DSCI 4520 – Introduction to Data Mining
Fall 2024

Medical Insurance Premium Prediction

Group 3

Deepesh Bashyal

George Bechara

Samantha Perez

Maahum Sattar

Nathan Stepp

Zack Warriner

1. Business Case

For our project, we will create a regression model to predict an individual's insurance premium price based on the patient's medical history. This model will provide valuable insights into how certain predictors impact premium costs, allowing for more informed decision-making by both individuals and insurance firms.

Solving this problem is important because it could enable people with extensive medical histories to plan for potential costs of medical insurance premiums. This analysis could also highlight disparities in premium pricing and give individuals an opportunity to advocate for more equitable treatment. Alternatively, insurance firms could use this data to optimize pricing and ensure that premium price directly and fairly correlates to an individual medical history.

Target variable: Insurance premium price

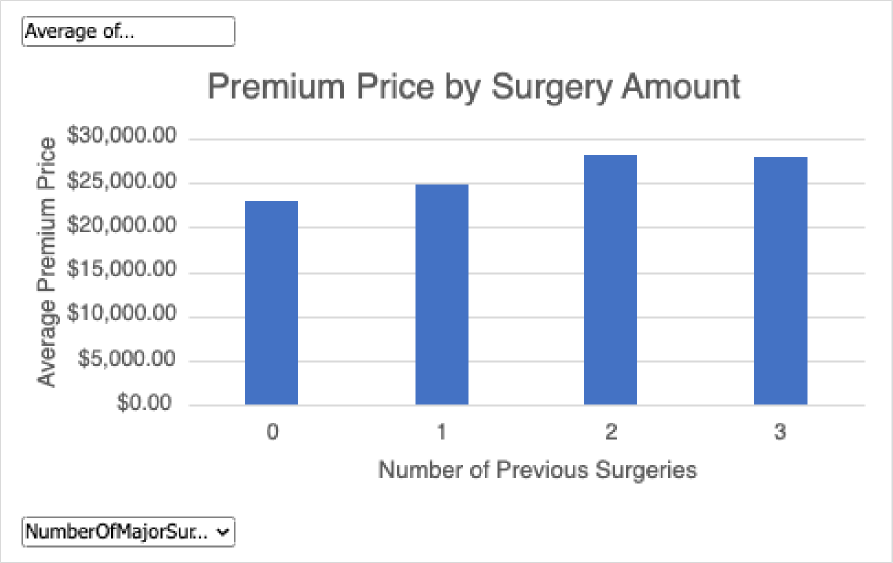
Predictors included in our dataset are

- Age,
- Height
- Weight
- Diabetes
- Blood pressure problems
- Transplants
- Chronic Disease
- Known Allergies
- History of cancer in family
- Number of major surgeries

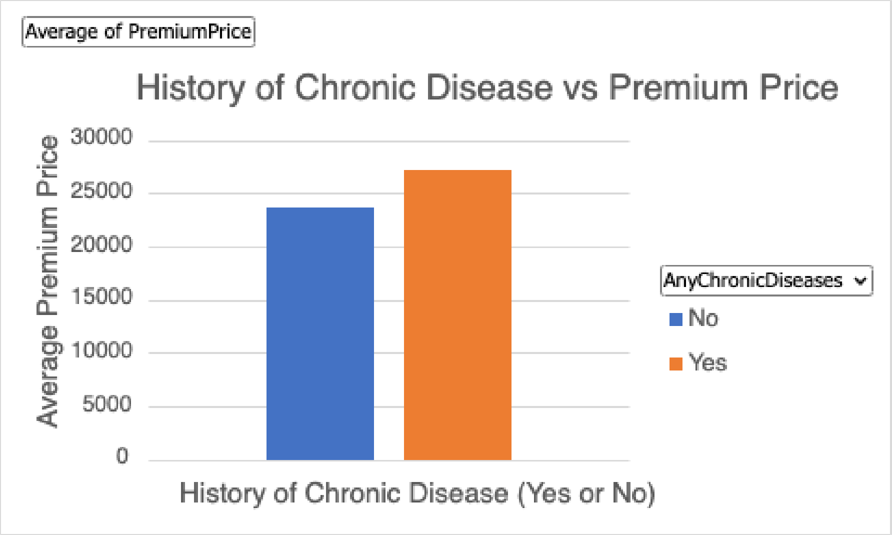
There are 986 observations in our dataset. We will start by using Excel and Tableau to create data visualizations. The data visualizations will help us determine which variables affect and do not affect our target variable. This will also tell us if we need to exclude anything from our dataset. Secondly, we will use rapid miner to build and evaluate the model. We plan to use Linear Regression, K-Nearest Neighbors (K-NN), and Random Forest techniques. 70% of the dataset will be used as a training model and the remaining 30% of the dataset will be used to validate our model's performance.

