Tutorial of Steps for a PREDICTIVE ANALYSIS

Jinhee Kim

Predictive Analytics is an advanced analytics that makes predictions about future events. It uses many techniques including statistics, modeling, machine learning, artificial intelligence and more. By successfully identifying the relationships among many factors and interpreting data, business (or individual) is effectively able to assess risks and opportunities and be preemptive for the future.

Predictive analysis can be done much more efficiently when it takes step by step and iterative operations followed by CRISP-DM. Business understanding is the first phase. Some of the things you want to achieve at this stage include:

- Find the questions you want to answer
- Include stakeholders in the discussion from the very beginning
- Define analytical goals
- Make it clear about expectation of success
- Communicate how others should help for analytical projects
- Define the target variables to make sense to all stakeholders
- Create project plan with timeline and milestones

Business Understanding

In this tutorial, a health related topic has been chosen, and the goals have been narrowed down to

Discover analytical insights of healthy life and longevity:

(1) discover key health indicators to improve longevity

(2) predict life expectancy and healthier life

Data Understanding

Secondly we need to understand the data. The below is what needs to be considered during this phase.

- Learn existing data sources, practical problems
- Write how we will extract and assemble the data
- Assess if data is suitable for the intended outcome
- Enhance the data with internal data and external data
- Run descriptive statistics for all variables (mean, medians, standard deviations, etc.)
- Learn continuous and categorial variables
- Define target variables
- How to handle outliers
- Graph data distributions
- Convert data distribution to a form a normal curve (logistic regression)
- Calculate correlation coefficients
- How to handle missing data
- How to handle sample in larger data population, experimental bias, measurement bias, intentional bias
- Determine if data set samples should be performed prior to analysis (reduce data volume, use that into resampling or cross validation, undersampling or oversampling?)

- Data set has been obtained at https://catalog.data.gov/dataset/community-health-status-indicators-chsi-to-combat-obesity-heart-disease-and-cancer
- The data set has been relatively in a good shape and is composed of 7 csv files with data explanation.
- Interesting information has been detected many areas including rick factors, access to care and demographic information.

Data Understanding – important data description

When possible, it would be very helpful to obtain data dictionary or description.

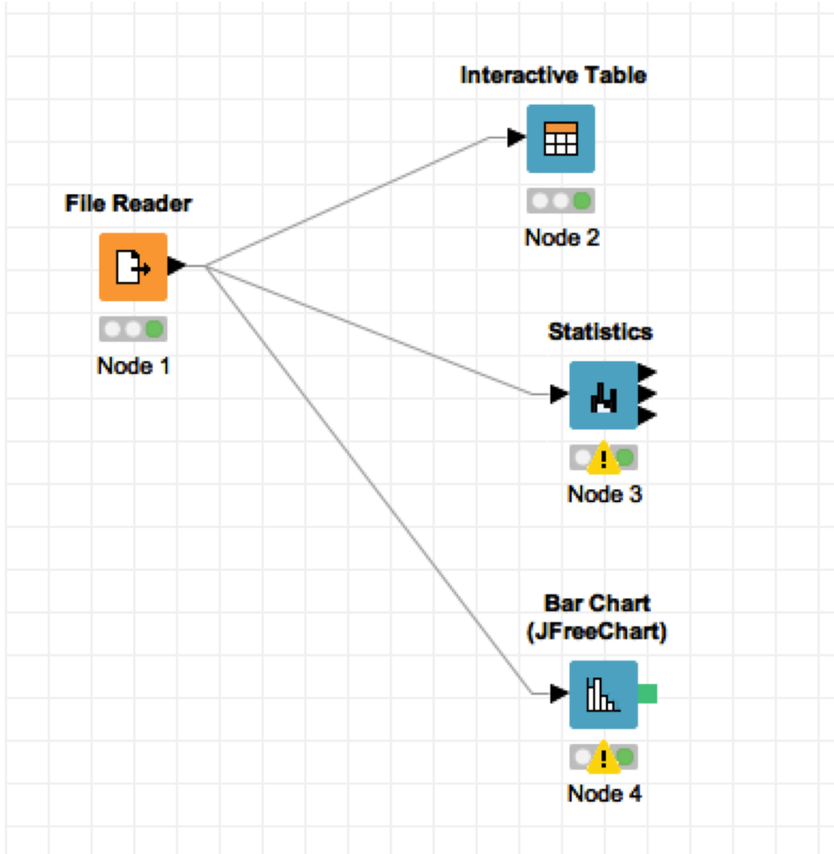
COLUMN_NAME	DATA_TYPE	IS_PERCENT_DATA	DESCRIPTION
ALE	Decimal	N	County data, average life expectancy
Health_Status	Decimal	Y	County data, self-rated health status
Unhealthy_Days	Decimal	N	County data, average number of unhealthy days in past month
No_Exercise	Decimal	Y	County data, no exercise
Few_Fruit_Veg	Decimal	Y	County data, few fruits/vegetables
Obesity	Decimal	Y	County data, obesity
High_Blood_Pres	Decimal	Y	County data, high blood pressure
Smoker	Decimal	Y	County data, smoker
Diabetes	Decimal	Y	County data, diabetes
Uninsured	Integer	N	County data, uninsured individuals
Elderly_Medicare	Integer	N	County data, medicare beneficiaries, elderly (age 65+)
Disabled_Medicare	Integer	N	County data, medicare beneficiaries, disabled
Prim_Care_Phys_Rate	Decimal	N	County data, primary care physicians per 100,000 pop.
Dentist_Rate	Decimal	N	County data, dentists per 100,000 pop.

Data Understanding

In order get good ideas about data, exploratory graphs and basic statistical information can be used. KNIME is an open source data mining program. Among KNIME nodes, Interactive table, bar chart, statistics are useful during the exploratory phase.

First the file has to be read, and then connect to interactive table, statistics and bar chart.

To move to the next step, key attributes should be identified. Basic statistics should be computed to find meaningful information. Exploratory graphics can be used to gain further insights to formulate hypotheses.



Data Understanding – interactive table

Interactive table shows that there are a lot of -1,111 which should be handled. It looks like this is an arbitrary number to fill up missing value or un-determined value.

File Hilite Navigation View Output

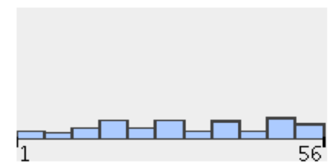
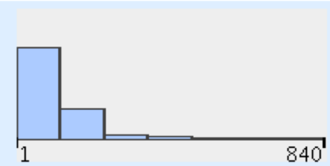
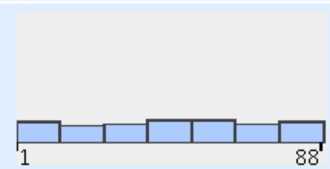
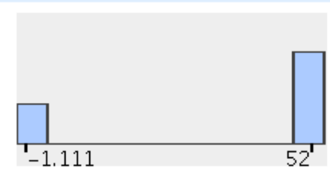
Row ID	State_...	Count...	CHSI_...	CHSI_...	CHSI_...	Strata...	No_Ex...	CI_Min...	CI_Ma...	Few_F...	CI_Min...	CI_Ma...
Row0	1	1	Autauga	Alabama	AL	29	27.8	20.7	34.9	78.6	69.4	87.8
Row1	1	3	Baldwin	Alabama	AL	16	27.2	23.2	31.2	76.2	71.2	81.3
Row2	1	5	Barbour	Alabama	AL	51	-1,111.1	-1,111.1	-1,111.1	-1,111.1	-1,111.1	-1,111.1
Row3	1	7	Bibb	Alabama	AL	42	-1,111.1	-1,111.1	-1,111.1	86.6	77.8	95.4
Row4	1	9	Blount	Alabama	AL	28	33.5	26.3	40.6	74.6	66.1	83
Row5	1	11	Bullock	Alabama	AL	75	-1,111.1	-1,111.1	-1,111.1	-1,111.1	-1,111.1	-1,111.1
Row6	1	13	Butler	Alabama	AL	76	24.5	15.5	33.5	-1,111.1	-1,111.1	-1,111.1
Row7	1	15	Calhoun	Alabama	AL	6	29.2	25.1	33.3	81.9	77.2	86.7
Row8	1	17	Chambers	Alabama	AL	50	34.7	25.3	44	84.6	75.4	93.7
Row9	1	19	Cherokee	Alabama	AL	64	-1,111.1	-1,111.1	-1,111.1	-1,111.1	-1,111.1	-1,111.1
Row10	1	21	Chilton	Alabama	AL	32	30.3	23.1	37.5	82.8	75.2	90.4
Row11	1	23	Choctaw	Alabama	AL	66	-1,111.1	-1,111.1	-1,111.1	-1,111.1	-1,111.1	-1,111.1
Row12	1	25	Clarke	Alabama	AL	51	31.5	22	41.1	-1,111.1	-1,111.1	-1,111.1
Row13	1	27	Clay	Alabama	AL	63	-1,111.1	-1,111.1	-1,111.1	-1,111.1	-1,111.1	-1,111.1
Row14	1	29	Cleburne	Alabama	AL	41	-1,111.1	-1,111.1	-1,111.1	-1,111.1	-1,111.1	-1,111.1
Row15	1	31	Coffee	Alabama	AL	32	23.3	17.2	29.4	-1,111.1	-1,111.1	-1,111.1
Row16	1	33	Colbert	Alabama	AL	21	30.2	23.3	37.2	76.9	66.8	86.9
Row17	1	35	Conecuh	Alabama	AL	75	-1,111.1	-1,111.1	-1,111.1	-1,111.1	-1,111.1	-1,111.1
Row18	1	37	Coosa	Alabama	AL	41	-1,111.1	-1,111.1	-1,111.1	-1,111.1	-1,111.1	-1,111.1
Row19	1	39	Covington	Alabama	AL	35	28.8	21.1	36.6	-1,111.1	-1,111.1	-1,111.1
Row20	1	41	Crenshaw	Alabama	AL	71	-1,111.1	-1,111.1	-1,111.1	-1,111.1	-1,111.1	-1,111.1
Row21	1	43	Cullman	Alabama	AL	21	29.4	23.9	34.9	76.2	69.4	83

Data Understanding - statistics

This basic statistics table shows how data has been distributed and many ideas of what to do to prepare the data including missing value, outliers, and many more to build models..

