



# **PREDICTING STROKE RISK: A DATA ANALYTICS PROJECT**

**TEAM 1**

**Sun Afolabi,**

**CDIM 6308-70**

**PROFESSOR: DR. LIANG CHEN**

**SPRING 2020**

**WEST TEXAS A&M UNIVERSITY**

## Executive Summary

### Goal and Motivation

Strokes are a leading cause of serious, long-term disability in the United States and kill as many as 140,000 Americans every year (Centers for Disease Control and Prevention [CDC], 2020). Annually, more than 795,000 individuals in the United States will have a stroke, and over 75% of those strokes occur for the first time. Strokes can occur at any age; however, many people can decrease their risk of having a stroke by controlling existing medical conditions and making healthy lifestyle choices (CDC, 2020). In this analytics project, we seek to predict the likelihood that a given individual will have a stroke. Our predictions will be established through data mining models and based on several demographic and relational risk factors.

### Method and Data

Our project is in the healthcare analytics domain, and to accomplish our objectives, we used secondary data gathered through McKinsey Analytics and Analytics Vidhya. The two healthcare datasets were posted on Kaggle.com in April 2018 (Agarwal, 2018). On the original datasets, we conducted descriptive analytics and data visualization. After completing these steps, we prepared the data by removing variables we determined to be insignificant. We then imputed missing values with averages, created dummy variables, and balanced our target variable, “stroke”, using the oversampling technique performed through Microsoft Excel. To complete the predictive analytics component of the project, we used the training dataset to create a decision tree and logistic regression models in RapidMiner. Lastly, we applied logistic regression, the most accurate model on the prediction data set.

### Key Findings

We found that hypertension, heart disease, work type, residence type, smoking status, age, and average glucose level, have a positive and statistically significant predictive power over the likelihood that someone could have a stroke in the future. In contrast, BMI has zero predictive power for stroke while gender and marital status have a predictive power that is not significant at the 5% level.

### Conclusion

In conclusion, the decimation of human population by known diseases to unknown factors remains an object of concern. Certain factors are beyond the ken and comprehension of humans, but stroke, the 5<sup>th</sup> leading cause of death in the United States, is not an uncontrollable condition. Our predictive model can be deployed through a medical facility’s website or patient portal application; where users enter information, and instantly see their likelihood of having a stroke in the future. The output of our analysis will provide a useful metric for patients and medical practitioners to determine the need for any further medical and lifestyle modifications necessary to reduce stroke risk. By being cautious about those variables, the annual deaths occasioned by stroke will be significantly reduced. Additionally, with the increase in research and technology, the sensitivity to health advice will play a great role to stem the rising tide of stroke among the populace.

## **Introduction**

### **Business Understanding**

Strokes occur every 40 seconds in the United States killing more than 140,000 Americans each year. This makes it the fifth leading cause of death and disability in the United States (CDC, 2020). The Mayo Clinic states that a stroke is a clinical condition that results when there is a blockage or rupture of blood vessels that supply the brain. The clinical signs and symptoms will depend on the area of the brain that is affected by the lack of blood flow. It can lead to sudden weakness, numbness, or loss of vision on one or both sides of the body. There could also be an onset of altered speech, dizziness, double vision, and controlled body movement (Mayo Clinic, 2020). Strokes can cause major neurological and medical complications if not managed appropriately, but they can also be prevented altogether. Early action and prompt treatment can reduce brain damage and other complications (Hankey, 2017). Stroke reduces mobility in more than half of stroke survivors aged 65 and over, becoming the leading cause of serious long-term disability (CDC, 2020). Per annum, more than 795,000 individuals in the United States have a stroke, leading to an average health expenditure of 34 billion USD per year (Benjamin, et. al, 2017). This includes the cost of healthcare services, medications given to treat stroke, and missed days of work. Looking at these figures, it is apparent that strokes have substantial health and financial repercussions, affecting not only the individual survivor and their families, but also placing a huge burden on our entire healthcare system.

### **Problem to Solve**

The American Stroke Association (ASA) has classified high blood pressure (hypertension) as the leading cause of stroke and the most significant controllable risk factor (American Stroke Association, 2020). It is possible that other relational risk factors have a similar or greater impact on stroke prediction. Through this project, our aim is to identify other significant risk factors that contribute to an individual's likelihood of having a stroke in the future. According to the ASA, many risk factors that exacerbate or elevate the risk of developing a stroke have been identified. Risk factors such as hypertension (high blood pressure), smoking, diabetes (high blood sugar), poor diet, physical inactivity, obesity, hyperlipidemia (high cholesterol and triglycerides), carotid artery disease, peripheral artery disease, heart disease, atrial fibrillation, and sickle cell disease can be controlled through lifestyle modifications and/or pharmaceutical therapies. Non-controllable risk factors are age, family history, race, gender, prior stroke, or heart attack. However, they play an important role when combined with controllable risk factors (American Stroke Association, 2018).