2. Data Exploration and Visualization

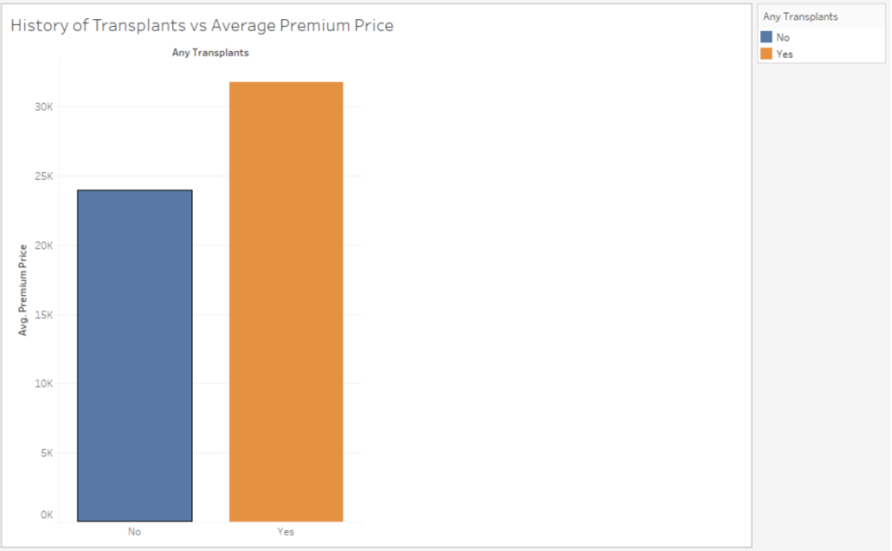
Before we created our model, we first wanted to understand our dataset. From our data dictionary, we found our dataset has a mix of both categorial and numerical predictor variables, but not all these variables affect our target variable. As such, we created the following visualizations to help determine what predictors to include in our dataset:



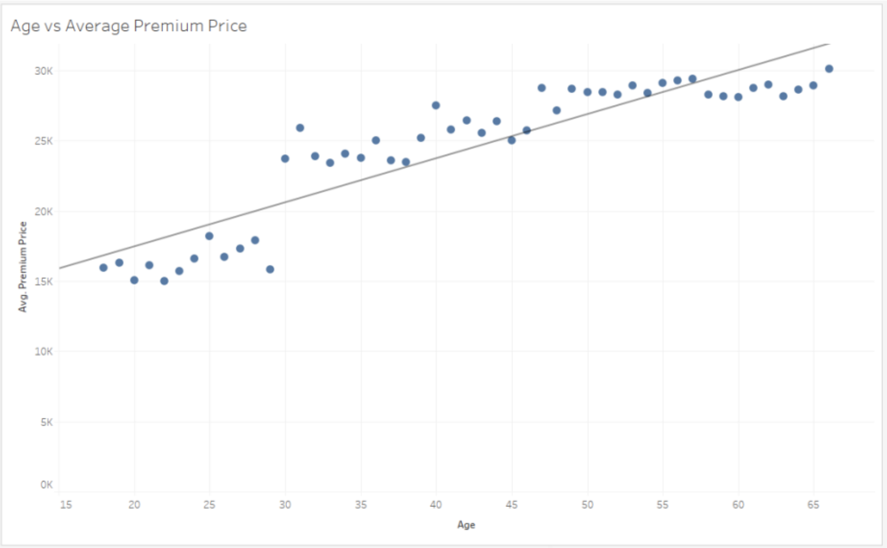
Average premium price by number of previous surgeries shows an approximately uniform distribution. However, average premium price for those within 2 or 3 surgeries are higher than those with 0 or 1 surgeries.