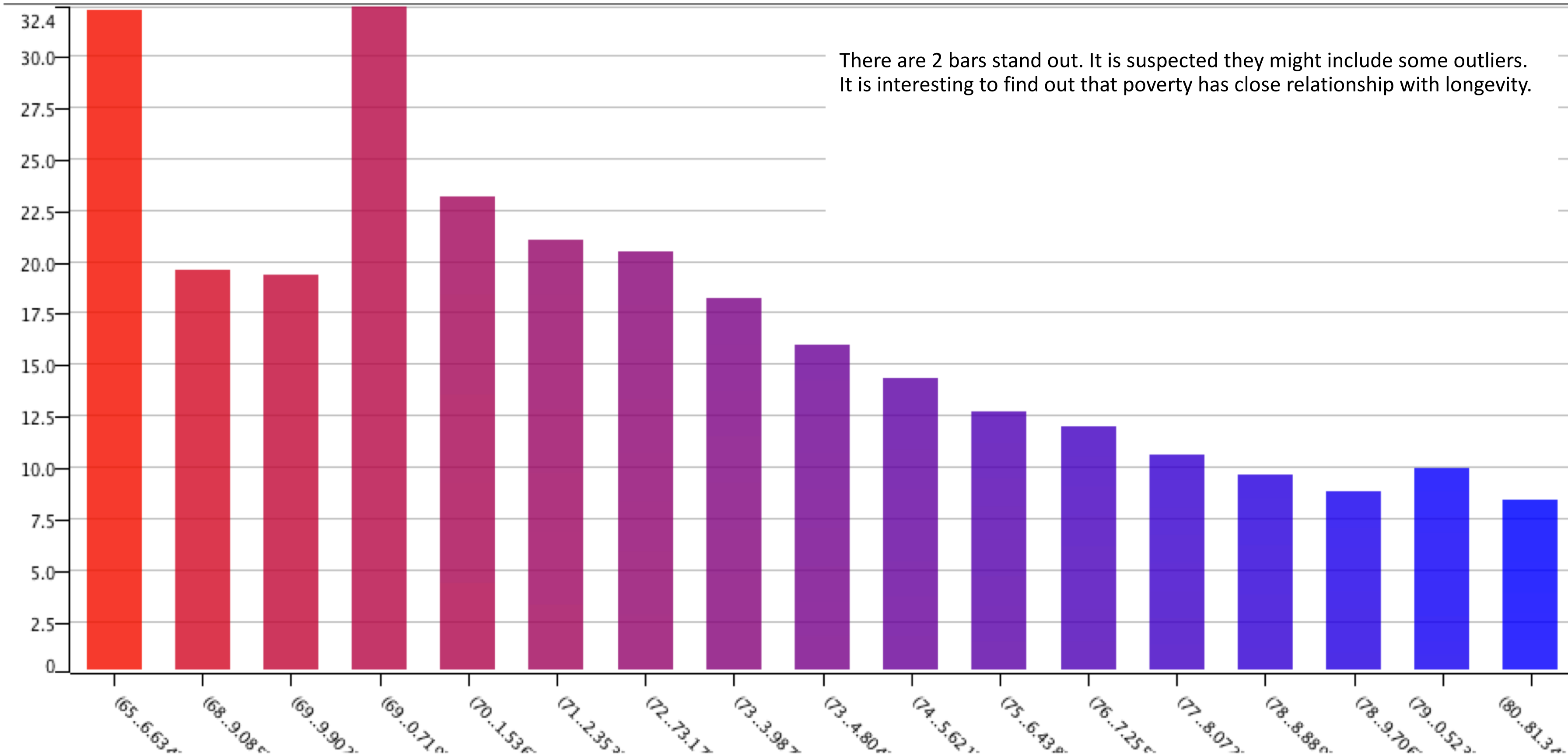
⚠ **Maximum number of unique possible values (1000) exceeds for column(s): "Uninsured","Elderly_Medicare","Disabled_Medicare","CHSI_County_Name"**

Numeric | Nominal | Top/bottom

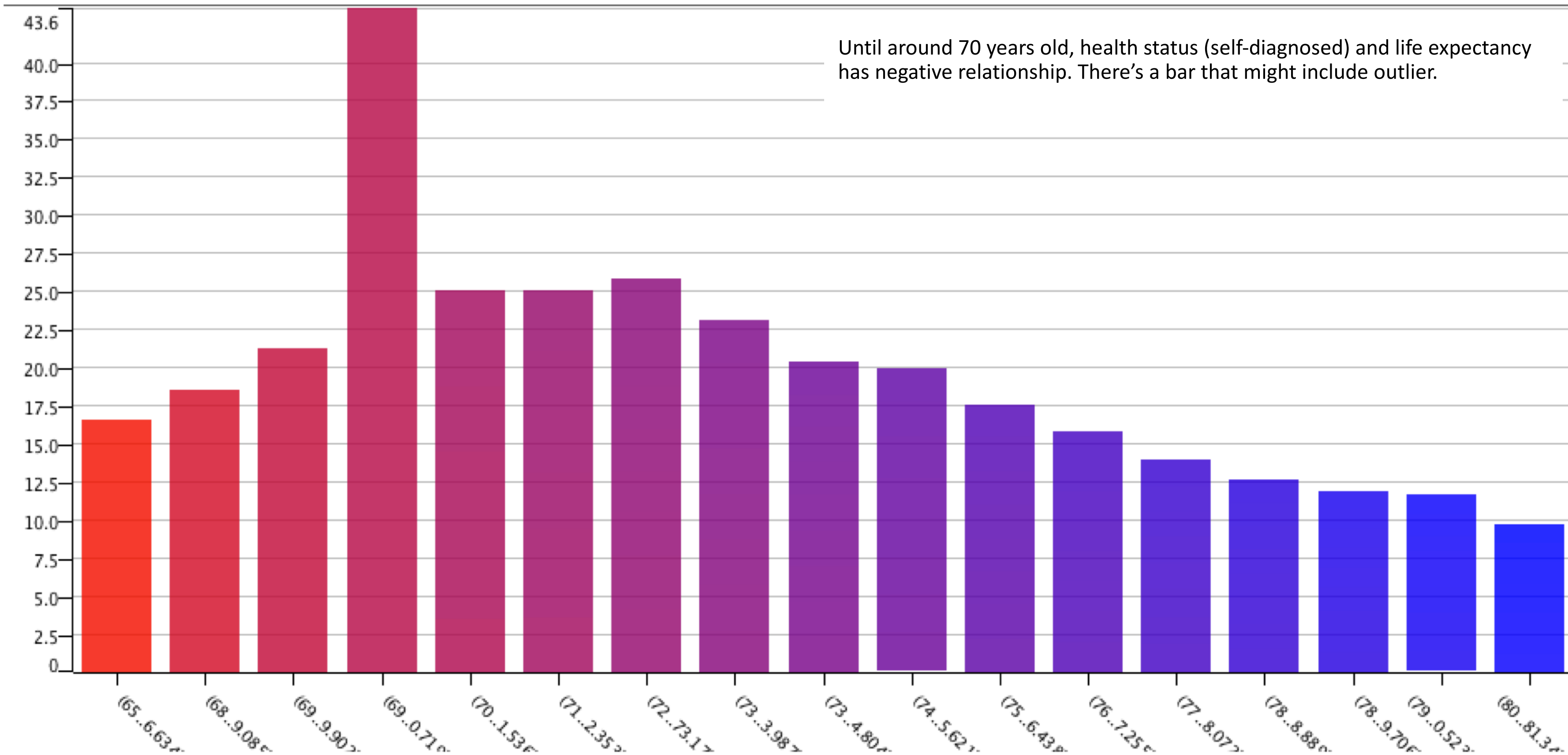
Column	Min	Mean	Median	Max	Std. Dev.	Skewness	Kurtosis	No. Missing	No. +∞	No. -∞	Histogram
State_FIPS_Code	1	30.3047	?	56	15.1344	-0.0818	-1.0986	0	0	0	
County_FIPS_Code	1	103.7167	?	840	107.9995	2.8325	11.2695	0	0	0	
Strata_ID_Number	1	44.6963	?	88	25.1184	-0.0226	-1.1614	0	0	0	
No_Exercise	-1,111.1	-312.1302	?	52.4	520.2688	-0.885	-1.2169	0	0	0	
CI_Min_No_Exercise	-1,111.1	-316.2394	?	43.6	517.5883	-0.8851	-1.2169	0	0	0	

Data Understanding – histogram (life expectancy and poverty)

There are 2 bars stand out. It is suspected they might include some outliers.
It is interesting to find out that poverty has close relationship with longevity.

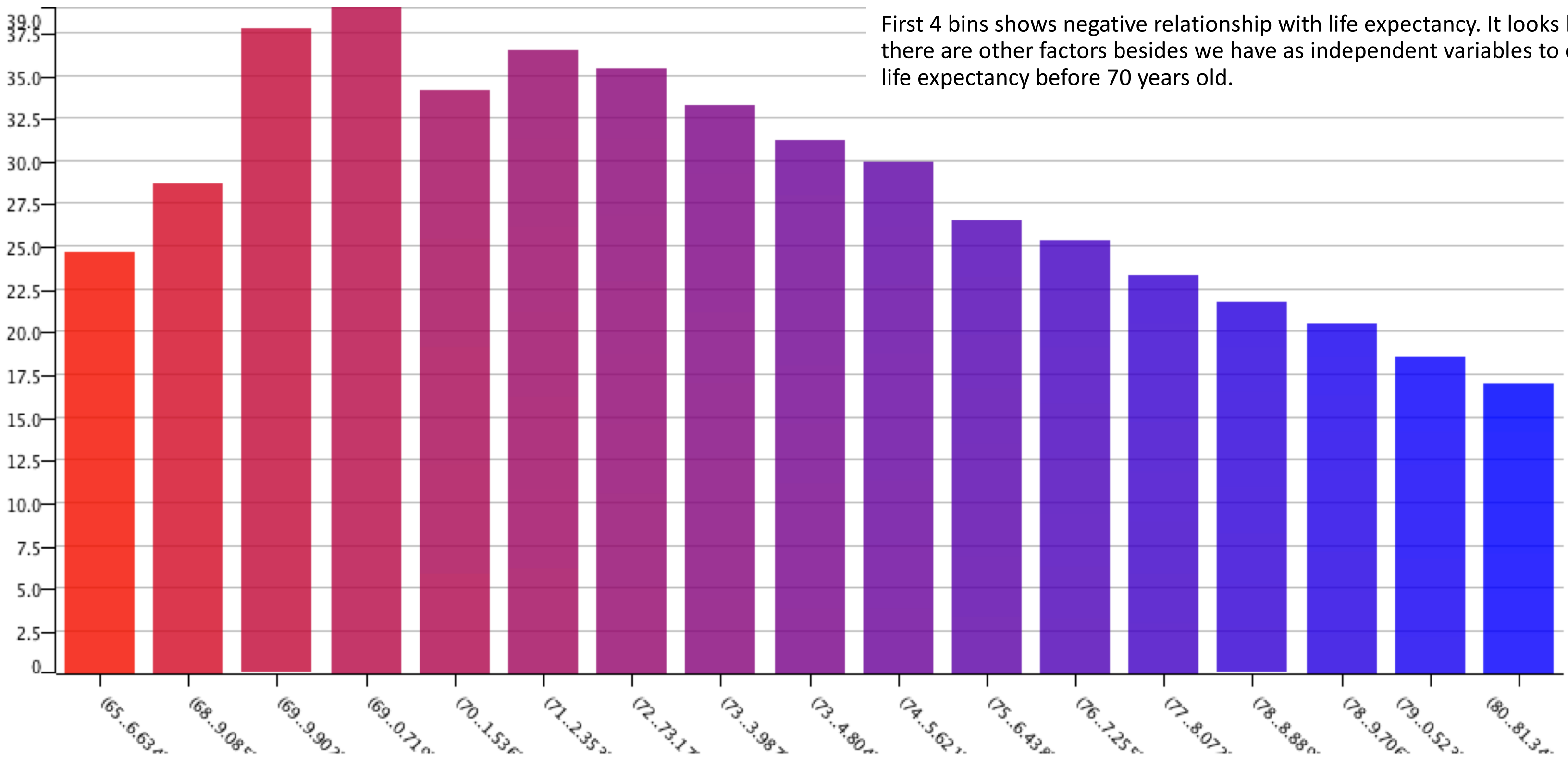


Data Understanding – histogram (life expectancy and health status)



Until around 70 years old, health status (self-diagnosed) and life expectancy has negative relationship. There's a bar that might include outlier.

