Young & Yousufuddin, 2019).

Obesity is an important factor in the development of coronary heart disease and also has a very important impact on cerebral infarction, as it can lead to hypertension, hyperlipidaemia and hyperglycaemia, which can accelerate cardiovascular disease and lead to stroke (Kernan et al., 2013).

Hypertension is a trigger for strokes, as blood pressure increases in the body and vascular tone expands, causing damage to the vessel walls in the blood vessels. As a result of this injury, blood fat in the blood vessels tends to penetrate the lining and build up too much pressure, leading to atherosclerosis, and the accumulation of lipids blocking the blood vessels can lead to stroke (Alloubani et al., 2018).

From the point of view of smoking, after reviewing some literature, it can be concluded that smoking tends to lead to endothelial dysfunction of blood vessels, which can easily lead to the formation of blood clots and blockage of blood vessels, thus triggering stroke, and also increases the incidence of coronary heart disease, cerebrovascular disease and the vascular diseases surrounding them, accounting for more than 18.9% of all risk factors for stroke. (Centers for Disease Control and Prevention, 2021).

From a work type perspective, factors such as smoking, alcohol consumption and obesity are not the only causes of stroke; factors such as stress and late nights also contribute to an increased risk of cerebral infarction.

8.2 Data Visualization

Figure 80 is showing the overall structure of the model.

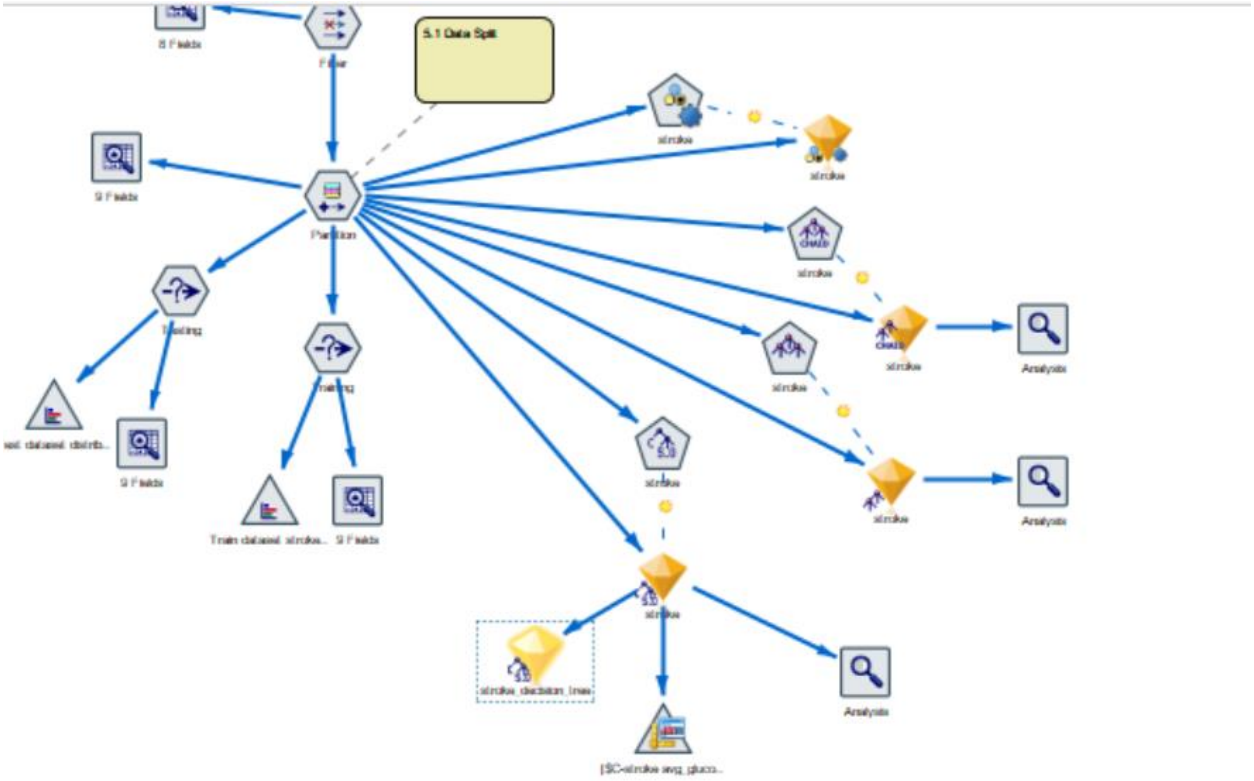


Figure 80. The overall structure of the model

Figure 81 is showing the dataset before the data split to train and test sets.

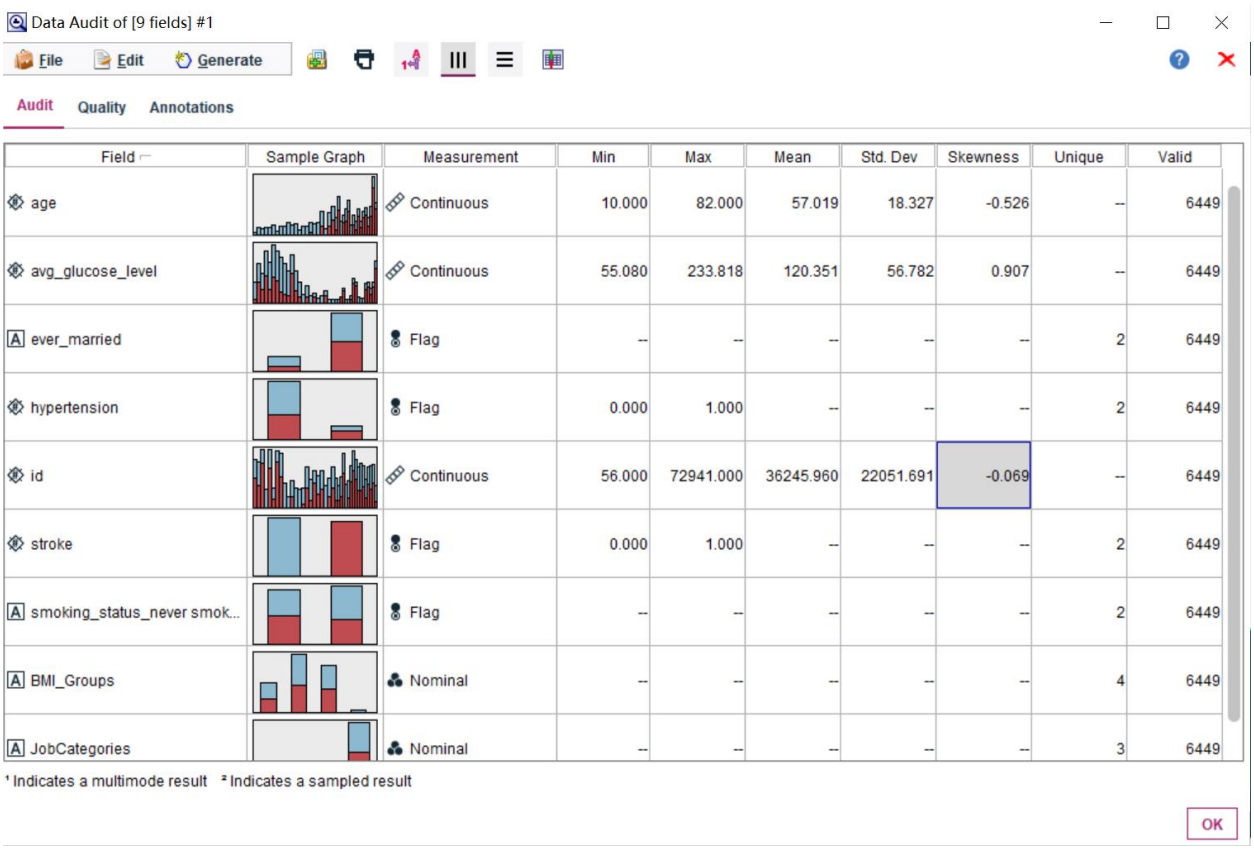


Figure 81. dataset before the data split to train and test sets

Figure 82 is showing the dataset after the data split to train and test sets. Before dataset split process and run the model, we delete the attribute id, to avoid any impact on model building and prediction accuracy later on.