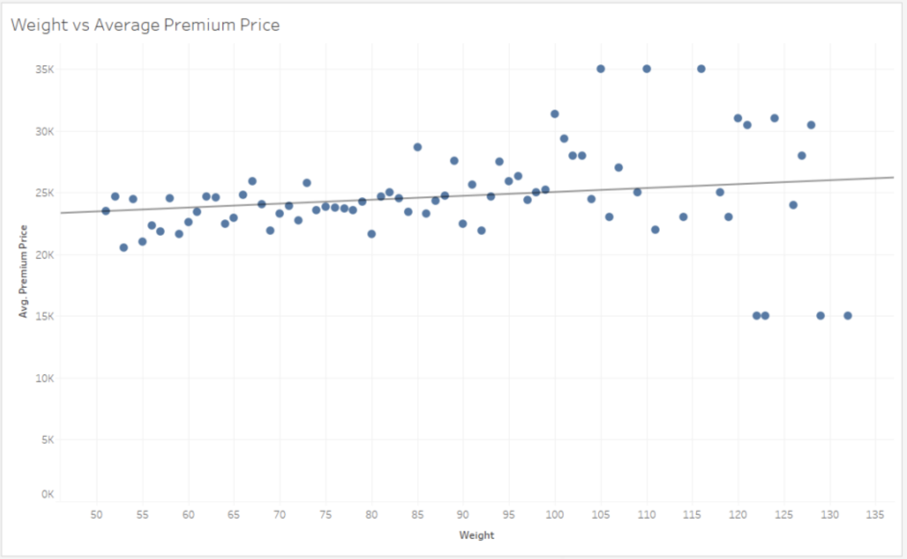
History of chronic diseases vs average premium price shows that people with a history of chronic diseases have a slightly higher average premium price than those with no history of chronic diseases.



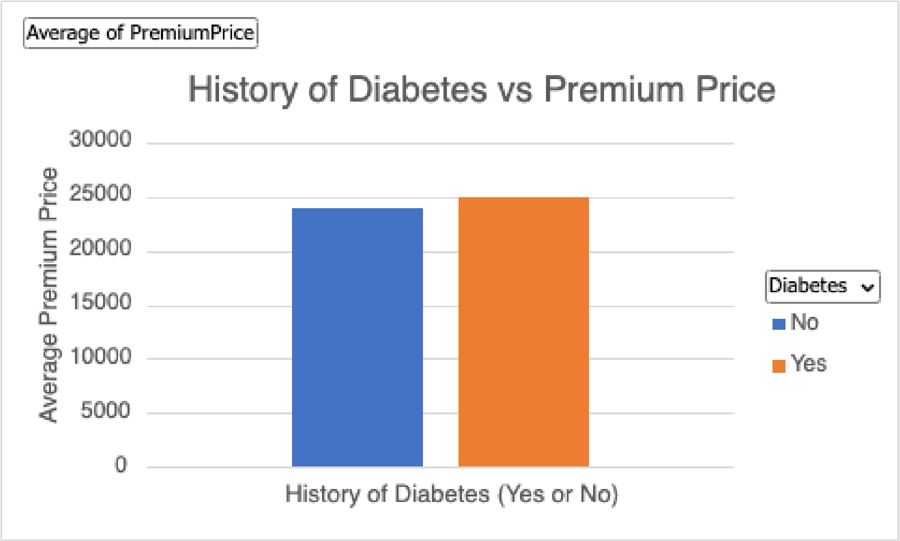
History of prior transplants vs average premium price shows that people who have had prior transplants have a noticeably higher average premium price than those with no prior transplants.