Data Understanding – histogram (life expectancy and no exercise rate)



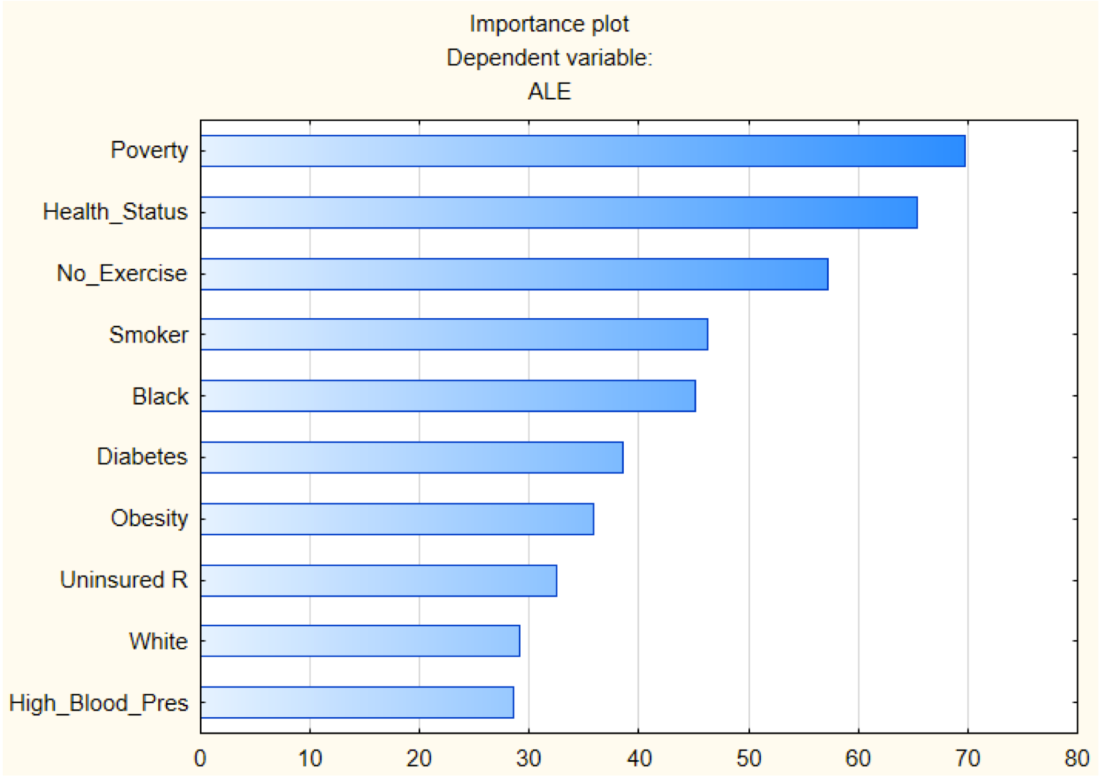
First 4 bins shows negative relationship with life expectancy. It looks like there are other factors besides we have as independent variables to cause life expectancy before 70 years old.

Feature Selection (1)

Feature selection is also known as variable, attribute, predictor selection. This is a selecting process to use them in predictive model construction. It is important to simplify the attributes in order to interpret the models easily, reduce processing time, enhance generalization, and reduce overfitting.

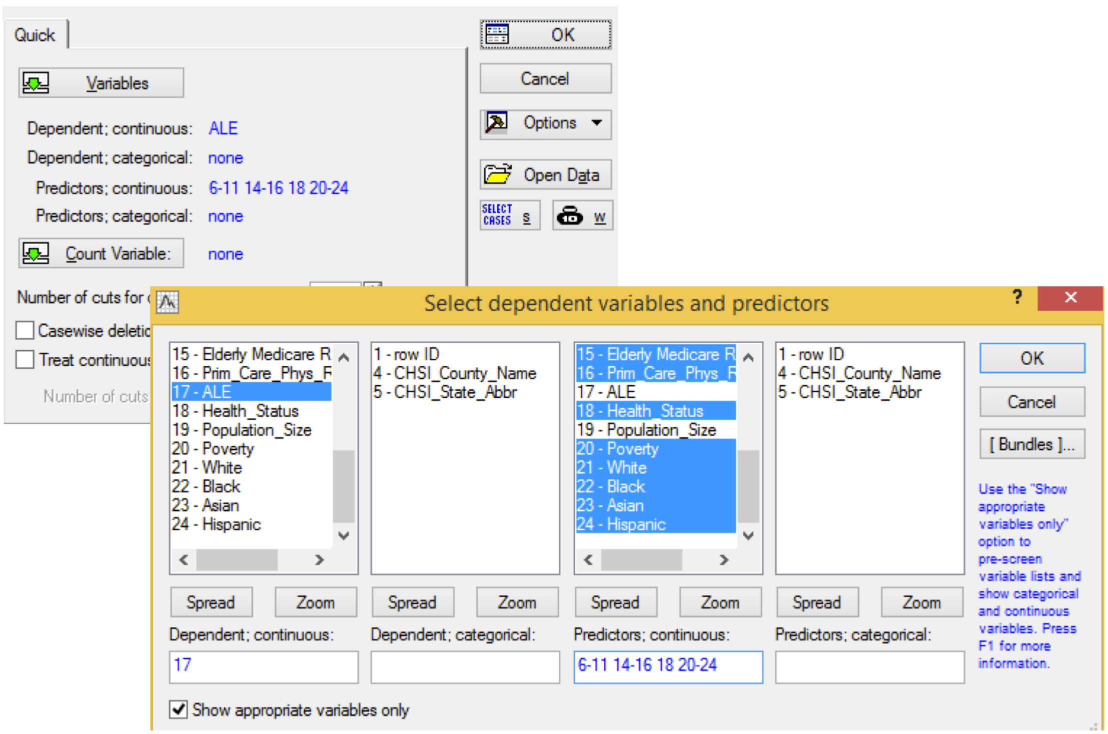
This analytical activities focus on ALE (Average life expectancy) as a main class attribute.

- Insights: It shows that poverty, health status (self diagnosed), exercise and smoking are strong indicators to factor average life expectancy followed by diabetes, diabetes, obesity. It would be not unsafe that what people think about their health status is more likely their actual health status. It is interesting to know that poverty is #1 factor and "Black" race is also one of important factors.



Feature Selection (2)

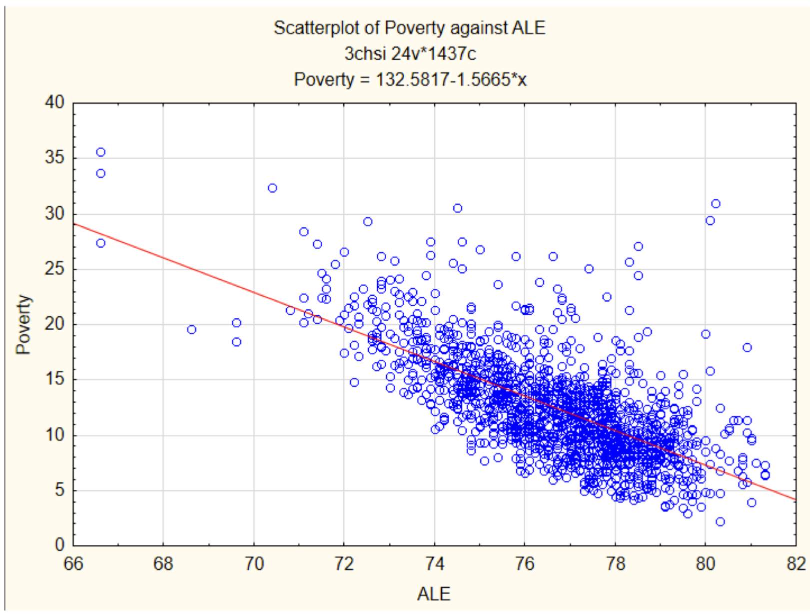
In order to get F-value to see how importantly variables are related to the dependent variable, you click ”Data Mining “ tab, then “Feature Selection”. After you specify the variables, you can click “summery” for F-value. When you click “histogram”, you can get importance plot.

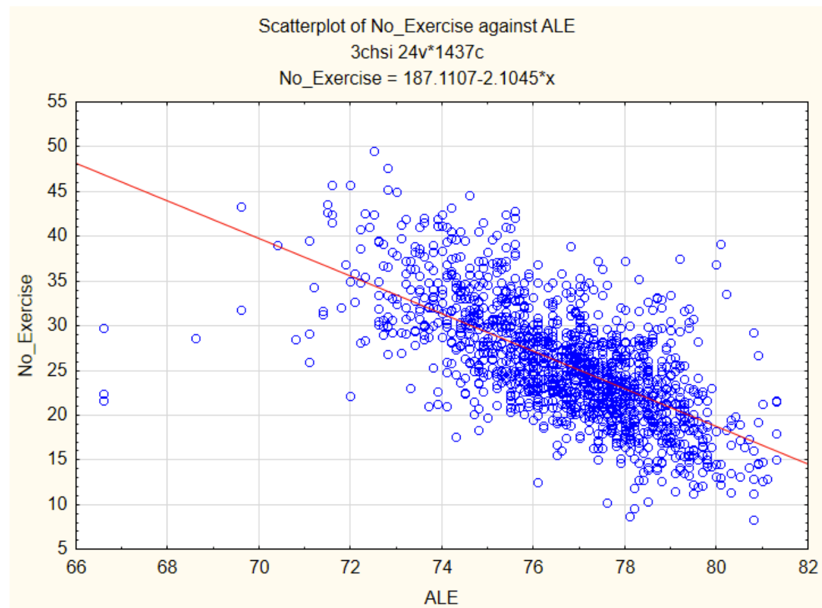


	F-value	p-value
Poverty	92.95342	0.000000
Health_Status	90.14998	0.000000
No_Exercise	76.70593	0.000000
Black	59.02300	0.000000
Smoker	58.16560	0.000000
Diabetes	50.69831	0.000000
Obesity	46.18917	0.000000
Uninsured R	42.71937	0.000000
High_Blood_Pres	38.37362	0.000000
White	32.94523	0.000000
Asian	18.06126	0.000000
Few_Fruit_Veg	11.56978	0.000000
Hispanic	7.21772	0.000000
Elderly Medicare R	6.06386	0.000000
Prim_Care_Phys_Rate	3.96806	0.000005

Strong Predicters

The Scatterplots between average life expectancy and poverty and No-exercise show negative linear relationship.