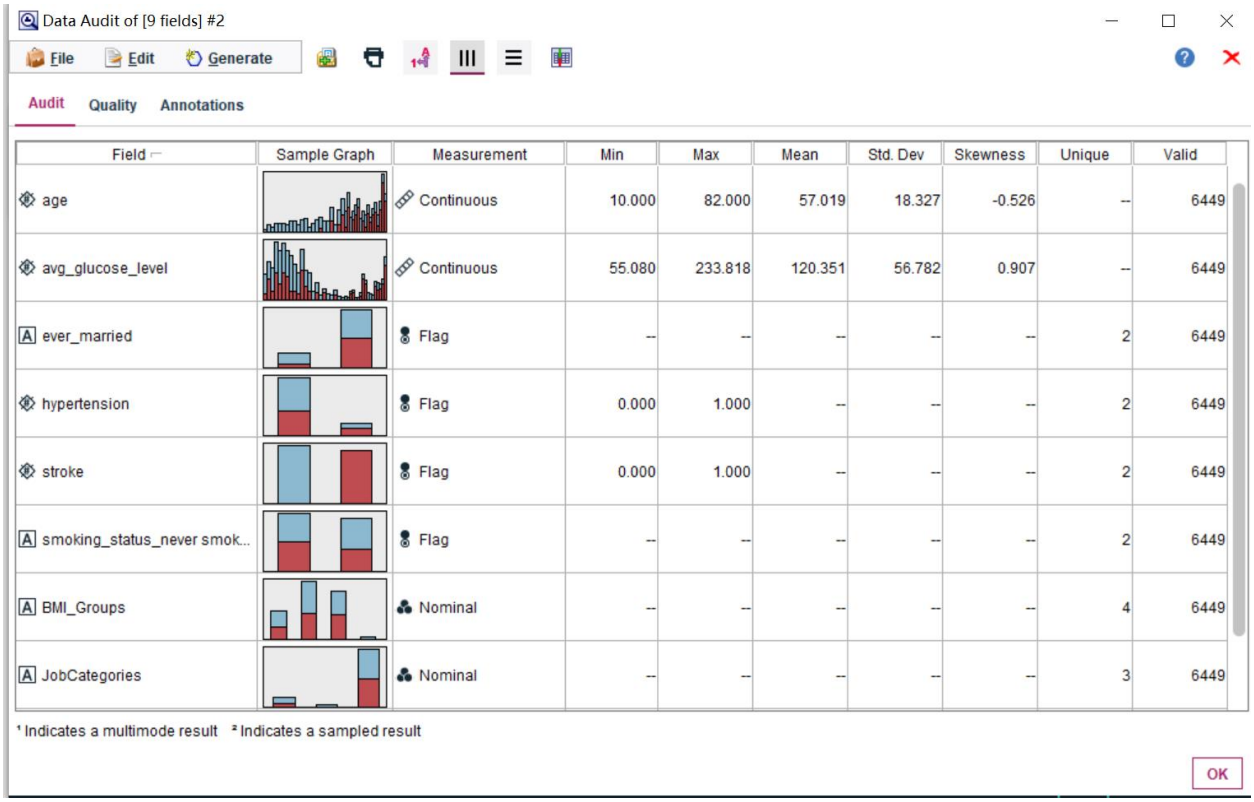


Figure 82. dataset after the data split to train and test sets

Figure 83 is showing the test dataset after the data split.

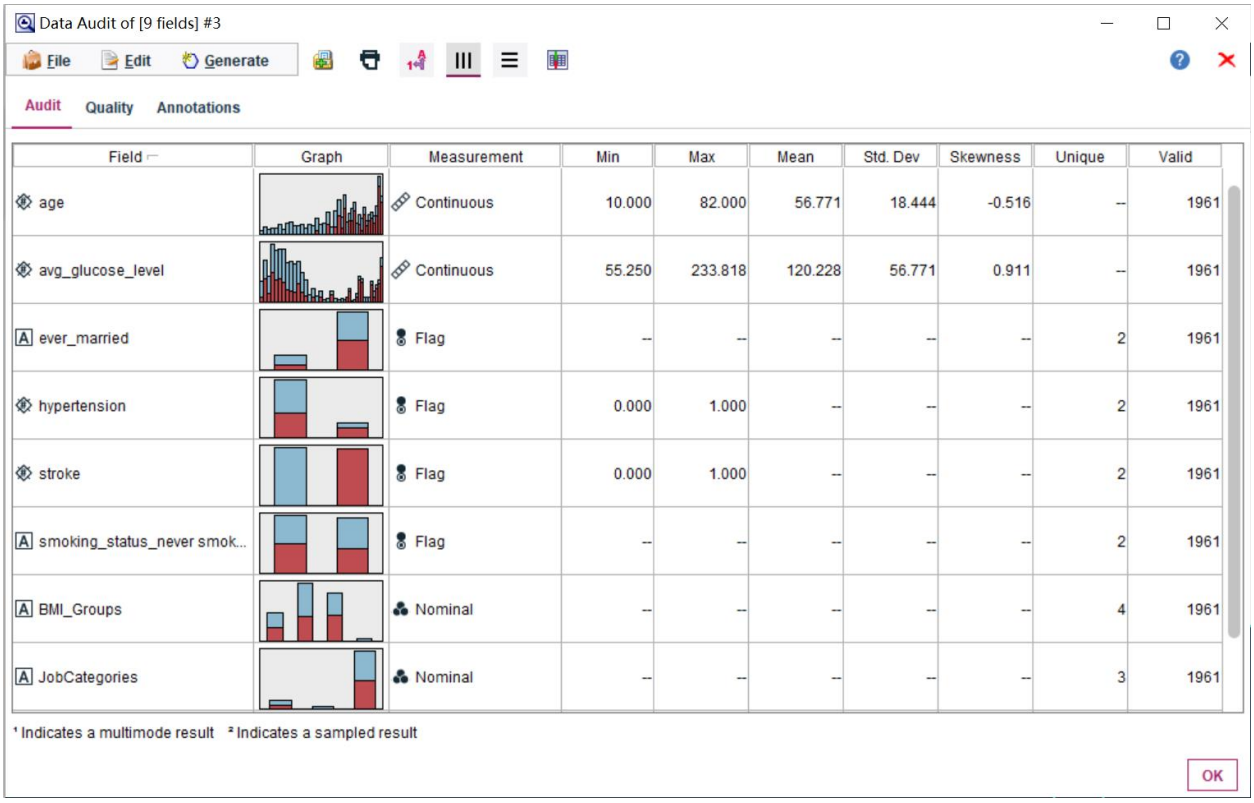


Figure 83. Test dataset after the data split

Figure 84 is showing the train dataset after the data split.

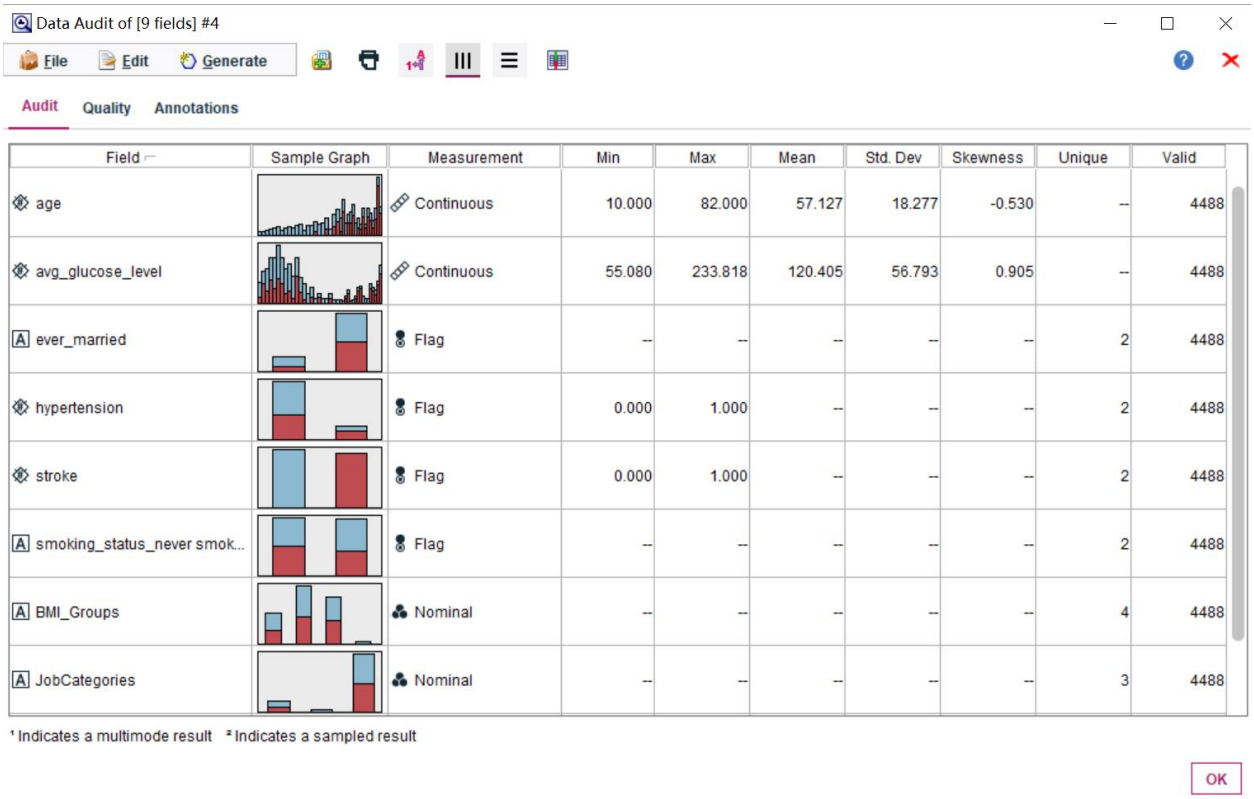


Figure 84. Train dataset after the data split

We can see that each attribute has a similar distribution in the training and test datasets, demonstrating that the dataset is evenly distributed between the training and test sets.

Figure 85 is showing the rule sets generated by C5.0. model. We can see that there are 96 rules are generated by C5.0. model. Also, the predictor(attributes) importance identified by the model. Age is the most important predictor, BMI group is the second important predictor, average glucose level is the third important predictor, following with smoking status, marriage status, job category and hypertension. We will clarify the relationship between these important predictors with stroke as following.