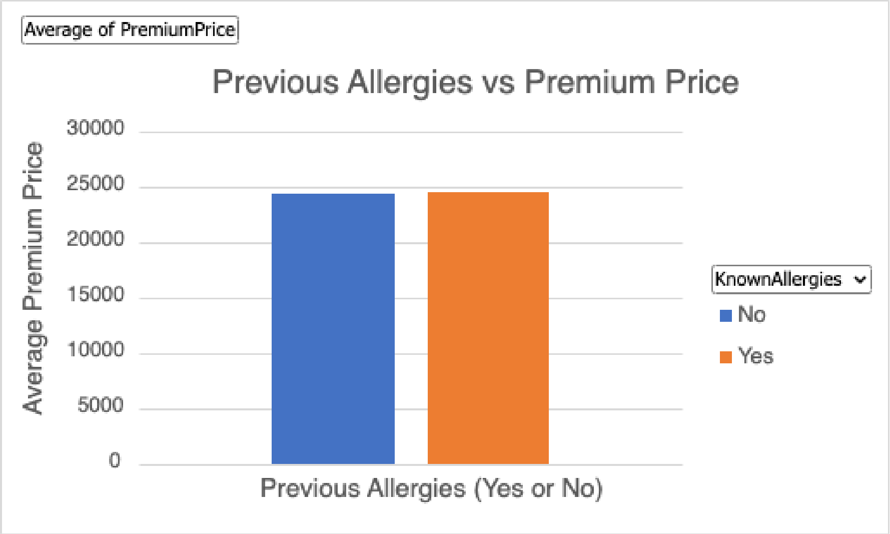
Age vs average premium price scatterplot shows a positive correlation between age and premium price. This plot has an R squared value of 0.8188 which tells us that 81.88% of the variation in premium price can be explained by age.



Weight vs average premium price scatterplot shows no correlation between weight and premium price. This plot has an R squared value of 0.0347 which tells us weight does not explain any variation in our data set.



History of diabetes vs average premium shows almost no difference in average premium price for those with or without a history of diabetes



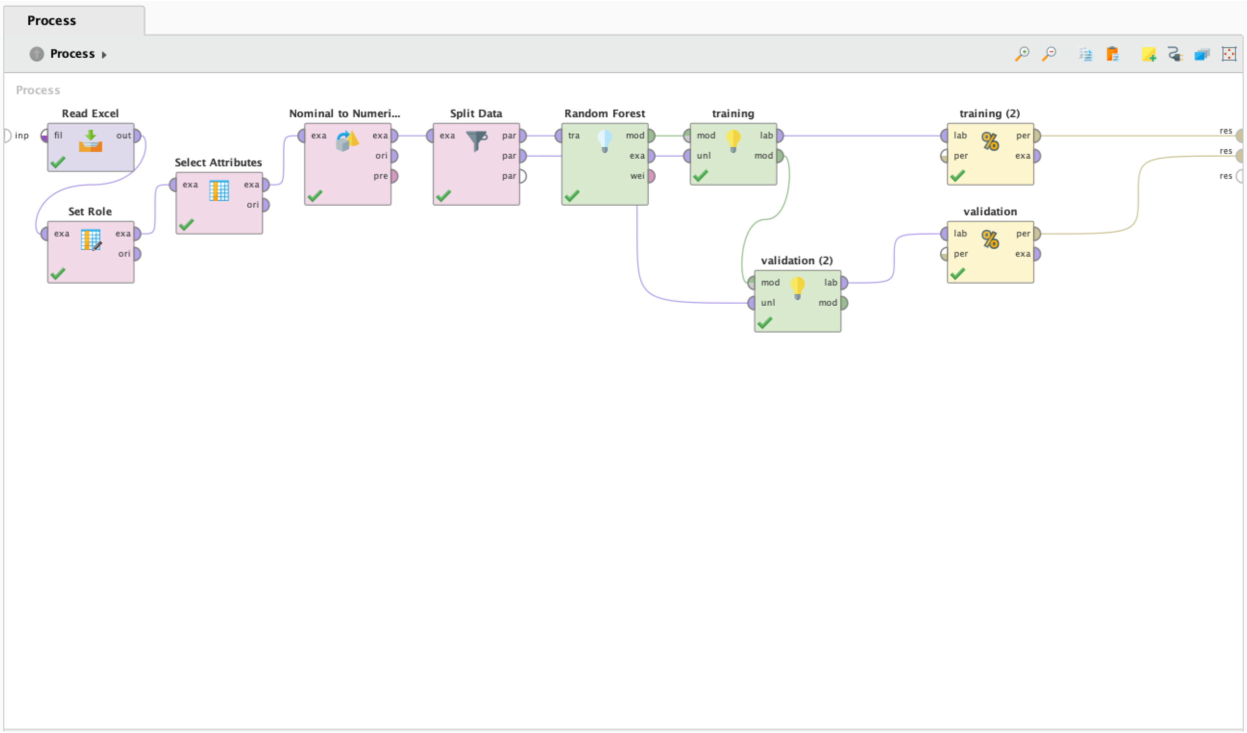
History of known allergies vs average premium shows almost no difference in average premium price for those with or without known allergies

Conclusion: our model will consider predictors like number of surgeries, history of chronic disease, history of transplants, history of cancer, history of blood pressure, and age. Our model will not consider predictors like weight, history of diabetes, and known allergies.