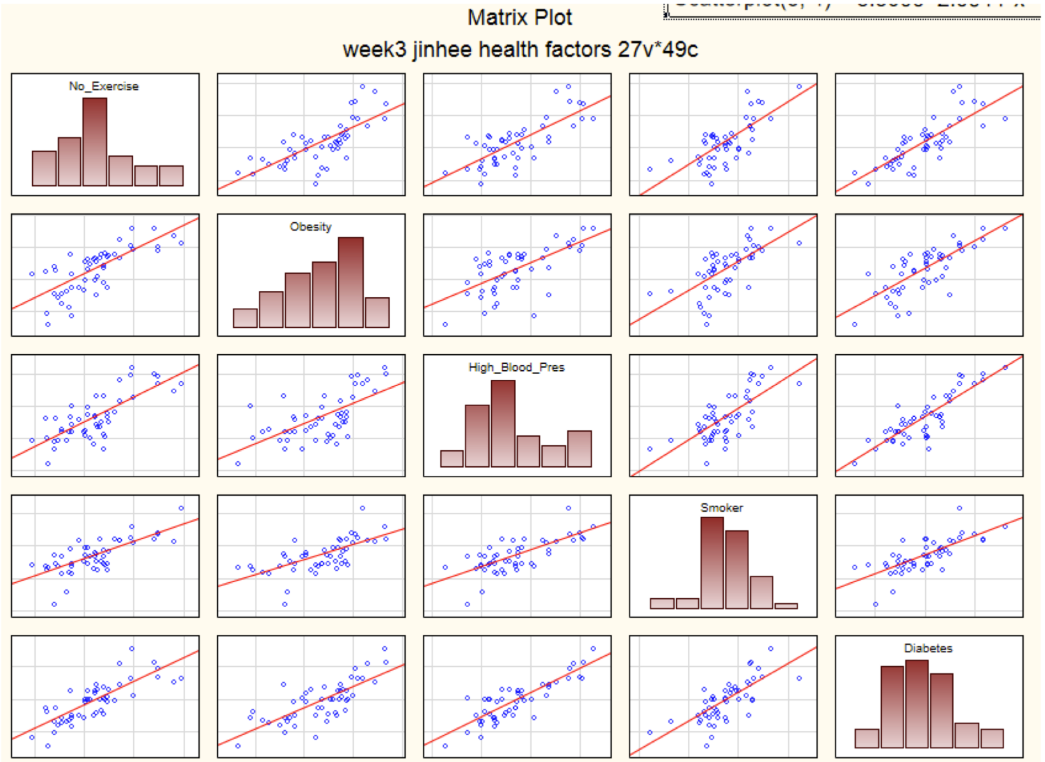




Correlations

Interesting findings between factors

- Almost perfect linear relationship between diabetes and high blood pressure
- Smoking related very strongly all to diabetes, high blood pressure, obesity, and even exercise. There is something unexpected discovery that smokers less likely spend time to exercise.
- All heath risk factors are showing positive linear relationship each other.



Data Preparation

During this phase, some of the following should be determined and managed.

- Sampling
    - Random sampling
    - Stratified sampling (more than 2 groups)
    - Oversampling and undersampling
    - Assign case weights or prior probabilities to specific target classes
- Cleaning
- Reduce variables to reduce complexity for models to work efficiently, and reduce noise
- Reduce numbers (neural network accepts categorial value only to numbers, it's better for decision tree as well)
- Clustering – reduce data volume
- Derive "dummy" variables from categorial variables
- Develop hierarchy generation
- Standardization – for statistical algorithms
- Recoding
- Filtering
- Missing value imputation
- Derived variables
- Summarize, calculate, make dummy variables
- Handle outliers
- Handle temporal data

For this predictive analysis, in order to construct the final dataset,

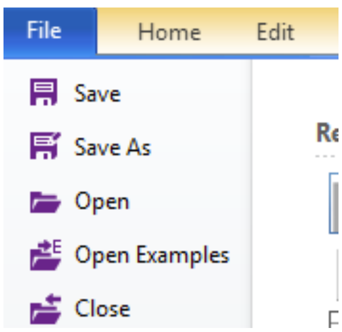
- Less valuable variables have been deleted
- 3 csv files have been consolidated and the data has been simplified from county level to state level for easy manipulation
- Missing data case has been handled
- Some data has been calculated (sum, average, etc.)

STATISTICA Data Miner Recipes (DMR)

This is a systematic process to build analytic models that relate dependent variables to independent variables. Usually when dependent variables are continuous, it is likely involved with regression models. When they are categorical, it creates classification models. DM Recipe is a step-by-step process starting with data analysis and ends with model evaluation. It includes various predictive models including neural networks, support vector machines, trees, and more.

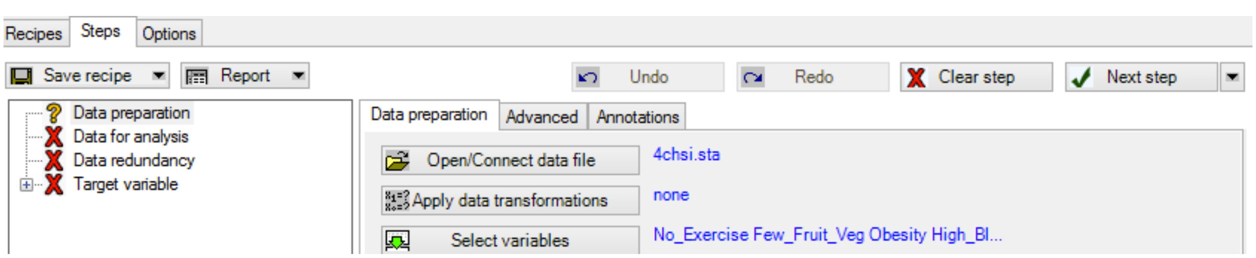
HOW TO START

1. Click "File" tab, "Open" source file. Before we start, we need to have "prepared" (i.e., cleaning, transformation) data ready in order to avoid any problems.

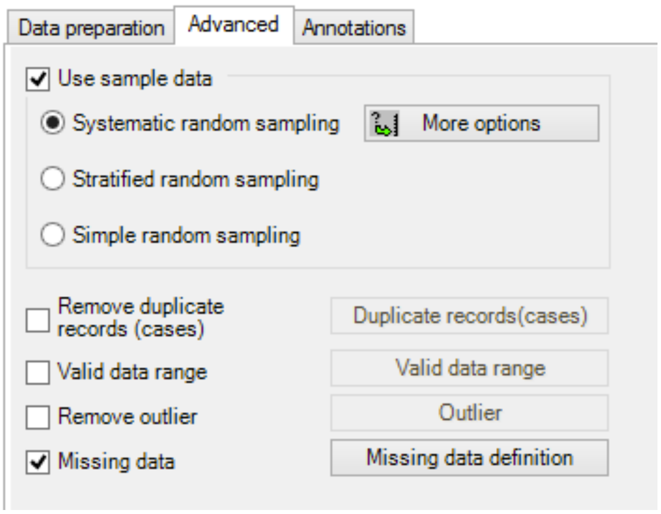


2. Click "Data Miner Recipes" from "Data Mining" tab. Then data mining dialogue will be opened.

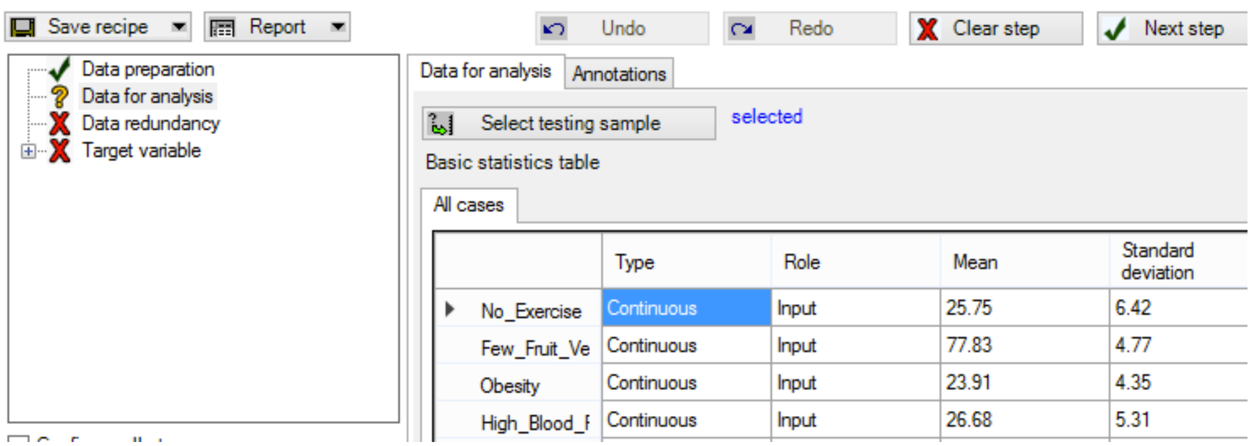
3. Once you select New or Open file, it starts from Data preparation. If you are here, it is very self-explanatory from here and basically you can just follow the following "Next step". First you can connect the data file. Just as you have done at Feature selection, variables should be specified.



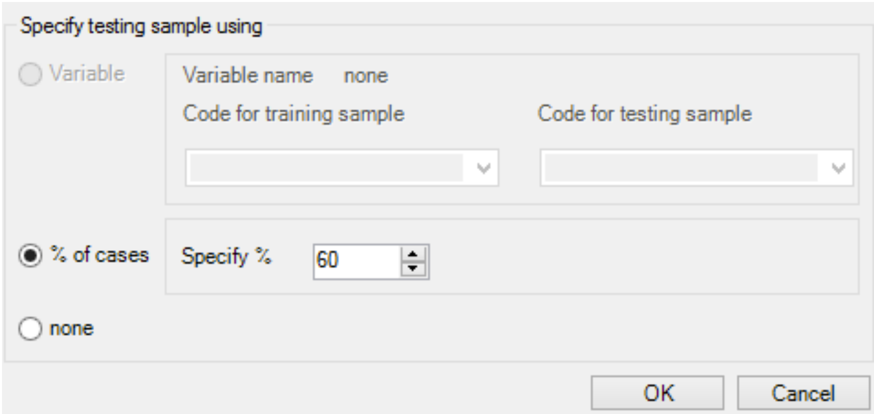
You have an option to choose "sample data" and specify how you want to do sampling in "Advanced" tab.



4. At Data for analysis, you get to see the basic statistic data and choose the training sample size.

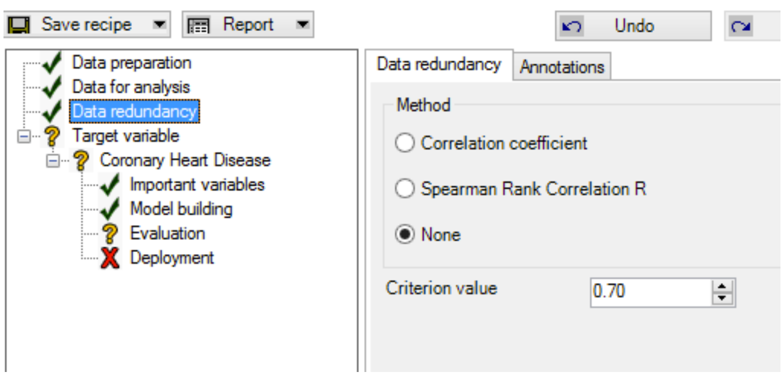


You are able to review statistical data such as mean, standard deviation, skewness etc. Also, you should specify the test sampling. This is not selected in default. But testing sample would be highly recommended to test the accuracy of trained models.