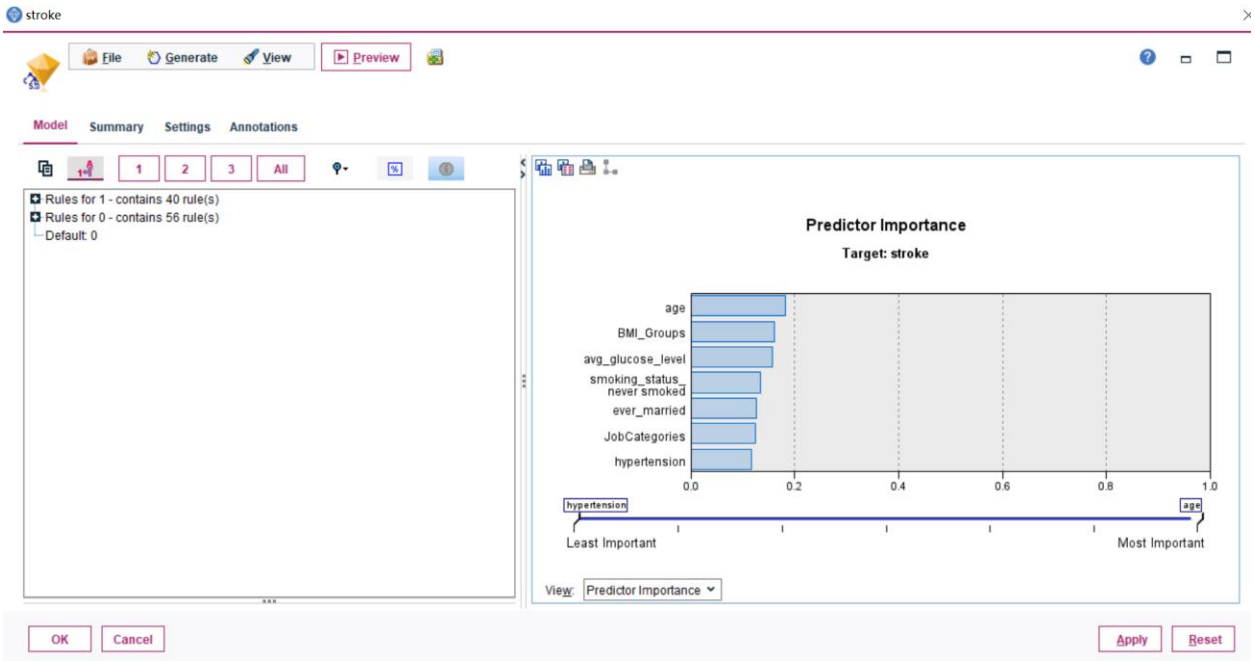
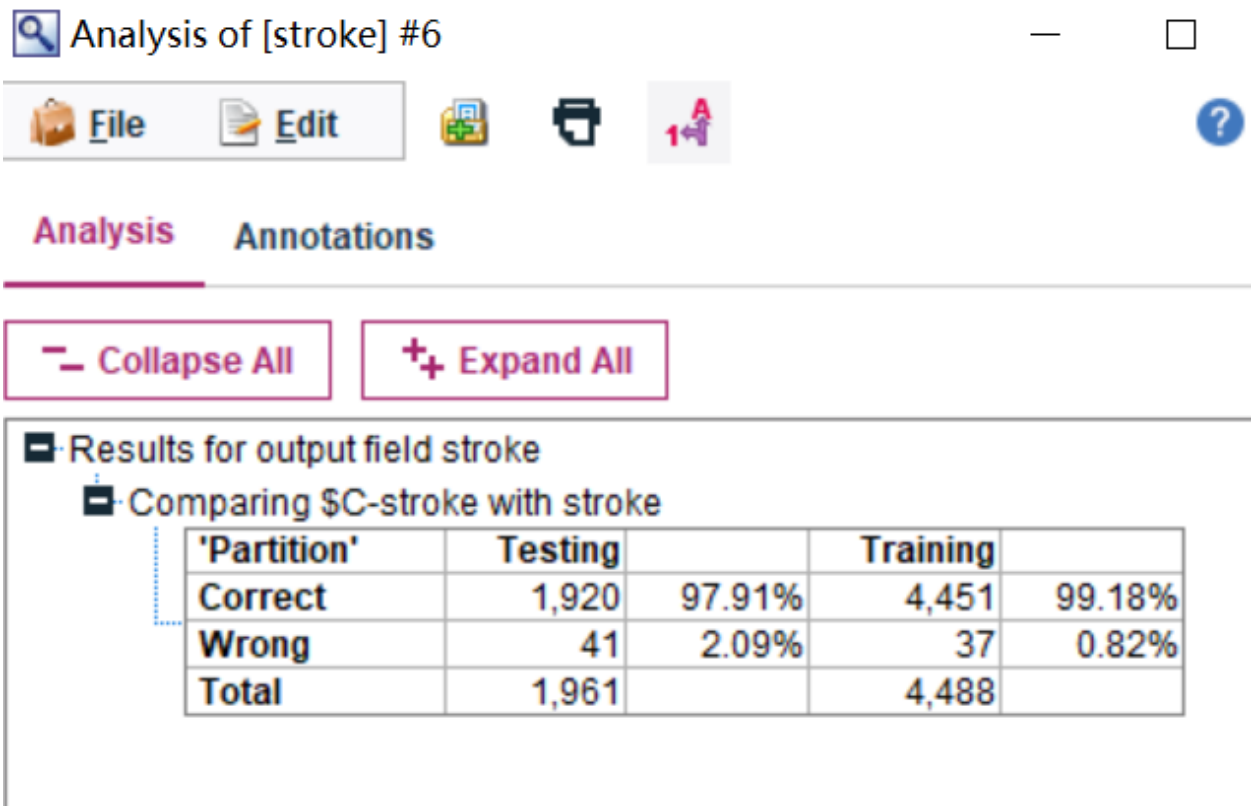


Figure 85. Rule sets generated by C5.0. model

Figure 86 is showing the performance evaluation of C5.0. model.



'Partition'	Testing		Training	
Correct	1,920	97.91%	4,451	99.18%
Wrong	41	2.09%	37	0.82%
Total	1,961		4,488	

Figure 86. Performance evaluation of C5.0. model

Figure 87 is showing the Relationship between age and stroke.

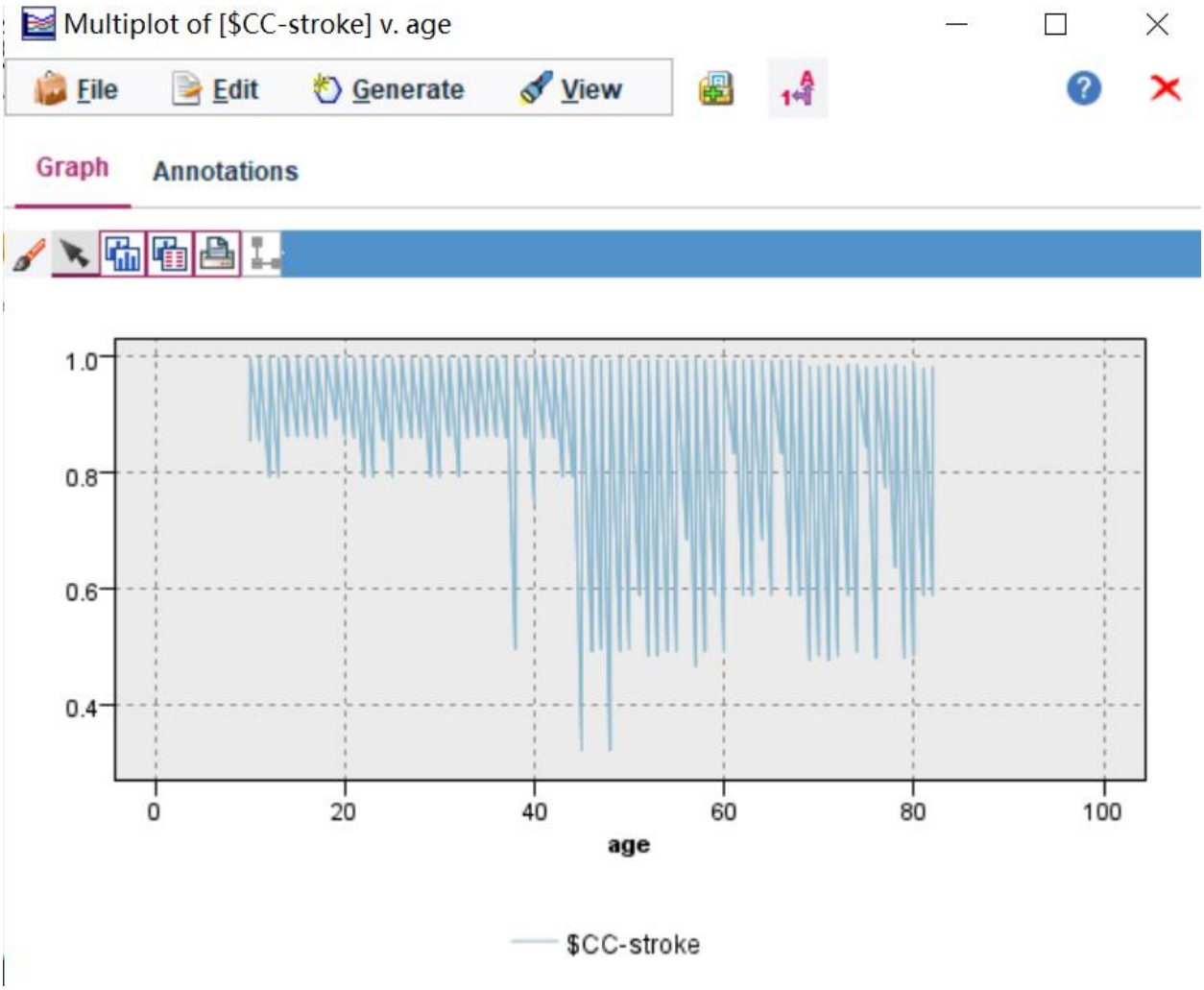


Figure 87. Relationship between age and stroke

Figure 88 is showing the Relationship between BMI group and stroke.