3. Predictive Model

For our predictive model, we used RapidMiner and created Random Forest, K-NN and linear regression models.

1. Random Forest Model



Our random forest model reads the excel input, sets our target variable, selects our desired attributes (based on our data exploration results), converts all variables from nominal to numerical, and splits data into two groups; training and validation. The random forest is then performed on the training group & a regression performance operator is applied to both the training and validation groups so that we can evaluate how our model performed. We performed this model multiple times with different numbers of trees and compared the root mean squared error (RMSE) and R^2 of our validation group. The below tables summarizes our results:

Table 1: No attributes excluded

Number of trees in Model	RMSE	R^2
20	3225.832	0.722
30	3208.788	0.726
45	**3166.227**	**0.733**
50	3169.527	0.732
75	3176.748	0.731

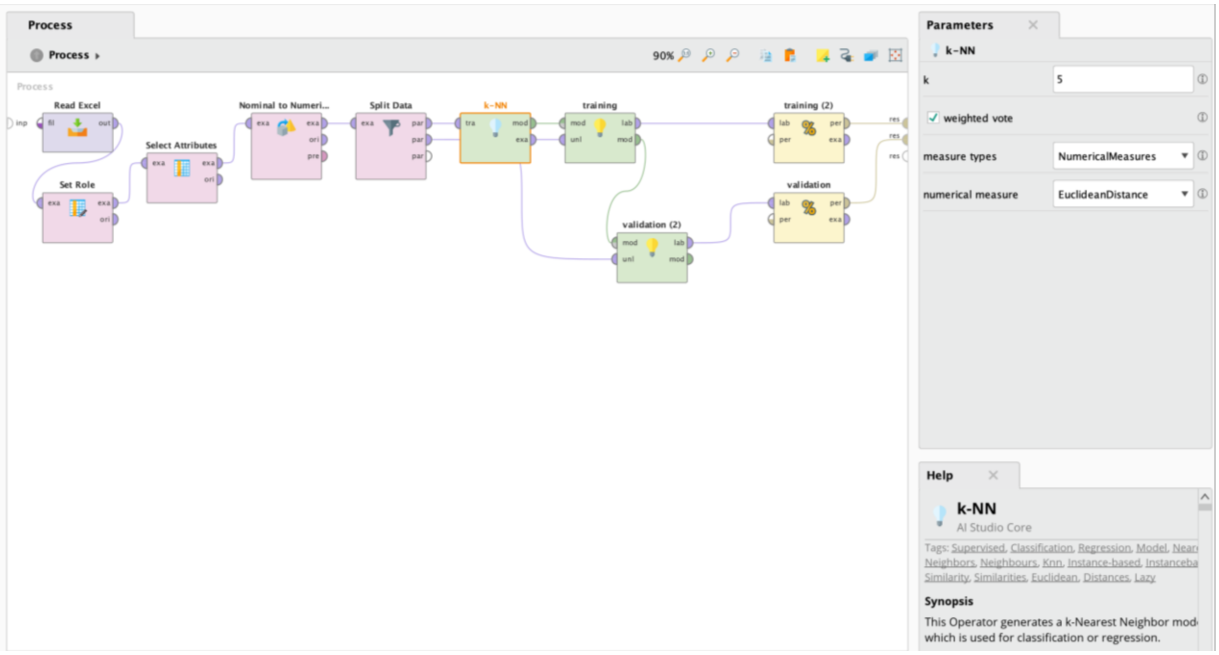
100	3186.767	0.729

Table 2: Known Allergies, Weight, and History of Diabetes excluded

Number of trees in Model	RMSE	R^2
20	3426.342	0.685
30	3154.343	0.737
45	3140.086	0.739
50	3139.242	0.739
75	**3120.021**	**0.742**
100	3135.478	0.739

As seen in Table 1, when no variables are excluded the optimal number of trees is 45 where RMSE is minimized at 3166.227 and R^2 is maximized at 0.733. As seen in table 2, when some variables are excluded, the optimal number of trees is 75 where RMSE is minimized at 3120.021 and R^2 is maximized at 0.742. When our model has fewer attributes, more trees are required to accurately capture the relationships between variables.