### **Motivation**

According to the National Institutes of Health (NIH), many individuals survive a stroke, but may be left with disabilities that can often take weeks, months, or even years to overcome. Unfortunately, in some cases patients are left with permanent disabilities, requiring them to rely on their caregivers or family members for routine activities of daily living. Stroke survivors may be unable to return to their accustomed lifestyle, or perform their usual functions at work or at home. They may require constant therapies and medications to control their blood pressure and/or blood coagulation (NIH, n.d.). Despite medical efforts to control modifiable risk factors, once individuals have had a stroke, the chances of recurrence are elevated. One out of four stroke survivors have another stroke within five years (Longman & Jackson, 2017), so we feel it is

important to identify these correlational underlying conditions before a patient suffers a stroke. Additionally, the data provided by this analysis can ultimately help reduce the \$34 billion price tag associated with annual stroke care in the United States (Benjamin, et. al, 2017).

### **Opportunity/Challenge**

The goal of this project is to predict an individual's risk for stroke based on a number of known risk factors. By analyzing datasets containing thousands of records through data analysis in tools such as Excel, Tableau, and RapidMiner, we will highlight the number of stroke patients in the dataset, predicting which individuals have a higher risk of developing a stroke in the future. Being able to predict the likelihood that an individual will have a stroke can be an extremely useful warning tool for both healthcare providers as well as individuals. Once a person is deemed at risk for developing a stroke, certain modifications may be taken, such as dietary changes, beginning an exercise regime, adopting healthy lifestyle habits, and controlling other risk factors with medication, if warranted (American Stroke Association, 2018). Having insight may provide the individual with motivation to make changes that will reduce the likelihood of a stroke in their future and can lessen the burden on our healthcare system.

### **Brief Action Plan**

To achieve our goals, we will construct a research method and work to understand, describe, and prepare the healthcare data obtained from Kaggle. Once the data is prepared, we will use the training dataset to create decision tree and logistic regression models. Next, we will apply the best performing model on the prediction data set to predict the likelihood that a given individual in the prediction data set will have a stroke.

## **I. Research Method/ Design**

Our research method is a “classic machine learning procedure” that includes the scientific paradigm of induction and deduction (University of Eastern Finland, n.d.). In the inductive phase, we train and test two machine learning models on secondary datasets retrieved through evidence acquisition on Kaggle.com. The main goal during the inductive phase is to create models that learn general rules from our dataset; general rules which accept “positive examples” and reject “negative examples.” This process is called supervised machine learning. Given the structure of our dataset, we found decision tree and logistic regression models to be the most suitable. In the deductive phase (predictive phase), we will use the best performing model from the inductive phase to make predictions using new data (second dataset). A schema of our research method can be found in Figure 1. This schema includes data collection, data preparation, model building, model evaluation, and model deployment to make predictions.

Using descriptive analytics techniques, we conducted data analysis and data visualization to gain a better understanding of the individual attributes contained within our dataset. For instance, we discovered numerous amounts of missing or insignificant values. However, no outliers were identified in our data. Our unit of analysis apropos of this analytics project are individuals/patients. The target or dependent variable in our data analytics project is “stroke”; a dummy variable that takes on the value “1” for patients who had a stroke, and “0” for those who

did not have a stroke. Our predictor or independent variables are age, hypertension, heart disease, average glucose level, body mass index (BMI), smoking status, and gender as a control variable. Please refer to Table 1 for a complete description of our attributes. Through supervised machine learning we will identify the attributes that are important predictors of our target variable “stroke”, and then use those attributes to predict the likelihood that a given individual will have a stroke. Our domain is healthcare analytics. We are using descriptive analytics (D) to visualize and better understand the data and then using predictive analytics (P) to make a prediction. Our efforts will bring Innovation (I) and Agility (A) to healthcare practitioners, which will increase their Productivity (P). By developing a predictive model, healthcare providers can expeditiously and accurately predict which individuals are at an increased risk for stroke, allowing them to quickly provide recommendations. This, in turn, may reduce the physical and emotional impact to individuals, and their families, while concurrently reducing healthcare costs.