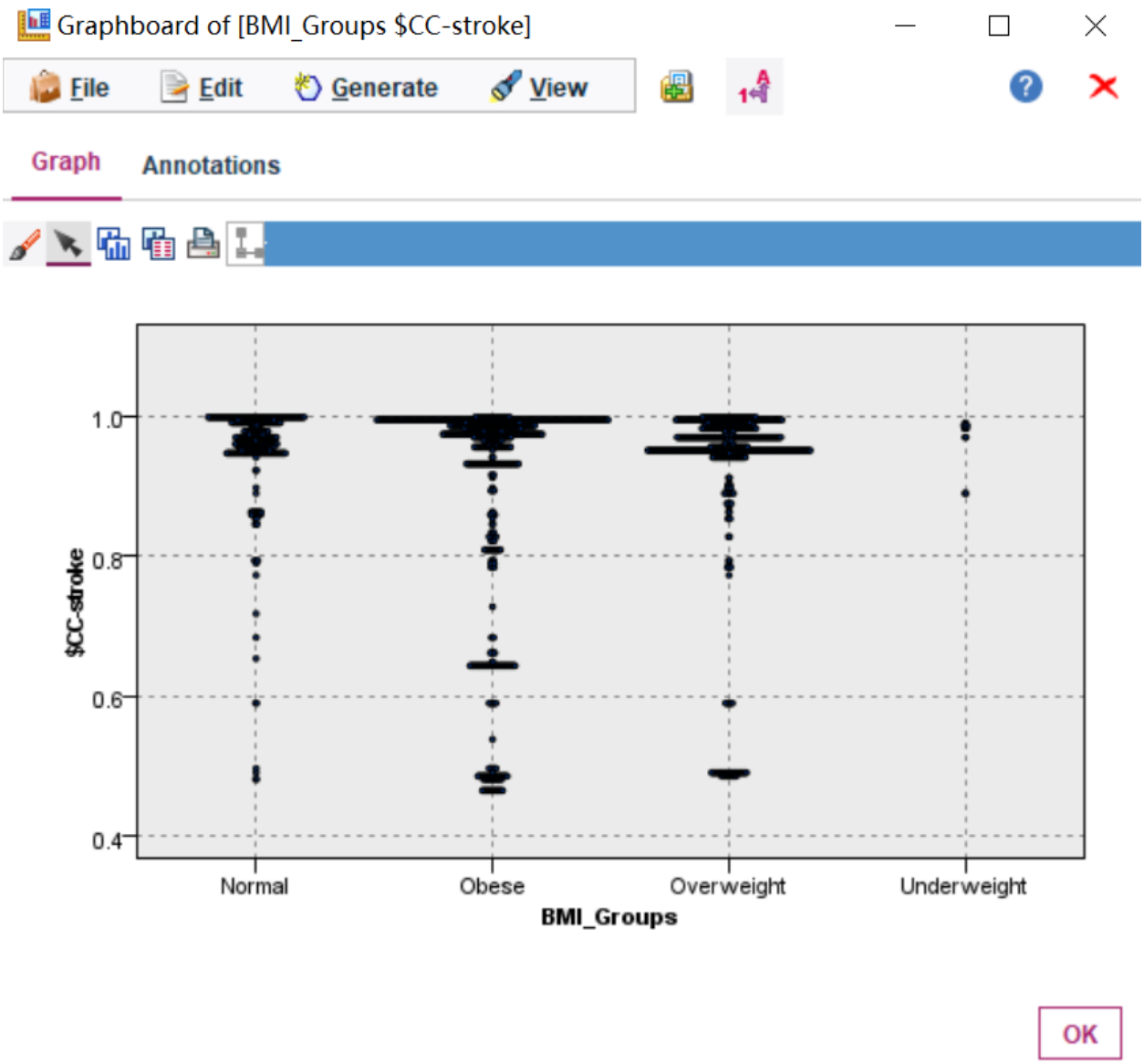


Figure 88. Relationship between BMI group and stroke

Figure 89 is showing the Relationship between average glucose level and stroke.

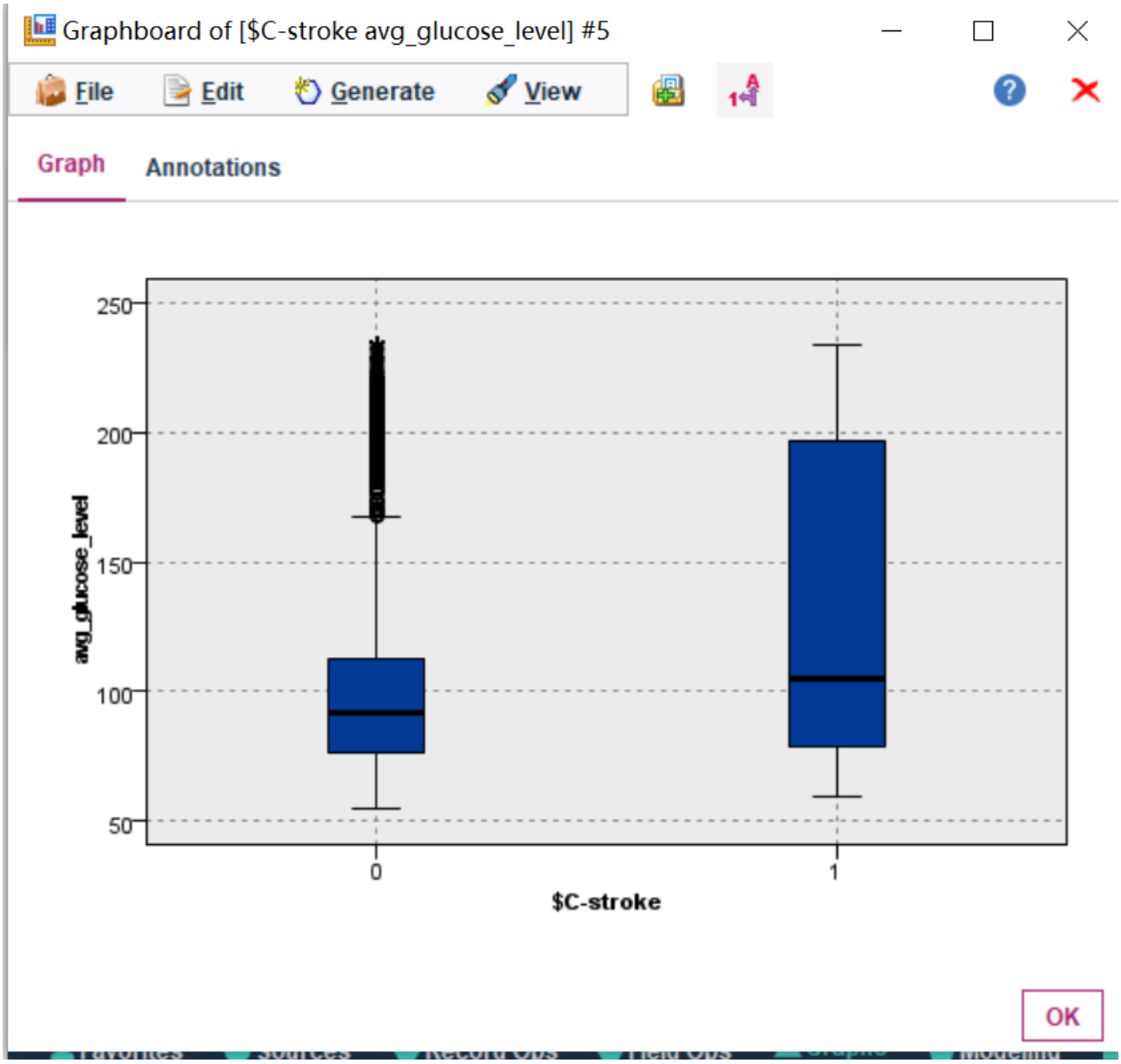


Figure 89. Relationship between average glucose level and stroke

Figure 90 is showing the Relationship between smoking status and stroke.