2. K-NN Model



k-NN

k: 5

☑ weighted vote

measure types: NumericalMeasures

numerical measure: EuclideanDistance

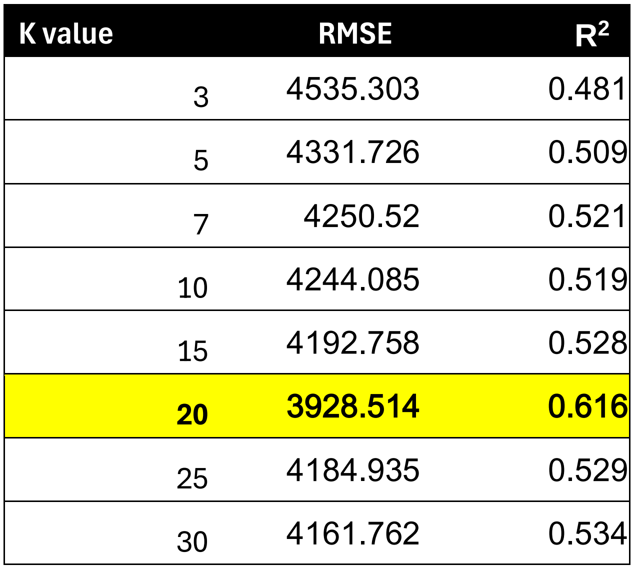
k-NN
AI Studio Core

Tags: Supervised, Classification, Regression, Model, Neare Neighbors, Neighbours, Knn, Instance-based, Instanceba Similarity, Similarities, Euclidean, Distances, Lazy

Synopsis

This Operator generates a k-Nearest Neighbor mod which is used for classification or regression.

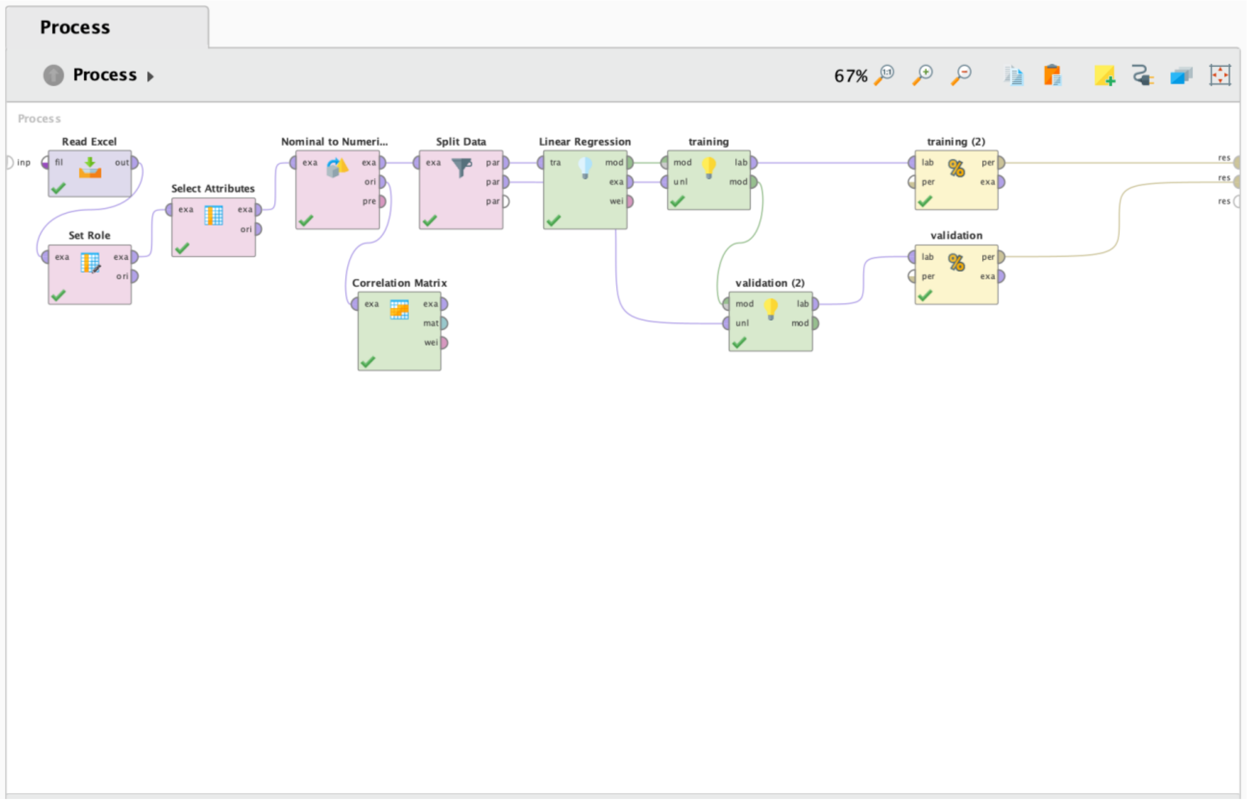
Our K-NN model reads the excel input, sets our target variable, selects our desired attributes (based on our data exploration results), converts all variables from nominal to numerical, and splits data into two groups; training and validation. The K-NN operator is then performed on the training group & a regression performance operator is applied to both the training and validation groups so that we can evaluate how our model performed. We performed this model multiple times with different numbers of k and compared the root mean squared error (RMSE) and R^2 of our validation group. The below table summarizes our results:



K value	RMSE	R^2
3	4535.303	0.481
5	4331.726	0.509
7	4250.52	0.521
10	4244.085	0.519
15	4192.758	0.528
20	**3928.514**	**0.616**
25	4184.935	0.529
30	4161.762	0.534

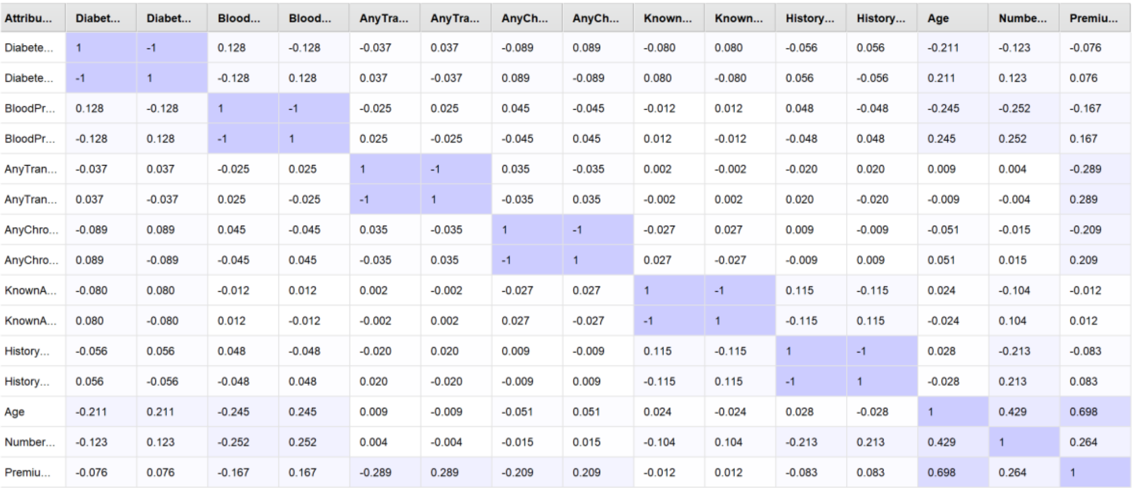
As we see in our table, the optimal value of k is 20 as this value minimizes RMSE at 3928.514 and maximizes R^2 at 0.616. This tells us that there is no underfitting or overfitting occurring.

3. Linear Regression



Our linear regression model reads the excel input, sets our target variable, selects our desired attributes (based on our data exploration results), converts all variables from nominal to numerical, and splits data into two groups; training and validation. The linear regression operator is then performed on the training group & a regression performance operator is applied to both the training and validation groups so that we can evaluate how our model performed. We performed this model multiple times with different variables being excluded/included and compared the root mean squared error (RMSE) and R^2 of our validation group. The below table shows our results:

Model	RMSE	R^2
All Attributes included (baseline)	4000.872	0.572
Weight excluded	4113.368	0.548
Known Allergies excluded	3984.241	0.575
History of Diabetes excluded	4008.429	0.571
All 3 excluded	4108.872	0.549