5. Data redundancy

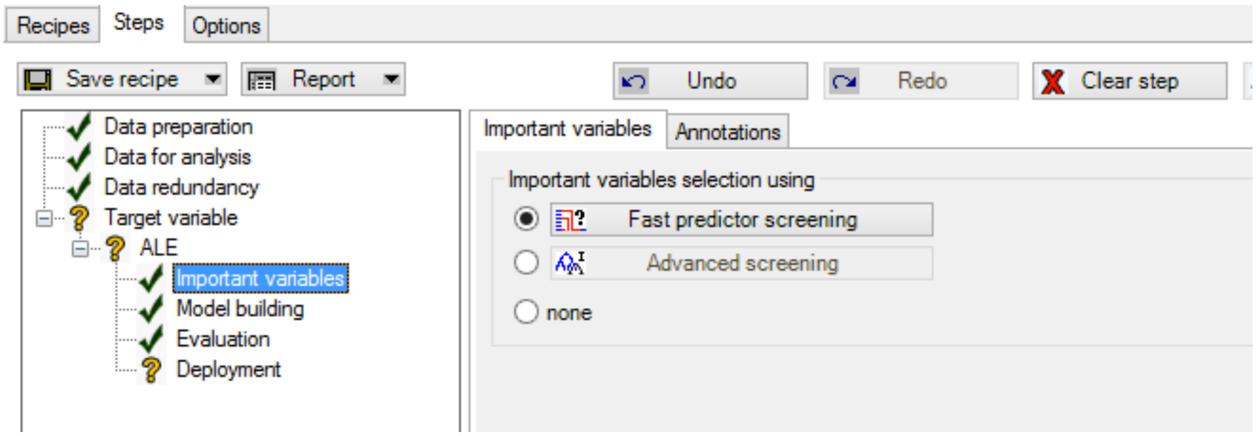
   Often, we may feel like the more the predictor variables we have, the better or more accurate predictable models we can build. Because all those variables would help to build models with better accuracy and less error. However, often when the dimensionality increases, the huge number of combinations of values grow exponentially. It becomes harder to support to the outcome of the models. Actually, simple variables can help to gain more than to lose some information from less variables as a result.



   In this case, "curse of dimensionality" has been considered during data preparation stage. We did not do anything.

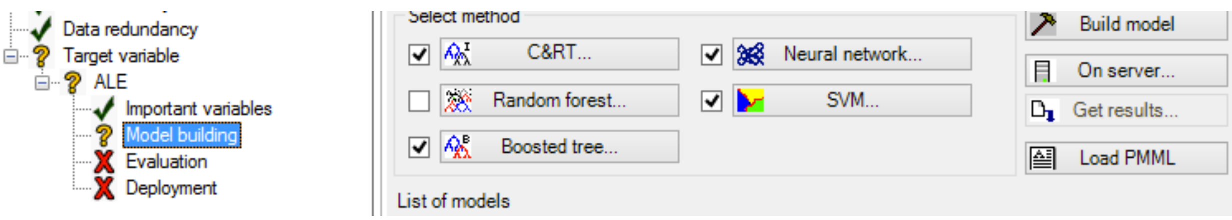
6. Important variables

STATISTICA DMR uses tree-based algorithms for finding important input predictor variables and interactions among them even after we have handled "data redundancy" previously to make sure simple interactions between variables.



Fast predictor screening has been chosen to make sure that the data has been optimized for models at best.

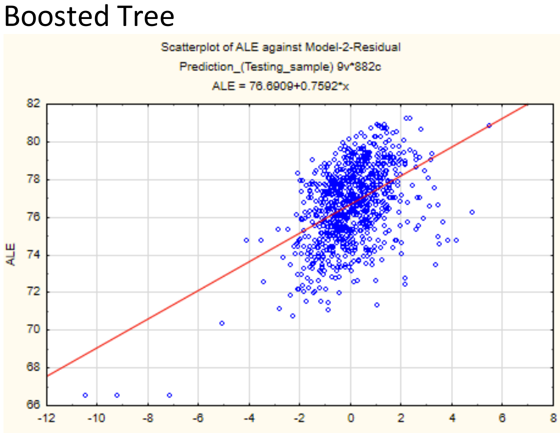
7. Model building.



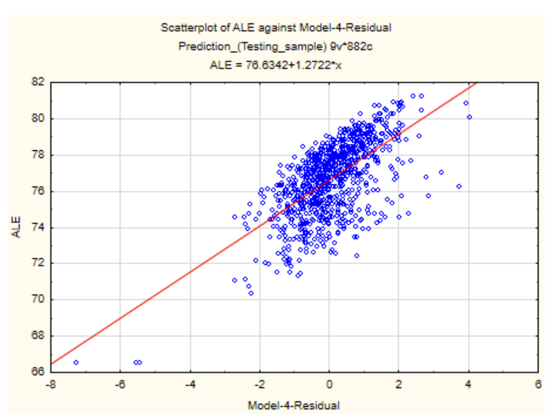
By default, the program automatically searches some predictive models, and is ready to “automate” the process them. Large-size data can take a while to build models.

It is also very useful to check a number of graphical displays to review how each model perform to predict the target of interest.

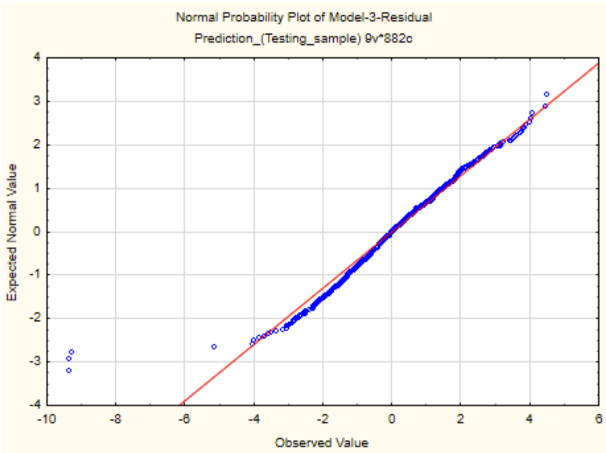
Boosted Tree



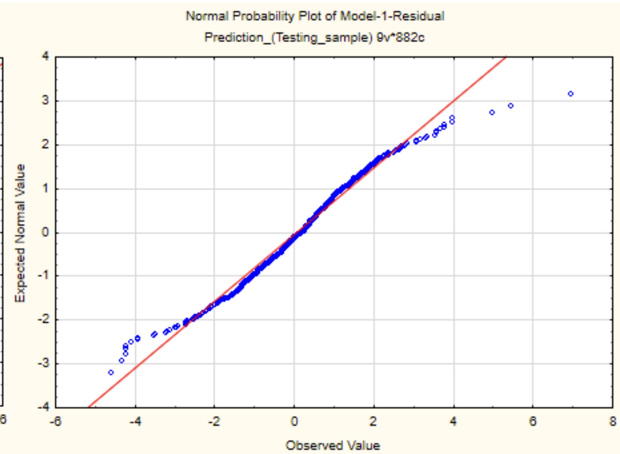
Neural Network



SVM



C&RT



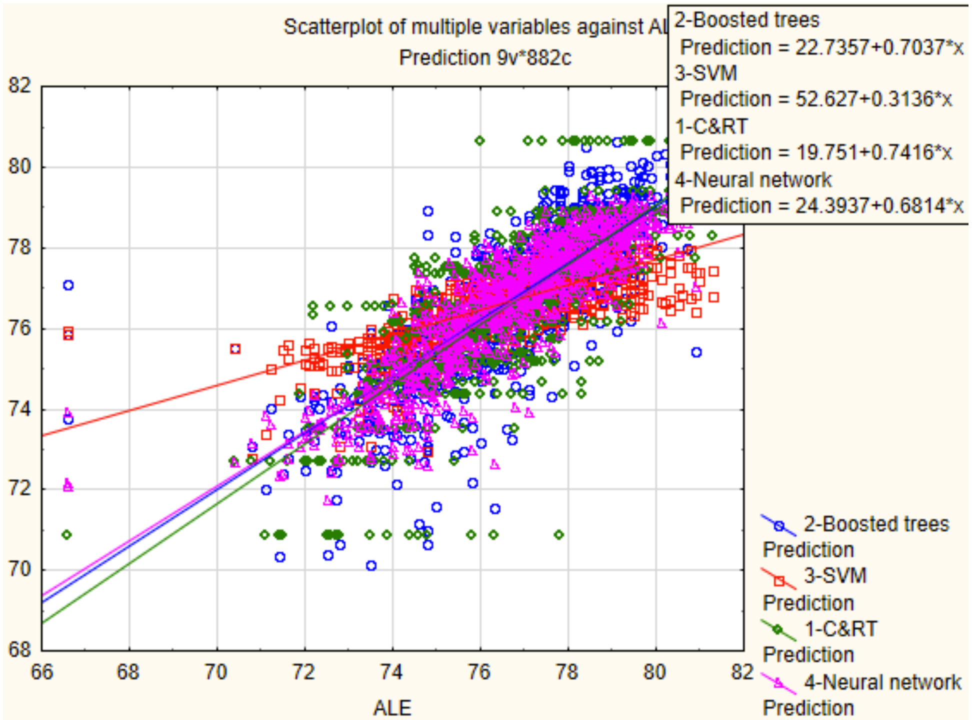
Once the models have finished building, we can get a table to show accuracy of each model.

List of models

Model ID	Name	Training residual (mean sum of square)	Testing residual (mean sum of square)	Correlation Coefficient (Training)	Correlation Coefficient (Testing)	Select for evaluation
4	Neural ...	0.93	1.04	0.87	0.87	TRUE
2	Boosted...	0.46	1.62	0.94	0.79	TRUE
1	C&RT	0.62	1.68	0.91	0.79	TRUE
3	SVM	1.85	2.26	0.80	0.76	TRUE

8. Model evaluation.

The models have been built. This is the process to test “trained models” with data sets that were not used before. The ability to predict new data is a very important part of predictive analysis. If the models do not perform, we will need to go back and investigate the data set and settings of models and try to re-build or change the existing models to meet the needs and goals.



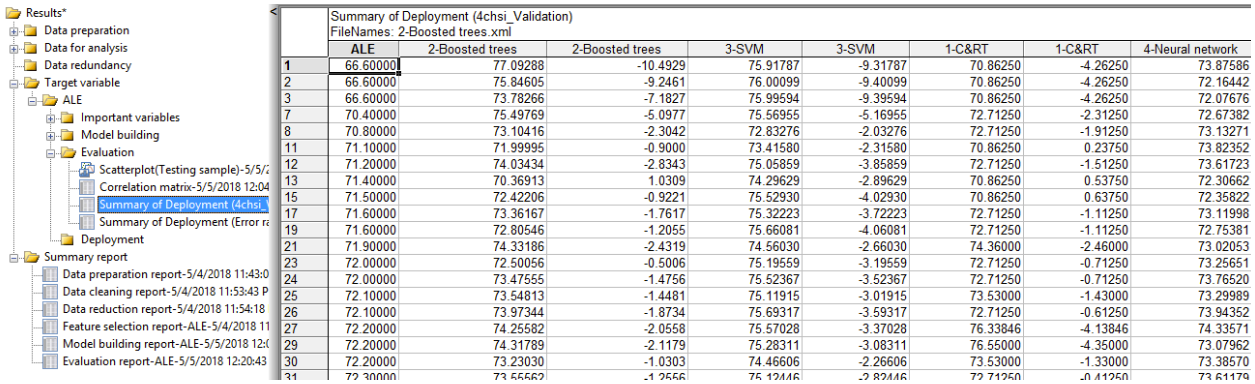
The scatterplots show that SVM model performs different from 3 other models. The other 3 models show very similar linear relationship.

Variable	Correlations (Prediction) Marked correlations are significant at p < .05000 N=882 (Casewise deletion of missing data) ALE
2-Boosted trees Prediction	0.787927
3-SVM Prediction	0.755165
1-C&RT Prediction	0.787848
4-Neural network Prediction	0.870155

Summary of Deployment (Error rates) (4chsi_Validation)	2-Boosted trees	3-SVM	1-C&RT	4-Neural
Error rate	1.620973	2.264428	1.677089	1.041139

Neural network shows the strongest correlations and smallest error rate.

After you click “Evaluate models” button, the result will tell in many ways about the models. You can also review “Summery of Deployment” to check the difference between observed data and predicted data for each model in case you need to check the data in a granular level.