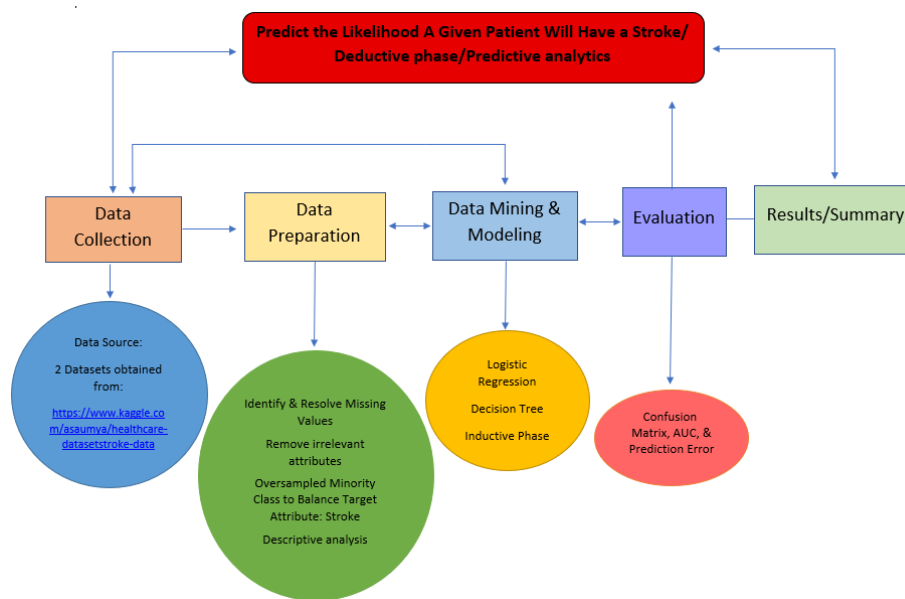


Figure 1: Predictive analytics schema

Variable	Definition	Measurement
ID	Patient ID	Unique ID
Marital Status	Patient marital status	Yes/No
Work Type	Patient work type	Private/Self-employed/ Govt. employed/Children
Residence Type	Patient residence type	Urban/Rural
Gender	Patient gender	Male/Female/Other
Age	Patient age	Number in Years
Hypertension	Patient hypertension status	0/1
Heart Disease	Patient heart disease status	0/1
Average Glucose Level	Patient average glucose level	Glucose level of the patient (mg/dL)
Body Mass Index (BMI)	Patient body mass index	Body fat based on height and weight (Kg/m2)
Smoking Status	Patient smoking status	Never/Formerly/Currently Smokes
Stroke	Patient stroke status	0/1

Table 1: Variable definition and measurement

## II. Data Description

Among the two datasets, all attributes include “id”, and 9 independent variables: “age”, “hypertension”, “heart\_disease”, “ever\_married”, “work\_type”, “Residence\_type”, “avg\_glucose\_level”, “bmi”, “smoking\_status”, and one control variable: “gender”. There is one target variable: “stroke” which is present in the training set, but not in the prediction set. Our training dataset contains 43,400 records, and our prediction dataset contains 18,601 records. The attributes can be visualized in Table 2.

Category	Attribute Name	Description	Data Type
Demographic Information	id	A unique ID to identify each patient	Numerical
	ever_married	Marital Status of the Patient	Categorical/Binomial
	work_type	The type of employment of the patient	Categorical/Polynomial
	Residence_type	Whether the patient lives in an urban or rural area	Categorical/Binomial
	gender	The gender of each patient	Categorical/Binomial
Risk Factors	age	Age of the patient	Discrete/ Integer
	hypertension	Whether or not the patient suffers from hypertension	Categorical/Binomial
	heart_disease	Whether or not the patient suffers from heart disease	Categorical/Binomial
	avg_glucose_level	The average glucose level the patient had when it was measured	Numerical/Continuous
	bmi	Measure of body fat based on height and weight (Kg/m2)	Numerical/Continuous
	smoking_status	Whether the patient is a current smoker, or has smoked before, or has never smoked.	Categorical/Polynomial
Target	stroke	Whether the patient has suffered a stroke or not	Categorical/Binomial