Attribu...	Diabet...	Diabet...	Blood...	Blood...	AnyTra...	AnyTra...	AnyCh...	AnyCh...	Known...	Known...	History...	History...	Age	Numbe...	Premiu...
Diabete...	1	-1	0.128	-0.128	-0.037	0.037	-0.089	0.089	-0.080	0.080	-0.056	0.056	-0.211	-0.123	-0.076
Diabete...	-1	1	-0.128	0.128	0.037	-0.037	0.089	-0.089	0.080	-0.080	0.056	-0.056	0.211	0.123	0.076
BloodPr...	0.128	-0.128	1	-1	-0.025	0.025	0.045	-0.045	-0.012	0.012	0.048	-0.048	-0.245	-0.252	-0.167
BloodPr...	-0.128	0.128	-1	1	0.025	-0.025	-0.045	0.045	0.012	-0.012	-0.048	0.048	0.245	0.252	0.167
AnyTran...	-0.037	0.037	-0.025	0.025	1	-1	0.035	-0.035	0.002	-0.002	-0.020	0.020	0.009	0.004	-0.289
AnyTran...	0.037	-0.037	0.025	-0.025	-1	1	-0.035	0.035	-0.002	0.002	0.020	-0.020	-0.009	-0.004	0.289
AnyChro...	-0.089	0.089	0.045	-0.045	0.035	-0.035	1	-1	-0.027	0.027	0.009	-0.009	-0.051	-0.015	-0.209
AnyChro...	0.089	-0.089	-0.045	0.045	-0.035	0.035	-1	1	0.027	-0.027	-0.009	0.009	0.051	0.015	0.209
KnownA...	-0.080	0.080	-0.012	0.012	0.002	-0.002	-0.027	0.027	1	-1	0.115	-0.115	0.024	-0.104	-0.012
KnownA...	0.080	-0.080	0.012	-0.012	-0.002	0.002	0.027	-0.027	-1	1	-0.115	0.115	-0.024	0.104	0.012
History...	-0.056	0.056	0.048	-0.048	-0.020	0.020	0.009	-0.009	0.115	-0.115	1	-1	0.028	-0.213	-0.083
History...	0.056	-0.056	-0.048	0.048	0.020	-0.020	-0.009	0.009	-0.115	0.115	-1	1	-0.028	0.213	0.083
Age	-0.211	0.211	-0.245	0.245	0.009	-0.009	-0.051	0.051	0.024	-0.024	0.028	-0.028	1	0.429	0.698
Number...	-0.123	0.123	-0.252	0.252	0.004	-0.004	-0.015	0.015	-0.104	0.104	-0.213	0.213	0.429	1	0.264
Premiu...	-0.076	0.076	-0.167	0.167	-0.289	0.289	-0.209	0.209	-0.012	0.012	-0.083	0.083	0.698	0.264	1

As seen in our table, our best performing model occurs when only Known Allergies is excluding as our RMSE is minimized at 3984.241 and our R^2 is maximized at 0.575. This tells us that known allergies may be introducing unnecessary variation to our model. This also tells us that Weight and History of Diabetes may be positive contributions to the models performance as the RMSE increased and R^2 decreased when these variables where removed. The correlation matrix shows us that the variable Age has a strong positive correlation with insurance premium price and variable Known Allergies has a very weak correlation to premium price, as confirmed by our model.

4. Conclusion

In this project, we developed a predictive model to estimate an individual's insurance premium price based on their medical history. Through data exploration and visualization, we identified key predictors such as age, number of surgeries, history of chronic disease, transplants, cancer, and blood pressure problems, while excluding variables like weight, diabetes history, and known allergies due to their limited impact on the target variable. Using RapidMiner, we implemented three regression models:

Random Forest, K-Nearest Neighbors (K-NN), and Linear Regression. Our findings indicate that the Random Forest model performed best with minimized root mean squared error (RMSE) and maximized R^2 when using 75 trees and excluding less relevant attributes. The results highlight the potential of regression models to provide insights into insurance pricing, revealing strong correlations between medical history and premium costs. We recommend leveraging Random Forest for this analysis due to its superior predictive performance and advocating for the use of such models to optimize insurance pricing, address disparities, and promote fairness in premium determination.