Summary of Deployment (4chsi_Validation)
FileNames: 2-Boosted trees.xml

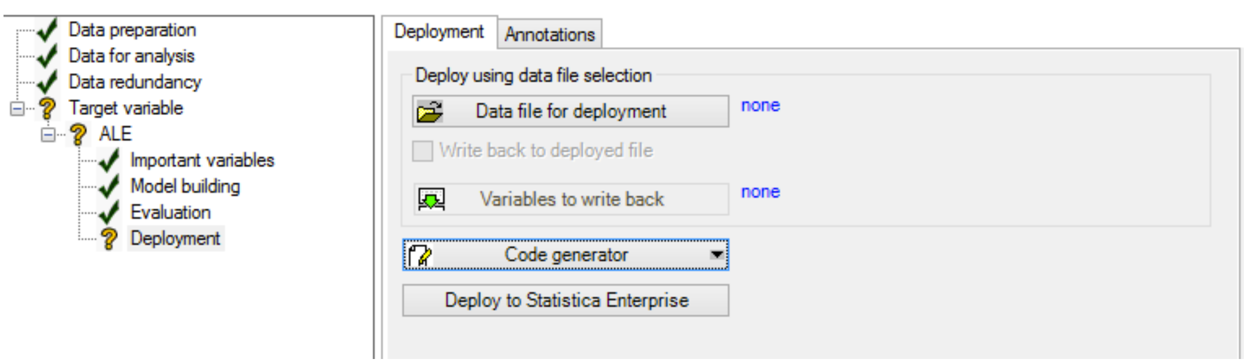
	ALE	2-Boosted trees	2-Boosted trees	3-SVM	3-SVM	1-C&RT	1-C&RT	4-Neural network
1	66.60000	77.09288	-10.4929	75.91787	-9.31787	70.86250	-4.26250	73.87586
2	66.60000	75.84605	-9.2461	76.00099	-9.40099	70.86250	-4.26250	72.16442
3	66.60000	73.78266	-7.1827	75.99594	-9.39594	70.86250	-4.26250	72.07676
7	70.40000	75.49769	-5.0977	75.56955	-5.16955	72.71250	-2.31250	72.67382
8	70.80000	73.10416	-2.3042	72.83276	-2.03276	72.71250	-1.91250	73.13271
11	71.10000	71.99995	-0.9000	73.41580	-2.31580	70.86250	0.23750	73.82352
12	71.20000	74.03434	-2.8343	75.05859	-3.85859	72.71250	-1.51250	73.61723
13	71.40000	70.36913	1.0309	74.29629	-2.89629	70.86250	0.53750	72.30662
15	71.50000	72.42206	-0.9221	75.52930	-4.02930	70.86250	0.63750	72.35822
17	71.60000	73.36167	-1.7617	75.32223	-3.72223	72.71250	-1.11250	73.11998
19	71.60000	72.80546	-1.2055	75.66081	-4.06081	72.71250	-1.11250	72.75381
21	71.90000	74.33186	-2.4319	74.56030	-2.66030	74.36000	-2.46000	73.02053
23	72.00000	72.50056	-0.5006	75.19559	-3.19559	72.71250	-0.71250	73.25651
24	72.00000	73.47555	-1.4756	75.52367	-3.52367	72.71250	-0.71250	73.76520
25	72.10000	73.54813	-1.4481	75.11915	-3.01915	73.53000	-1.43000	73.29989
26	72.10000	73.97344	-1.8734	75.69317	-3.59317	72.71250	-0.61250	73.94352
27	72.20000	74.25582	-2.0558	75.57028	-3.37028	76.33846	-4.13846	74.33571
29	72.20000	74.31789	-2.1179	75.28311	-3.08311	76.55000	-4.35000	73.07962
30	72.20000	73.23030	-1.0303	74.46606	-2.26606	73.53000	-1.33000	73.38570
31	72.30000	73.55562	-1.2556	75.12446	-2.82446	72.71250	-0.41250	73.61179

Overall, it is likely that Neural Network model will predict more accurately compared to other 3 models, boosted tree, SVM and C&RT. Among the 4 models, STATISTICA shows that Neural Network has the most significance as it has the least residual and the strongest correlation coefficient. Residual error is the difference between observed value and the estimated value. The smaller the number is the better the prediction is. Correlation coefficient is statistical relationship between variables. Usually it lays from 1 to -1. The closer to 1 or -1 is, the stronger relationship is.

7. Deployment.

Now it is the stage to actually use the model to predict with real world cases. The data will be brand new which have not been used both for training and testing. Successful predictive models should be able to predict new data with the accuracy that stakeholders can accept. Unfortunately, only STATISCA enterprise version is able to experience actual deployment.

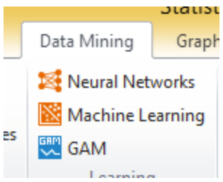
Additionally you may also like to download PMML xml code for single data mining algorithm. Click “Code generator”, select PMML and save XML file somewhere.



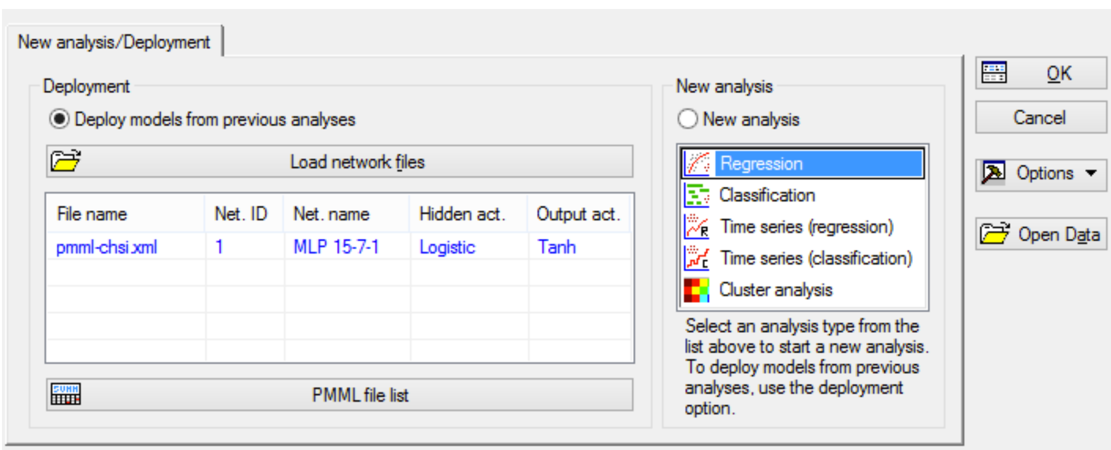
Neural Networks

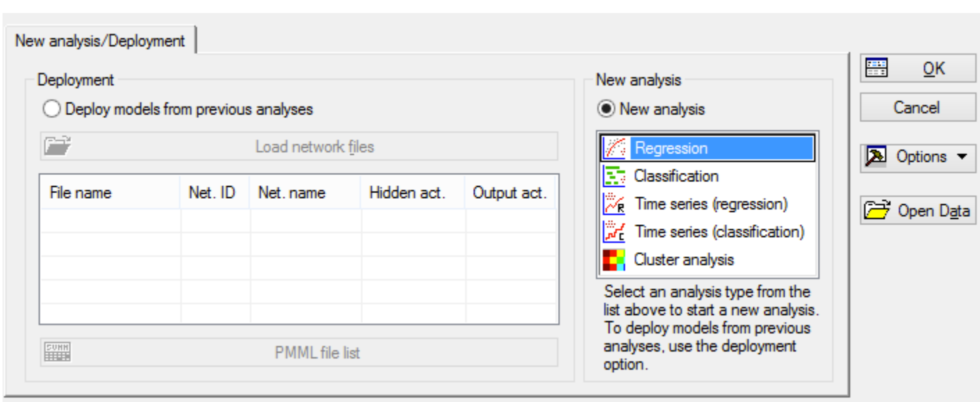
From DM recipe, we learned that neural network model is the best among 4 models. This algorithm will be run again individually with more control and optimization in order to see if this can be performed even better.

1. Click “Data Mining” tab, and click “Neural Networks”



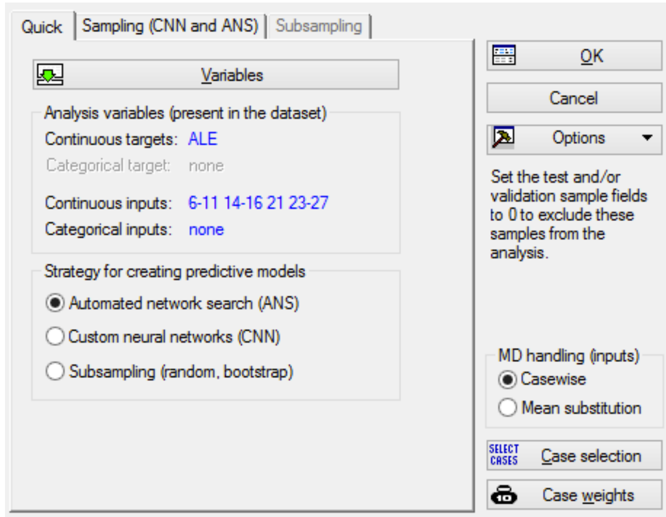
2. You have option to deploy from the previous analysis, or you can select new analysis.



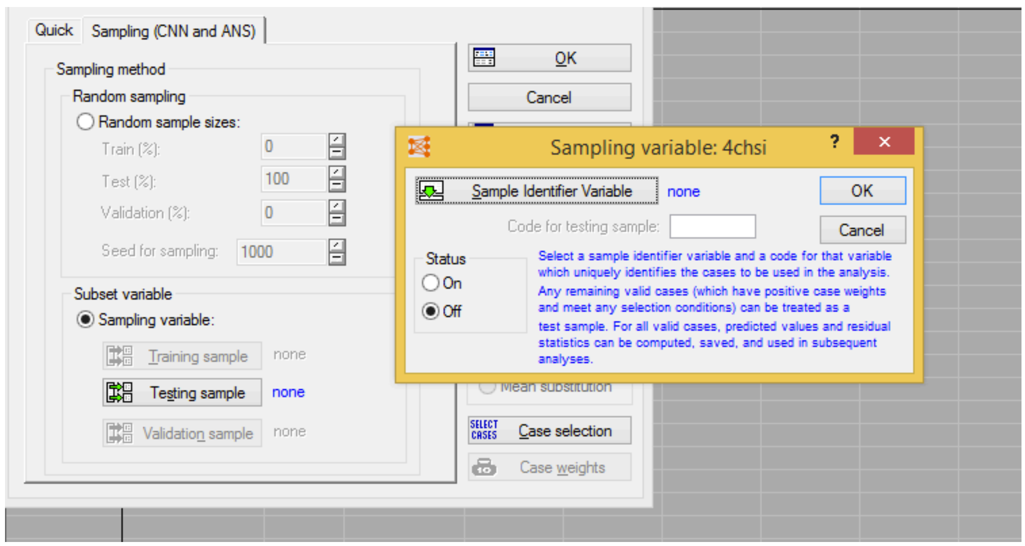


New analysis with “Regression” option has been started.