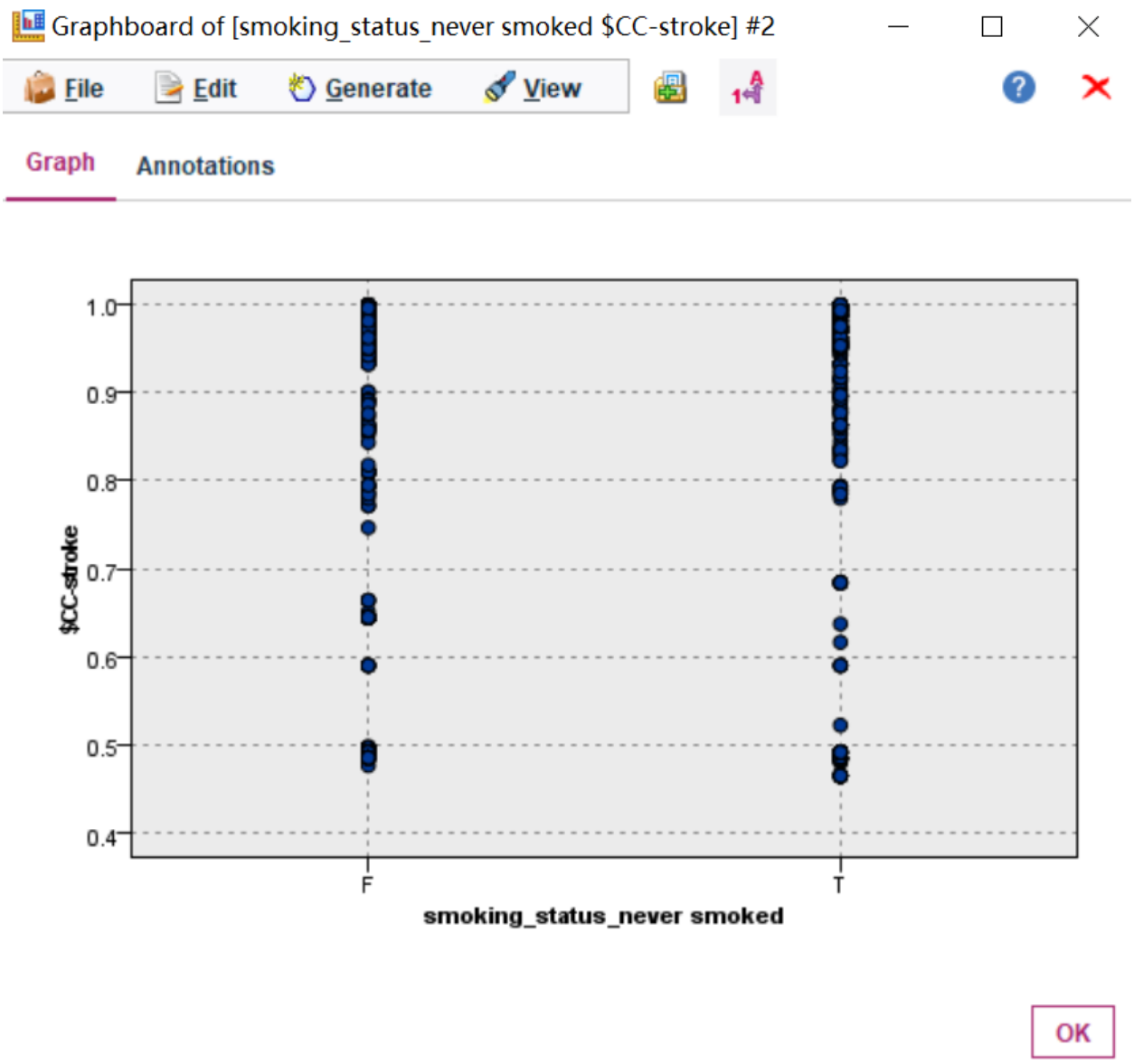


Figure 90. Relationship between smoking status and stroke

Figure 91 is showing the Relationship between marital status and stroke.

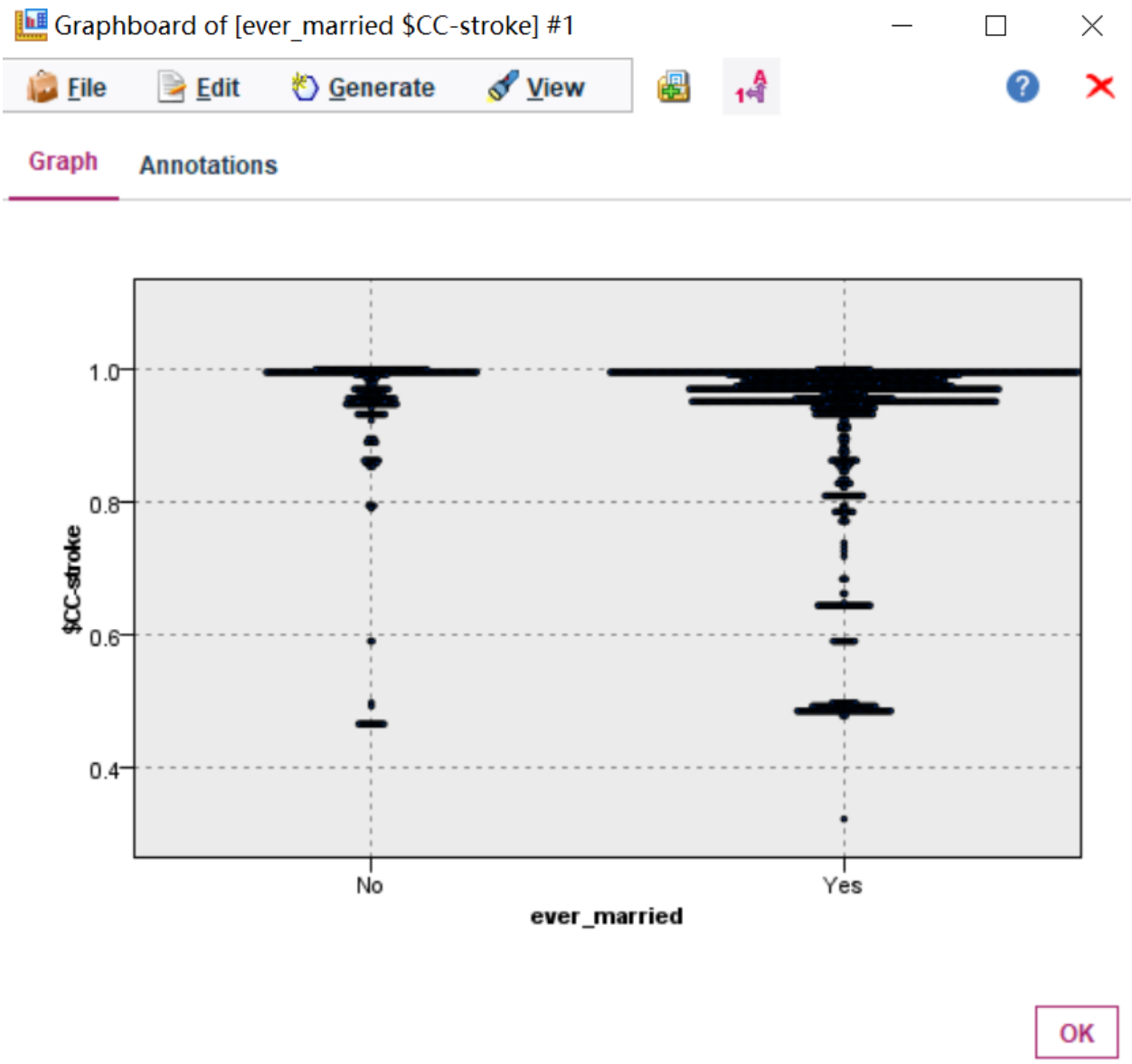


Figure 91. Relationship between marital status and stroke

Figure 92 is showing the Relationship between job category and stroke.