Table 2: Training dataset attributes

## III. Data Preparation

### Original Data Set Descriptive Analysis

We discovered that the original dataset contained missing values for the variables “smoking\_status” and “bmi” (see Figure 2). There were 1,462 records (3.4%) with missing BMI values, while in total there were 13,292 records (30.6%) with missing values. In addition, the proportion of “other\_gender” (0.025%) when compared to “male\_gender” and “female\_gender” represents a very small percentage. Drilling down further, the percentage of “other\_gender” vs “male\_gender” is 0.062% while “other\_gender” vs “female\_gender” is 0.043%. Patients aged 17 and younger, reported “work\_type” as “children”, likely meaning that they are underage and unemployed. Those 7,541 values represent 17.4% of the observations. Also, 81.1% of the individuals 17 years old and younger, have missing “smoke\_status” values. We find this logical given we would not expect many children to smoke. Another issue with the original datasets is that they contain several categorical variables such as “work\_type”, “Residence\_type”, “smoking\_status”, and “ever\_married”, which the logistic regression cannot handle because they are polynomials. Finally, the most complicated issue with the datasets is that the training dataset’s target variable “stroke” is imbalanced. Only 1.8% of the individuals in the dataset have reported a stroke.



Figure 2: attribute: bmi, smoking\_status & stroke in Tableau

## Data Preprocesses

According to Press (2016), data preparation is the most time consuming and least enjoyable part of data analytics accounting for 80% of the work performed by data scientists. However, it is considered an important process because the reliability and accuracy of the analysis depend, to a large extent, on the quality of the prepared data used to train the models. Thus, we carefully conducted our data cleansing and munging in a few steps. First, we transformed the CSV file into an Excel worksheet, and then removed all individuals aged 17 and younger. They originally comprised 17.4% of the “age” variable, and 81.1% of those children were missing “smoking\_status” values. Studies show that only .003% to 0.025% of children are likely to have a stroke (Ferriero, et al. 2019). After considering the infrequency of child stroke cases, we felt removing children aged 17 and younger, in addition to work type “children”, was appropriate. This decreased our records by 7,541 rows. Next, we removed the work type “never\_worked” since it only represents 0.41% of the work\_type attribute. In addition, the percentage of “other” gender comprised only 0.00025% of total genders reported. Due to “male” and “female” making up 40.84% and 59.14% of “gender” respectively, rows containing “other” were dropped from the dataset. Furthermore, missing “smoking\_status” records were removed. After dropping missing “smoking\_status” records, we then transferred the data to RapidMiner Turbo Prep and replaced missing “bmi” values by imputing the average BMI of patients 18 and older. All these processes decreased our example set records to 28,641.

The main issue we identified with the original dataset was the major imbalance of our target variable, “stroke”. The number of patients who reported having a stroke made up a small 1.8% of the attribute stroke, while 98.2% of patients did not report a stroke. We resolved this issue by bringing the dataset back to Excel and randomly oversampling the minority class. This resulted in a balanced target variable with 28,003 patients not reporting a stroke, and 28,003 patients reporting a stroke. Finally, we discovered another complication with the dataset when attempting to run logistic regression with categorical variables such as “work\_type”, “Residence\_type”, “smoking\_status”, “ever\_married”, and “gender”. We converted these categorical variables to dummy binomial variables to utilize in training the logistic regression. We applied all the above data preparation actions on both the example set and the prediction set, except the balancing of the target variable since our prediction dataset does not contain the target variable. Please refer to Figure 3 for a complete schedule of our data preparation processes.

		TRAIN				PREDICT			
		Rows	Columns	stroke YES	stroke NO			Rows	Columns
	original data	43,400	12	783	42,617			18,601	11
		(7,541)	-	(2)	(7,539)	CHANGE	MADE	(3,242)	-
	removed age 17 and under	35,859	12	781	35,078			15,359	11
		-	-	-	-	CHANGE	MADE	-	-
	removed all children job types	35,859	12	781	35,078			15,359	11
		(58)	-	-	(58)	CHANGE	MADE	(31)	-
	removed never worked	35,801	12	781	35,020			15,328	11
		(9)	-	-	(9)	CHANGE	MADE	-	-
	remove other gender	35,792	12	781	35,011			15,328	11
		(7,151)	-	(143)	(7,008)	CHANGE	MADE	(3,086)	-
EXCEL	removed missing smoke status	28,641	12	638	28,003			12,242	11
		-	-	-	-	CHANGE	MADE	-	-
RAPIDMINER TURBOPREP	bmi Impute missing with average	28,641	12	638	28,003			12,242	11
		27,365	-	27,365	-	CHANGE	MADE	-	1
EXCEL	Over sampling	56,006	12	28,003	28,003	Added Stroke		12,242	12
		-	8	-	-				
RAPIDMINER TURBOPREP	Create dummy variables	56,006	20	28,003	28,003				
	Data (loss)/gain	12,606	8	27,220	(14,614)			(6,359)	1
		29%	67%	3476%	-34%			-34%	9%

Figure 3: Data preparation processes

## IV. Results and Findings

### Descriptive Statistics

Our original dataset was carefully prepared to ensure that when the oversampling was carried out to balance our data. We would first amplify the target variable “stroke” and then closely match the percentages of the values of the corresponding attribute in the original dataset. We ended up with 56,006 records in our balanced dataset. A summary of the descriptive statistics for both datasets can be seen in Appendix 1.