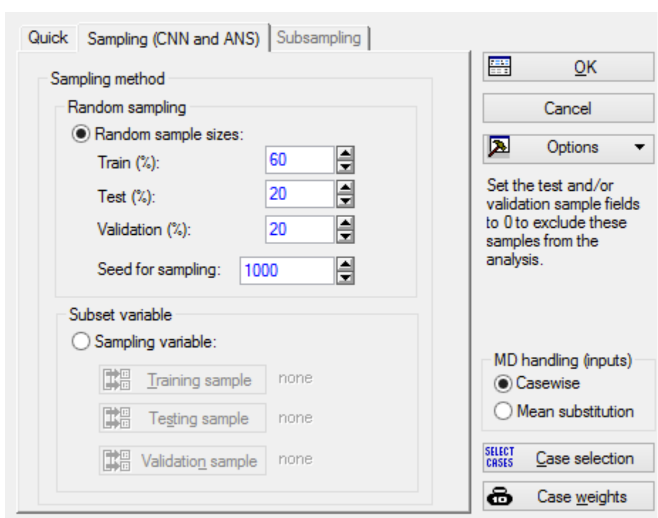
3. As you click the "ok", you will see a dialogue box like the below. Just like other times, variables need to be specified.



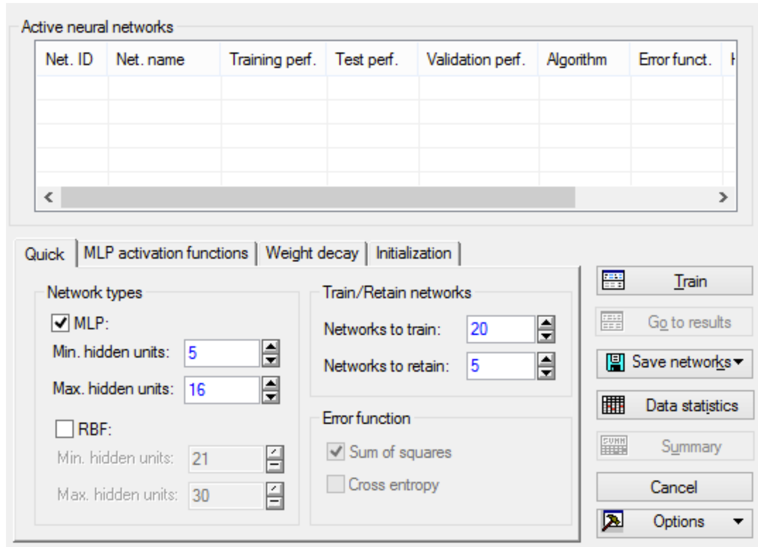
You have an option to choose sample differently for subsequent analysis.



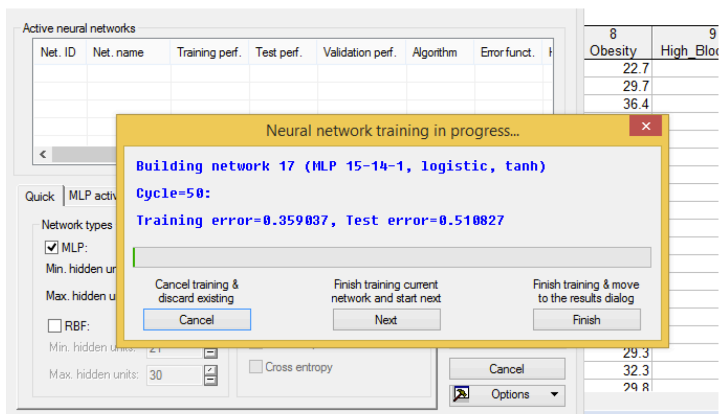
At this time, random sampling has been selected as below.



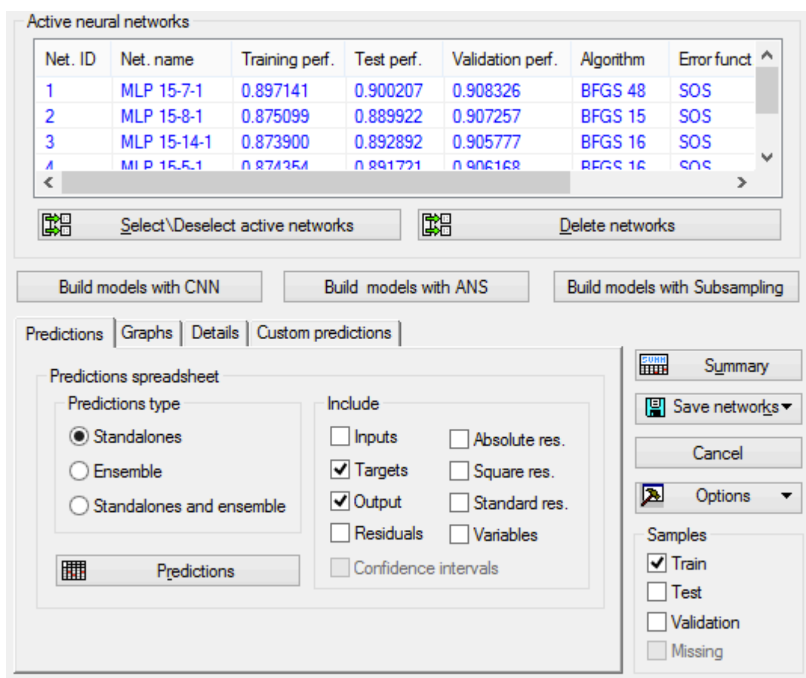
4. Now the setup has been done and we can start training. Click “Train” button.



It’s running



The result is like below. ANN will train and retain 5 networks. When training is complete, the ANN Results dialog box will be displayed.



5. Check the results

Summary of active networks (4chsi)

Index	Net. name	Training perf.	Test perf.	Validation perf.	Training error	Test error	Validation error	Training algorithm	Error function
1	MLP 15-7-1	0.897141	0.900207	0.908326	0.371792	0.416754	0.382915	BFGS 48	SOS
2	MLP 15-8-1	0.875099	0.889922	0.907257	0.446760	0.449590	0.388438	BFGS 15	SOS
3	MLP 15-14-1	0.873900	0.892892	0.905777	0.450461	0.444697	0.398477	BFGS 16	SOS
4	MLP 15-5-1	0.874354	0.891721	0.906168	0.449220	0.444231	0.392224	BFGS 16	SOS
5	MLP 15-10-1	0.893080	0.896645	0.912867	0.385652	0.431039	0.371473	BFGS 43	SOS

	Predictions statistics (4chsi) Target: ALE				
Statistics	1.MLP 15-7-1	2.MLP 15-8-1	3.MLP 15-14-1	4.MLP 15-5-1	5.MLP 15-10-1
Minimum prediction (Train)	66.89156	71.0671	70.8289	71.2535	66.6550
Maximum prediction (Train)	80.91928	80.7167	80.4183	80.4440	80.7441
Minimum prediction (Test)	71.42406	72.2005	72.1860	72.0951	71.1161
Maximum prediction (Test)	80.98818	81.1413	80.6312	80.8834	81.1974
Minimum prediction (Validation)	71.15719	71.7439	71.8423	71.9848	70.5609
Maximum prediction (Validation)	80.95542	80.9070	80.5511	80.6711	80.5098
Minimum prediction (Missing)					
Maximum prediction (Missing)					
Minimum residual (Train)	-3.37751	-4.4817	-4.2521	-5.0214	-3.3432
Maximum residual (Train)	2.85772	3.8457	3.8409	3.6414	2.6715
Minimum residual (Test)	-5.53569	-7.2552	-7.2853	-7.3585	-7.3152
Maximum residual (Test)	2.79133	2.8776	2.7121	2.8340	2.5044
Minimum residual (Validation)	-3.88968	-3.5671	-3.5935	-3.6985	-3.2611
Maximum residual (Validation)	2.78152	2.7225	3.1469	3.0620	2.7648
Minimum standard residual (Train)	-5.53919	-6.7051	-6.3353	-7.4920	-5.3836
Maximum standard residual (Train)	4.68672	5.7536	5.7228	5.4330	4.3019
Minimum standard residual (Test)	-8.57495	-10.8204	-10.9248	-11.0403	-11.1422
Maximum standard residual (Test)	4.32386	4.2916	4.0670	4.2520	3.8146
Minimum standard residual (Validation)	-6.28583	-5.7234	-5.6927	-5.9055	-5.3505
Maximum standard residual (Validation)	4.49501	4.3683	4.9852	4.8892	4.5363

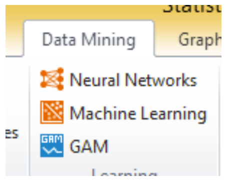
	Correlation coefficients (4chsi)		
	ALE Train	ALE Test	ALE Validation
1.MLP 15-7-1	0.897141	0.900207	0.908326
2.MLP 15-8-1	0.875099	0.889922	0.907257
3.MLP 15-14-1	0.873900	0.892892	0.905777
4.MLP 15-5-1	0.874354	0.891721	0.906168
5.MLP 15-10-1	0.893080	0.896645	0.912867

15-7-1 has better predicting performance compared to other networks. All 5 models are in the similar range. 15-10-1 has yield higher correlation in test and validation data set. Earlier in DM recipe, NN had correlation of 0.87 whereas ANN made it up to 91%. Adding hidden units and iterative training, the significance has been improved. You have an option to choose certain network to build and deploy the model.

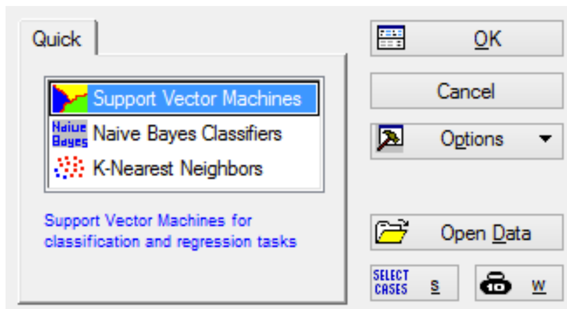
Support Vector Machine

We wanted to give it a try if support vector machine can be improved by building it individually, and possibly even better than ANNs have performed above.

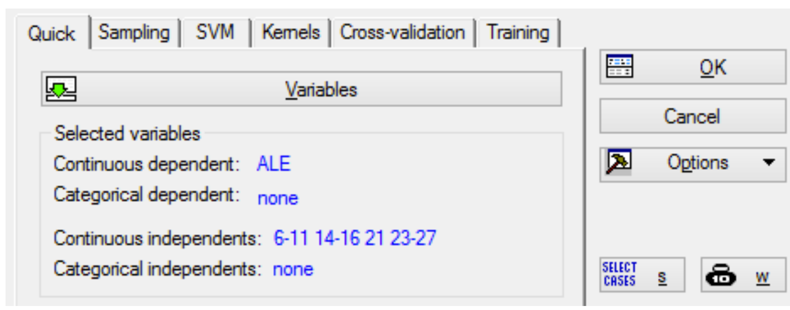
1. Click "Data Mining" tab, and click "Machine Learning"



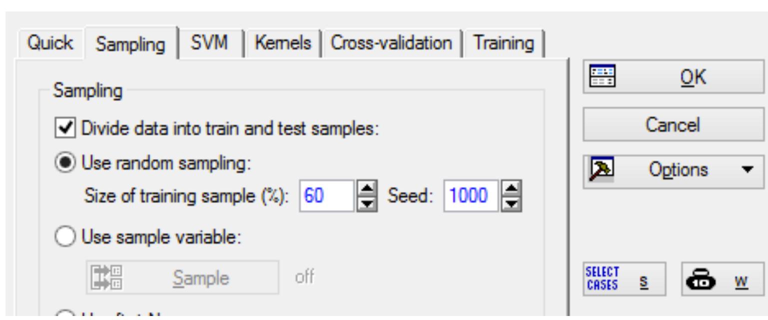
2. Click Support Vector Machine



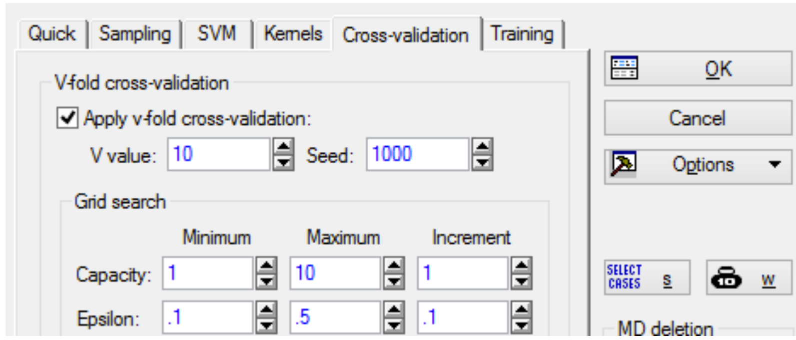
3. As you click the "ok", you will see a dialogue box like the below. Just like other times, variables need to be specified.