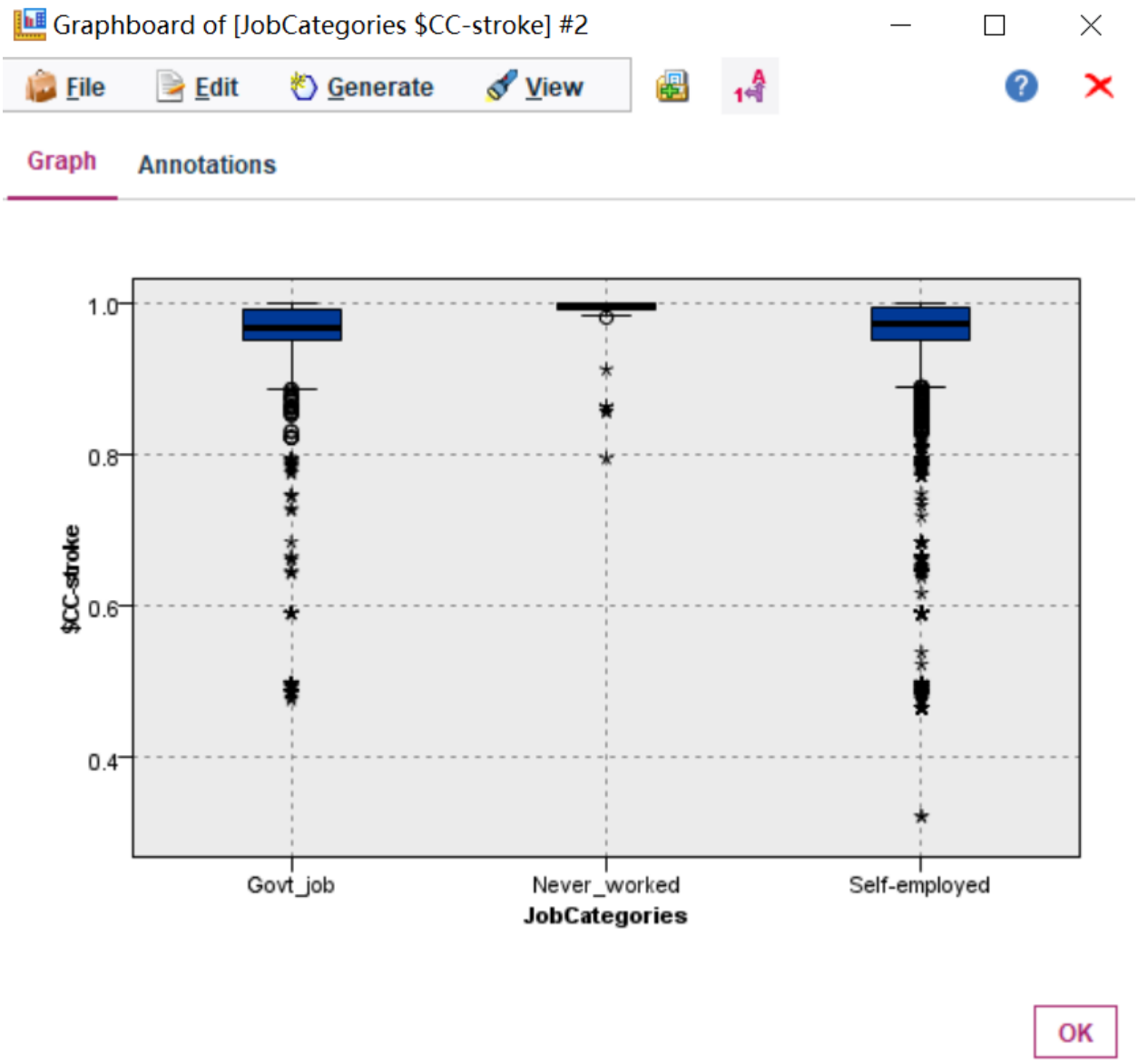


Figure 91. Relationship between job category and stroke

Figure 92 is showing the Relationship between hypertension and stroke.

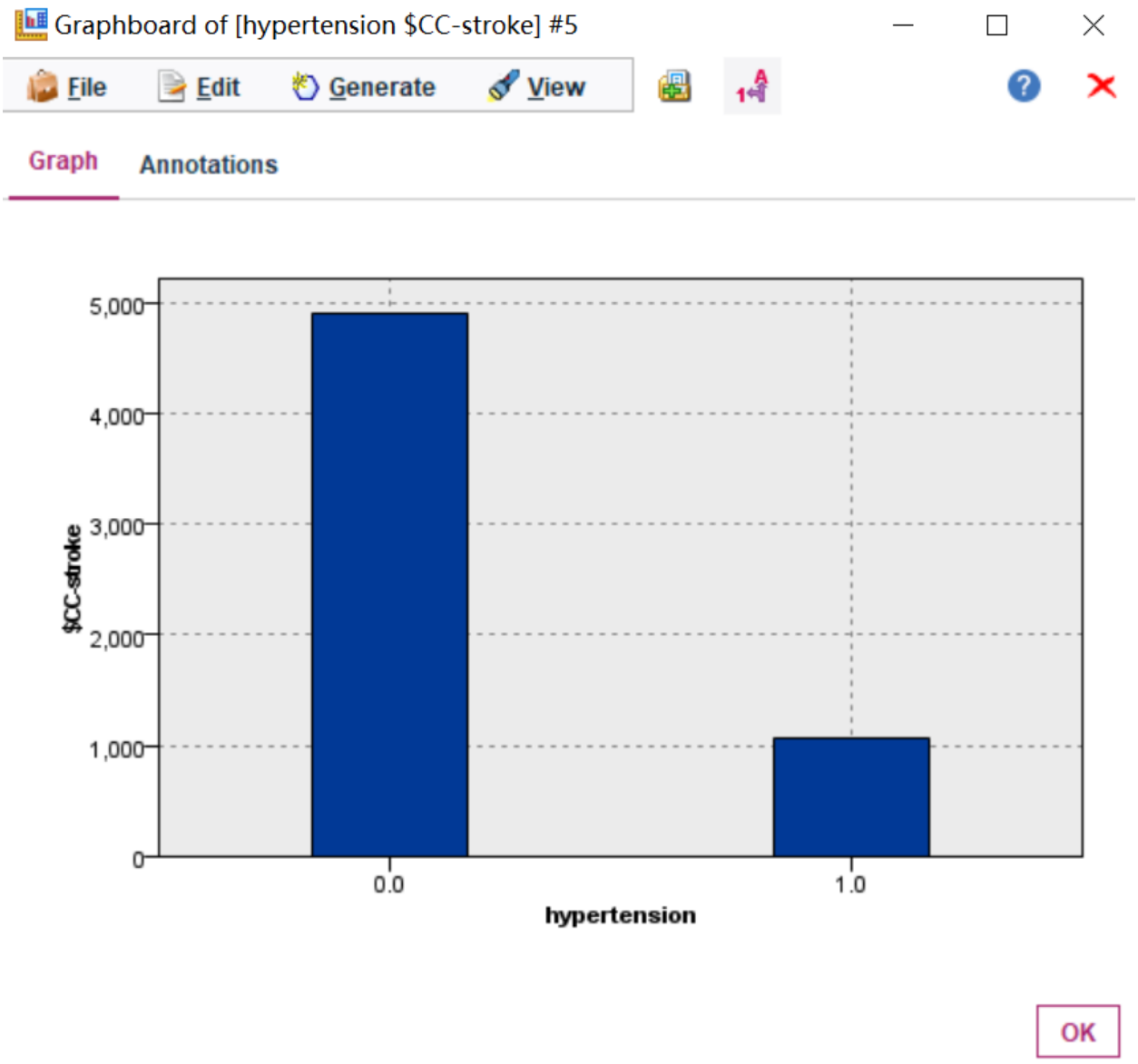


Figure 91. Relationship between hypertension and stroke

8.3 Interpretation of result, models and patterns

The patterns:

The patterns between age, BMI groups, average glucose level, smoking status, marital status, job categories and hypertension directly influence the chance of getting a stroke. Based on the deeper research, the average glucose level and marital status has been affected by age, which means those two attributes will increase accordingly.

Finally, age is being determined as the most influencing factor and most important factor to the stroke.

From the analysis of feature importance. The feature of age has the most significant for the decision. According to the research from (Margaret Kelly-Hayes,2010), over 70% of all strokes happen in those over the age of 65, and the risk rises with age, with the frequency doubling every ten years beyond the age of 45. About 145, 000 of the 795, 000 new or recurrent strokes predicted to occur in the United States each year will be fatal.

The second important feature for the decision is the BMI group. Obesity is an important factor in the development of coronary heart disease and also has a very important impact on cerebral infarction, as it can lead to hypertension, hyperlipidaemia and hyperglycaemia, which can accelerate cardiovascular disease and lead to stroke (Kernan et al., 2013).

The third important feature for the decision is average glucose level. From the research of (Yao, 2020), the blood arteries in the body can get damaged over time by high glucose levels, which raises the risk of stroke. The feature importance of the decision tree matches the clinical findings.

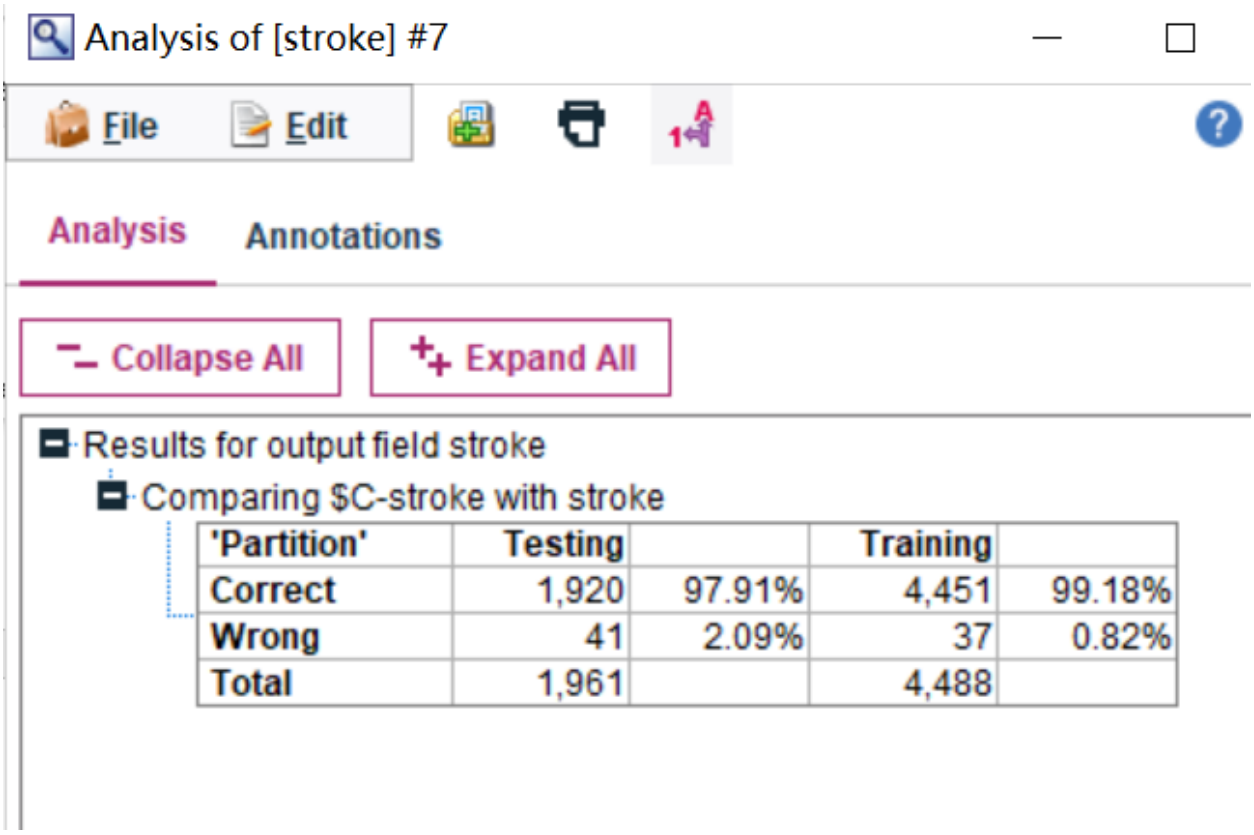
The fourth important feature for the decision is smoking status. From the point of view of smoking, after reviewing some literature, it can be concluded that smoking tends to lead to endothelial dysfunction of blood vessels, which can easily lead to the formation of blood clots and blockage of blood vessels, thus triggering stroke, and also increases the incidence of coronary heart disease, cerebrovascular disease and the vascular diseases surrounding them, accounting for more than 18.9% of all risk factors for stroke. (Centers for Disease Control and Prevention, 2021).

The fifth important feature for the decision is marital status. As we discussed before, marital status is changing with age growing. The person is more likely to get married with the growth of age. Therefore, the marriage status predictor is now being categorized as the ageing factor.

The sixth important feature for the decision is job category. From a work type perspective, factors such as smoking, alcohol consumption and obesity are not the only causes of stroke; factors such as stress and late nights also contribute to an increased risk of cerebral infarction.

The seventh important feature for the decision is hypertension. Hypertension is a trigger for strokes, as blood pressure increases in the body and vascular tone expands, causing damage to the vessel walls in the blood vessels. As a result of this injury, blood fat in the blood vessels tends to penetrate the lining and build up too much pressure, leading to atherosclerosis, and the accumulation of lipids blocking the blood vessels can lead to stroke (Alloubani et al., 2018). However, in the Figure 91 shows most patients who get stroke does not have hypertension, which is differs from the results of numerous studies. This might because this dataset is biased, which collect data from the place where most people do not have hypertension, as I know China is a region with a high prevalence of hypertension. Therefore, predictors of hypertension may require deeper data mining.

The models and results:



'Partition'	Testing		Training	
Correct	1,920	97.91%	4,451	99.18%
Wrong	41	2.09%	37	0.82%
Total	1,961		4,488	

Figure 92. C5.0 Model accuracy

The performance of the model can be evaluated by observing the accuracy when the model running the testing set. Figure 92 is stating the accuracy of the model is quite satisfying as the accuracy of prediction reaches 97.91% for the test set.

8.4 Assess and Evaluations of result, model and patterns

Evaluation of patterns:

The model finds patterns that provide relationships between attributes and goals, but the relationships themselves are not as clear. The model analyses the importance of other attributes to the target attribute, and the relationship between other attributes and the target attribute. However, the effect of attributes combining with each other on the target attribute stroke is not clear, for example age and hypertension are not assessed for the risk of stroke. This will therefore affect the accuracy of the results and conclusions. Potential solutions are to

add additional measurements to the current algorithm, or to use multiple algorithms to perform the task.

Performance evaluation of the model and results:

Based on the accuracy of the training and test sets, only a small number of people were incorrectly classified into the wrong category. All in all, the overall performance of the model was satisfactory as it was able to achieve prediction accuracies of over 95% for both the training and test sets. The next step could be to further improve the accuracy of the algorithm by adding more cases or using a larger database to do data mining, there or making a larger training dataset.

8.5 Iterations and Improving Models

We could use the method of re-split training and test data set to improve the model, for example, re-split the training model and test model by 7.5:2.5 or 8: 2 ratios. Also, data preprocessing is quite important for decision tree training. A suitable normalisation method could be used. In this dataset, a great number of smoking records are missing, a better way of dealing with the missing value should be explored, for example, we could set the smoke status of patients with an age smaller than 20 as non-smokers.

Iteration of the model will continue to improve the performance of the model and bring clarity to the data mining process. In order to improve the model, a review of the current steps is necessary.

Step 1: The first step in data mining is to understand the context of the dataset and to develop a business understanding of this dataset to have a grasp of the current situation. By understanding the dataset, the overall structure of this dataset will be clear, the business objectives will be clarified, and the expected results will be listed. In this step, the precise positioning of the business objectives is crucial.

Step 2: The second step in data mining. The initial data collection involves finding and identifying suitable datasets and importing them into SPSS. The overall characteristics and quality of the data will then be assessed and analyzed. In this step, an understanding of the structure and quality of the data set is an important step before data mining.

Step 3: The third step in data mining. This step involves data cleaning and improving the quality of the dataset. In this step all missing values will be removed and extreme this will also be processed to ensure the quality of the dataset. In addition, feature combination or integration will also take place in this step. The data cleaning and processing process will influence the final model selection and prediction discovery.

Step 4: The fourth step in data mining. In this step, imbalances in the data will be eliminated, especially in the target attributes. All unnecessary attributes with low relevance to the final result will be filtered out and removed and the unbalanced data will be rebalanced.

Step 5: The fifth step in data mining. In this step, we need to make a judgement based on the dataset and the expected results and determine the data mining method. The choice of data mining method will influence the choice of data mining algorithm. And the train and test set will be splinted before we do data mining.

Step 6: The sixth step in data mining. In this step, based on the training dataset and test dataset segmentation of the dataset, we perform model/algorithm selection and building. Data mining algorithms will be evaluated and selected based on business objectives and data mining goals.

Step 7: The seventh step in data mining. In this step, the data mining model/algorithm will be run, and the patterns are being identified.

Step 8: The eighth step of data mining. The results of the data mining, models and patterns are evaluated, and the results are analyzed for validity and correctness. Attempts are made to improve the accuracy of the model, and the most convenient and easy updates can occur in the training/test data split. As the number of training sets increases, the efficiency of the algorithm increases. The new data splitting ratio was set to 8:2 as in Figure 93.