4. Specify Sampling

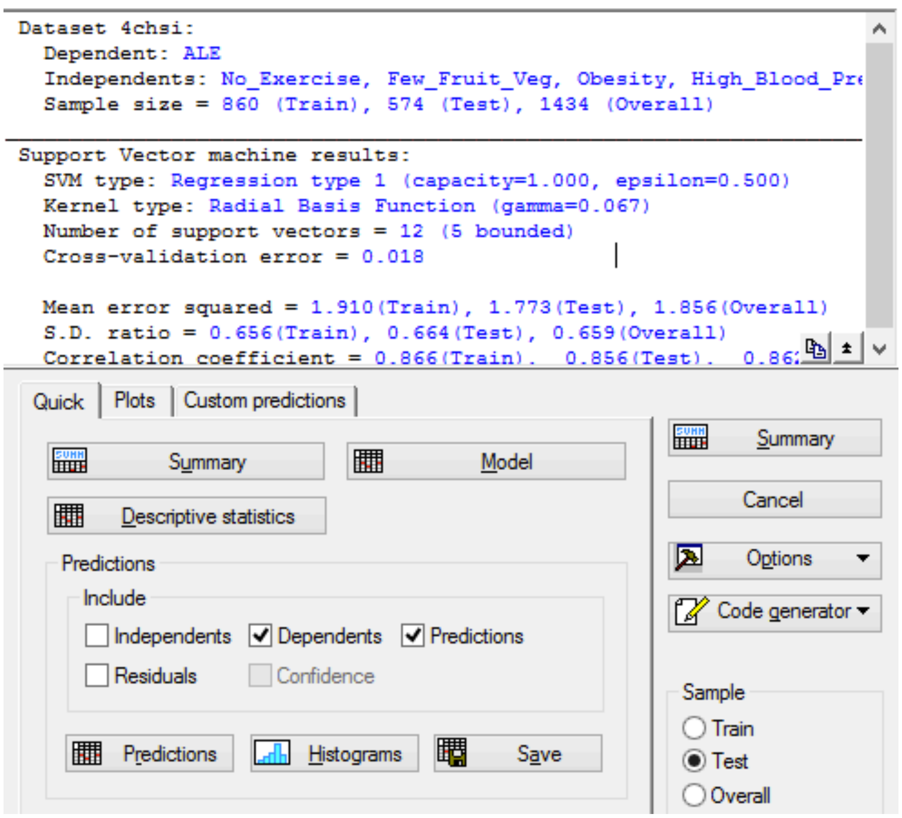


5. V-fold cross validation



Now, display the "Cross-validation" tab and select the "Apply v-fold cross-validation" check box. Click OK to initiate SVM training (model fitting), which is carried out in two stages. In the first stage, a search is made for an estimate of the capacity constant that achieves the highest classification accuracy. In the second phase of training, the estimated value of capacity is used to train an SVM model using the entire training sample. When training is finished, the Support Vector Machine Results dialog is displayed.

6. Results



Regression summary (Support Vector Machines), Test sample (4chsi)
SVM: Regression type 1 (C=1.000, epsilon=0.500), Kernel: Radial Basis Function (gamma=0.067)
Number of support vectors= 12 (5 bounded)

Regression summary	ALE
Observed mean	76.73606
Predictions mean	76.37348
Observed S.D.	1.93182
Predictions S.D.	0.84965
Mean squared error	1.77329
Error mean	0.36258
Error S.D.	1.28246
Abs. error mean	1.07073
S.D. ratio	0.66386
Correlation	0.85573

You can review the results of SVM at the result dialog. The summary box at the top show specification of model information including number of support vectors, types, parameters and more. The detail view can be seen at regression summery. The tab of Plots can make many graphs like histogram and scatterplots.

Mean error squared is 1.773 which is higher than Neural network model, correlation coefficient is 0.856 which is slightly less than NN model. After we deployed V-fold cross validation, the overall performance has been improved, but still NN model seems to be better.

Check out simple results from 2 other regression models below.

Boosted Tree Model

Risk estimates (4chsi)
Response: ALE

	Risk Estimate	Standard error
Train	0.763579	0.038092
Test	1.207956	0.170425

K-Nearest Neighbor Model

Regression s
Nearest neigh

Regression summary	ALE
Observed mean	76.66361
Predictions mean	76.64289
Observed S.D.	2.01199
Predictions S.D.	1.66382
Sum of squared error	1.03587
Error mean	0.02072
Error S.D.	1.01898
Abs. error mean	0.75894
S.D. ratio	0.50645
Correlation	0.86302

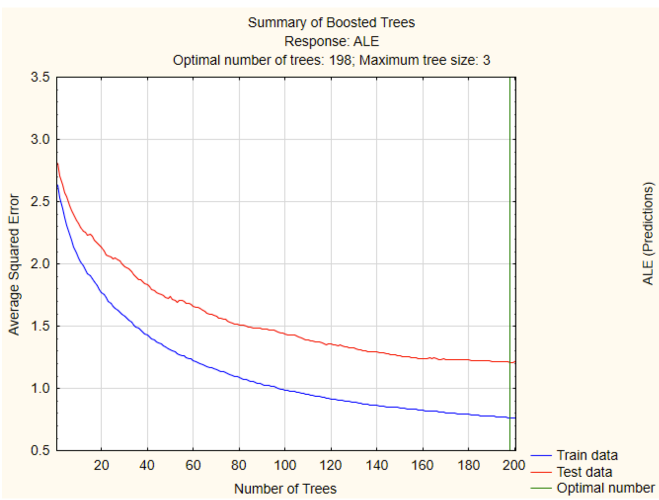
Model Comparison

After building 4 models, we should compare them in order to find the most optimal model. Since the dependent variable is continuous and they all are regression models, 3 matrixes have been chosen to measure their significance: mean squared error, correlation coefficient, residual diagnostic graphs.

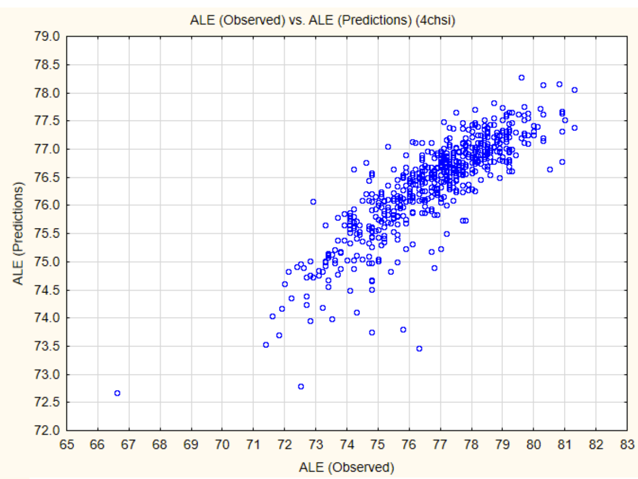
	Neural Network	SVM	Tree	K-Nearest
1. Mean Squared Error	0.67	1.77	0.76	1.03
2. Correlation Coefficient	0.91	0.85	0.79	0.86

3. Residual Diagnostics

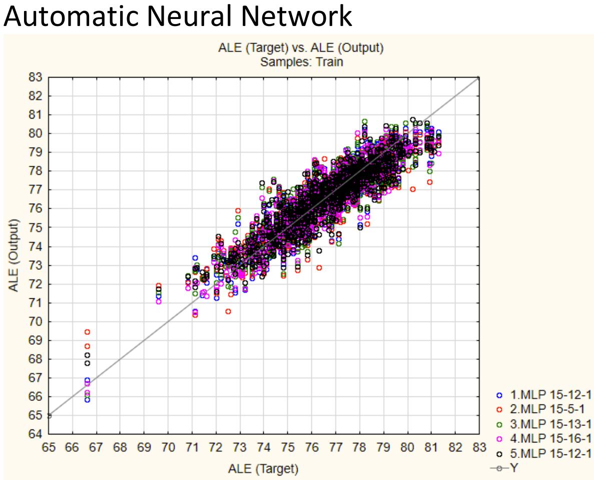
Boosted Tree



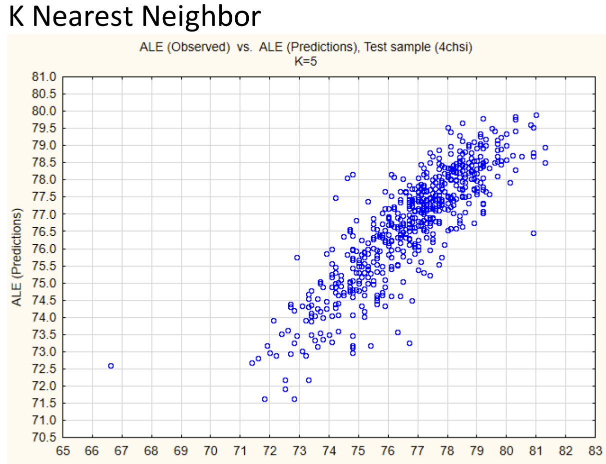
SVM



Automatic Neural Network



K Nearest Neighbor



Summery

Overall, it has been concluded that neural network model would be chosen to predict life expectancy as the neural network model has the smallest error and the greatest correlation coefficient.

Next page includes some learnings from the class and experiment with classification model.

Experiment Classification Model

Just to experience the case of classification model, bins have been created as another target class of life expectancy during data preparation stage. And decision tree classification model has been made at KNIME.

Some of the validation have been checked after the model has been built.

Confusion Matrix:
Accuracy is over 97%. This seems “too” good. It would be worth checking if bin size is too big, the model has overfitting problem, the data itself has been already too generalized, sample size is too small, or just the model is incredibly good.

ALE 2 \ Pre...	70	74	76	78	80
70	61	0	0	0	0
74	4	196	13	0	0
76	0	0	318	0	0
78	0	0	17	510	0
80	0	0	0	0	218

Correct classified: 1,303 Wrong classified: 34

Accuracy: 97.457 % Error: 2.543 %

Cohen's kappa (κ) 0.965

Cumulative Gain chart (Lift Chart):
The blue straight line is the base.
The red curved line is the actual.

The further the curved line is from the straight line, the better.

When there are more than 1 lift chart from various classification models, the model with bigger curve will be picked. And optimal point can be noticed from the curved line as well.