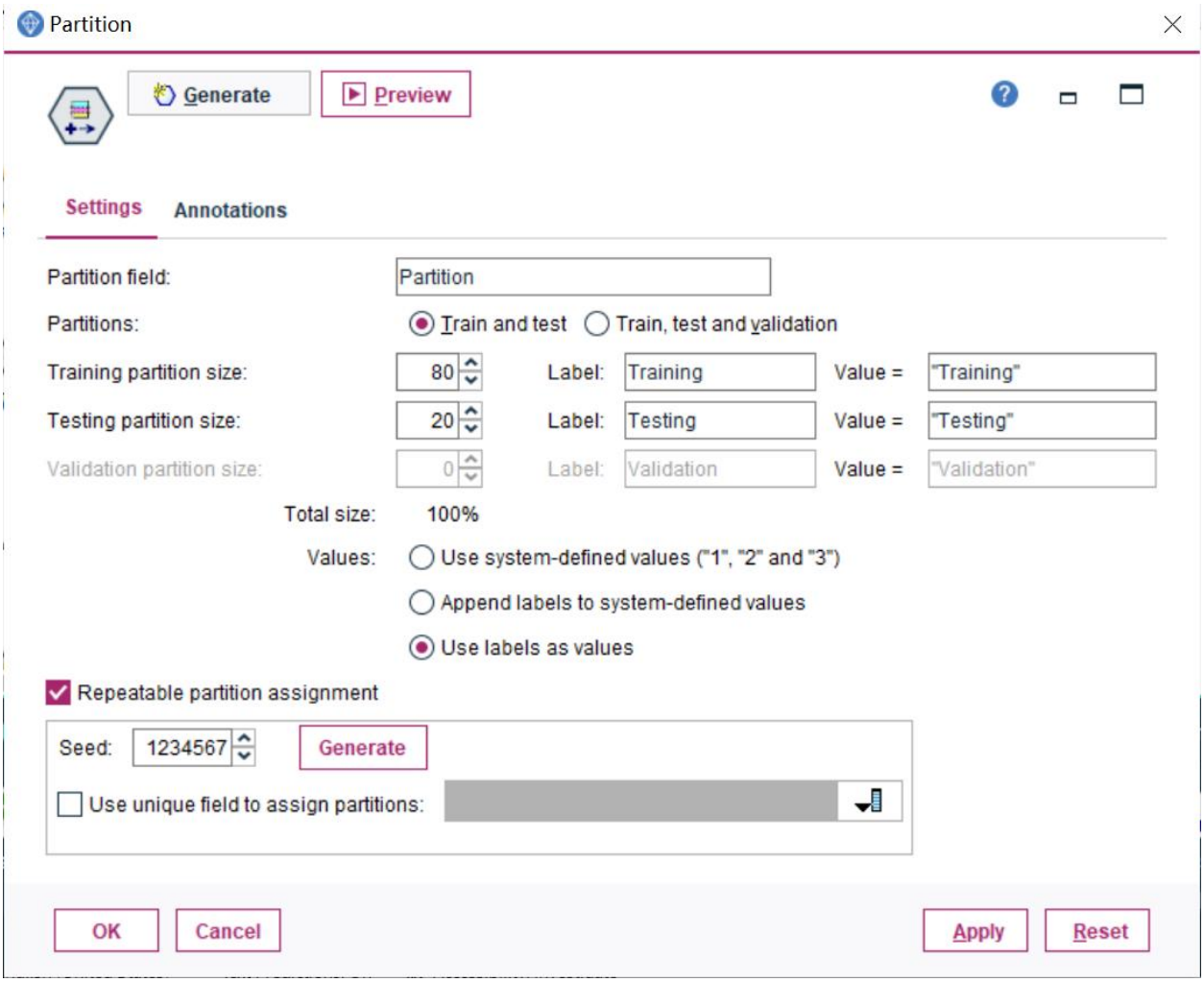


Figure 93. New data splitting 8:2

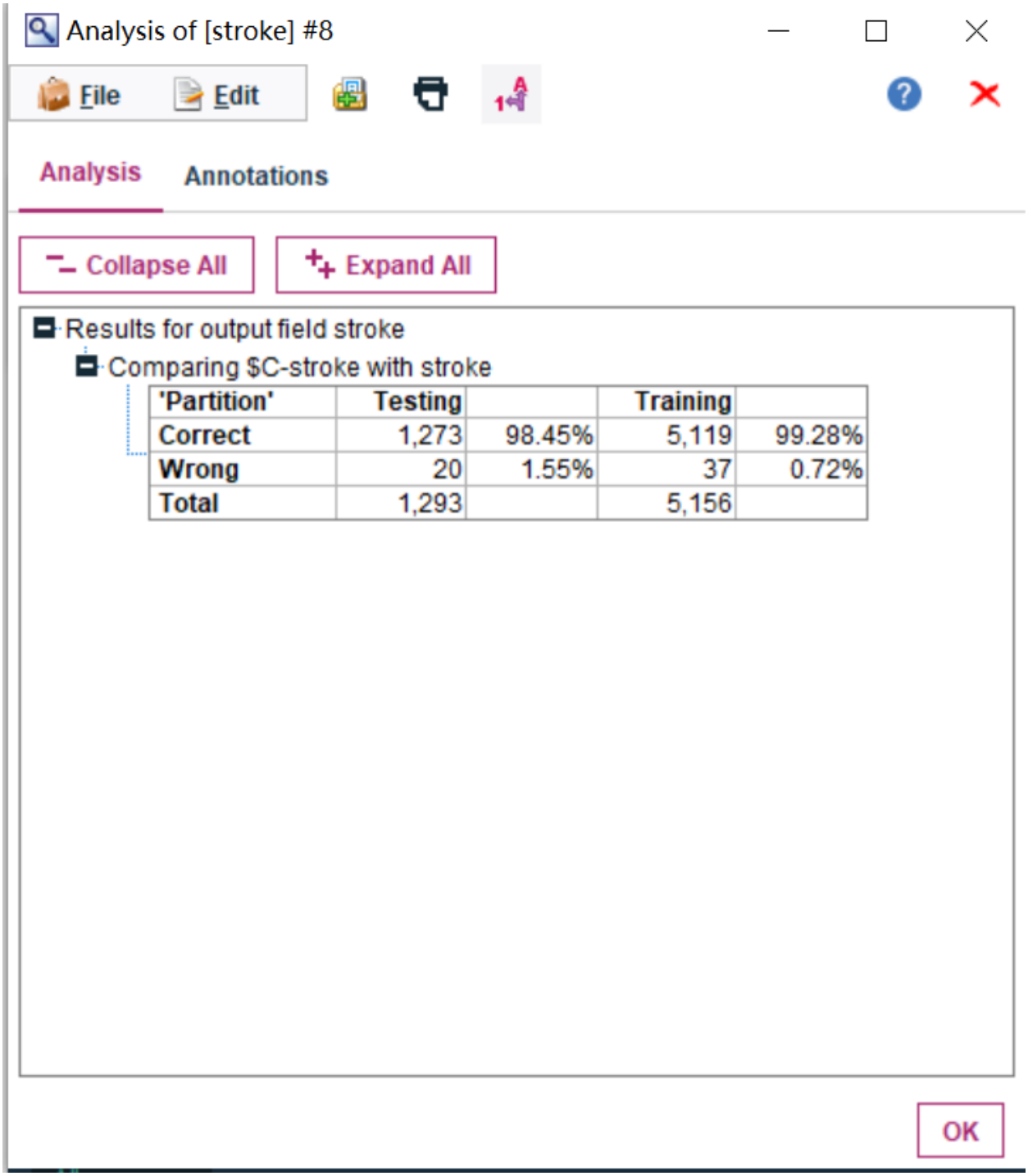


Figure 94. The accuracy report after new data splitting

The application of the new data proportions resulted in an increase in overall training accuracy and in the accuracy of the test set. It can be seen that more training data will give the model higher accuracy and also validate the stability of the C5.0. model.

9. Reference

2021 Guideline for the Prevention of Stroke in Patients With Stroke and Transient Ischemic Attack: A Guideline From the American Heart Association/American Stroke Association. *Stroke* 2021;May 24:[Epub ahead of print]. https://www.ahajournals.org/doi/10.1161/STR.0000000000000375.

Alloubani, A., Saleh, A., & Abdelhafiz, I. (2018). Hypertension and diabetes mellitus as a predictive risk factors for stroke. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 12*(4), 577-584.

Barnett, H. J. (2005). Stroke by Cause. *Stroke, 36* (12), 2523-2525. doi: 10.1161/01.STR.0000194560.65809.47.

Berry, M. W., Mohamed, A., & Yap, B. W. (Eds.). (2019). *Supervised and unsupervised learning for data science*. Springer Nature.

Body mass index (BMI) calculator. DiabetesCanadaWebsite. (n.d.). Retrieved August 21, 2022, from https://www.diabetes.ca/managing-my-diabetes/tools---resources/body-mass-index-(bmi)-calculator#:~:text=Body%20Mass%20Index%20is%20a,most%20adults%2018%2D65%20years.

Chong, J. Y. (2022, August 4). *Overview of stroke - brain, spinal cord, and nerve disorders.* MSD Manual Consumer Version. Retrieved August 15, 2022, from https://www.msdmanuals.com/home/brain,-spinal-cord,-and-nerve-disorders/stroke-cva/overview-of-stroke

Cleveland Clinic medical professional. (2018, February 21). *Blood glucose test: Levels & What They mean*. Cleveland Clinic. Retrieved September 19, 2022, from https://my.clevelandclinic.org/health/diagnostics/12363-blood-glucose-test#:~:text=A%20blood%20glucose%20test%20is,indicate%20pre%2Ddiabetes%20or%20diabetes.

Data mining methods: Top 8 types of data mining method with examples. EDUCBA. (2021, October 16). Retrieved August 21, 2022, from https://www.educba.com/data-mining-methods/

Editor, P. (2021, November 17). *Data Mining Techniques: Top 5 to consider*. Precisely. Retrieved August 21, 2022, from https://www.precisely.com/blog/datagovernance/top-5-data-mining-techniques

Harris, M. I., Hadden, W. C., Knowler, W. C., & Bennett, P. H. (1987). Prevalence of diabetes and impaired glucose tolerance and plasma glucose levels in US population aged 20–74 yr. *Diabetes*, *36*(4), 523-534.

Kelly-Hayes M. (2010). Influence of age and health behaviors on stroke risk: lessons from longitudinal studies. Journal of the American Geriatrics Society, 58 Suppl 2(Suppl 2), S325–S328. https://doi.org/10.1111/j.1532-5415.2010.02915.x

Kernan, W. N., Inzucchi, S. E., Sawan, C., Macko, R. F., & Furie, K. L. (2013). Obesity: a stubbornly obvious target for stroke prevention. *Stroke, 44*(1), 278-286.

U.S. Department of Health & Human Services. (2020, September 17). About Adult BMI. Retrieved from Centers for Disease Control and Prevention: https://www.cdc.gov/healthyweight/assessing/bmi/adult bmi/index.html

Wajngarten, M., & Silva, G. S. (2019). Hypertension and stroke: Update on treatment. *European Cardiology Review, 14(*2), 111–115. https://doi.org/10.15420/ecr.2019.11.1

WHO. (2020, December 9). *The top 10 causes of death*. World Health Organization. Retrieved July 29, 2022, from https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death.

Yao, T., Zhan, Y., Shen, J., Xu, L., Peng, B., Cui, Q., & Liu, Z. (2020). Association between fasting blood glucose and outcomes and mortality in acute ischaemic stroke patients with diabetes mellitus: a retrospective observational study in Wuhan, China. BMJ open, 10(6), e037291. https://doi.org/10.1136/bmjopen-2020-037291

Young, N., & Yousufuddin, M. (2019). Aging and ischemic stroke. *Aging* (Albany NY), 2542–2544

10. Disclaimer

"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.html, Links to an external site.).

I also acknowledge that I have appropriate permission to use the data that I have utilized in this project. (For example, if the data belongs to an organization 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."