### Decision Tree Results

Appendix 2 depicts, the attributes weight plot of the decision tree, which shows that avg\_glucose\_level has the largest weight at 0.520, followed by bmi with a weight of 0.359. This demonstrates that average\_glucose level is the most important factor for stroke while the second most important factor is bmi. In contrast, smoking\_status and heart\_disease are the least important factors for stroke with a weight of 0.003 and 0.011 respectively. Surprisingly, our decision tree output shows that hypertension, age, and gender have a moderate importance for stroke with a weight of 0.047, 0.036, and 0.023 respectively.



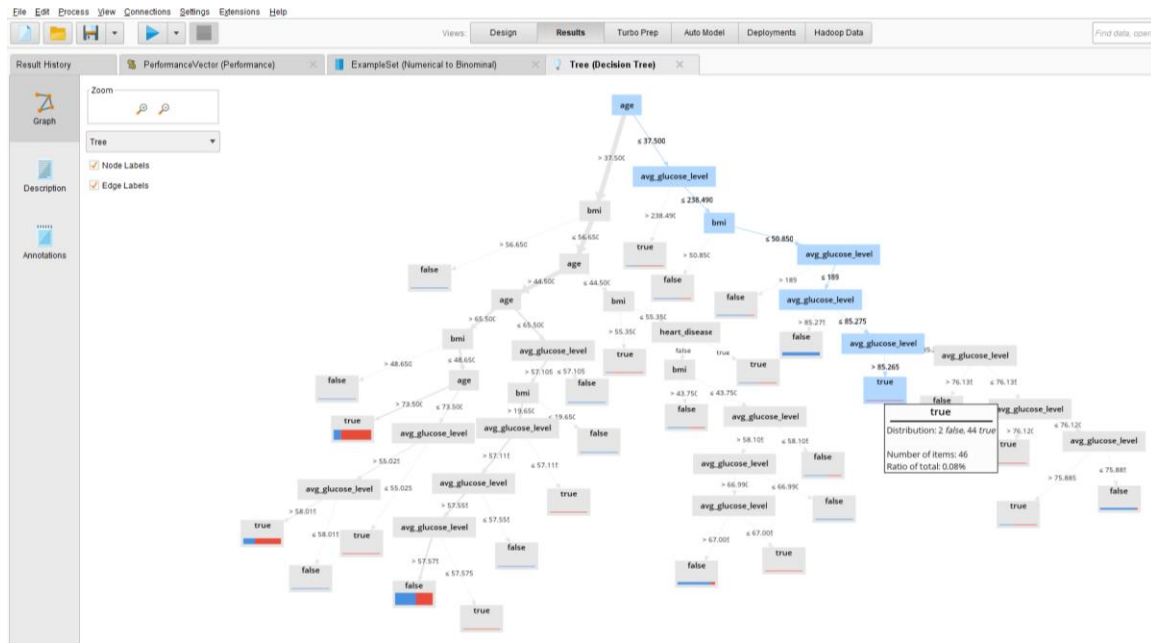


Figure 4: Decision Tree Model

Figure 4 depicts the tree of the decision tree model. We can see that age is the root node, but it also appears to be a split node at some other parts of the tree. According to the decision tree, a patient with an average glucose level higher than 85.3, but less than 238.5, with a bmi less than or equal to 50.1, who is not older than 37.5 years, has a higher probability of having a stroke. To be precise, that patient has a 95.7% chance of developing a stroke  $((44/46)*100)$ . However, only 46 patients fall into that leaf node which is approximately 0.08% of our sample size. Likewise, a patient who is older than 73 years and has a bmi of at most 48.7, has a high probability of getting a stroke. This probability is 80.6%, which is derived as follows:  $((2992 + 12408)/15400)*100$ . This probability is the percentage of patients with a stroke in the leaf node. This specific leaf node accounts for 27.5% of our sample size. Though this leaf represents a large portion of our sample size, it is not the largest leaf node. The largest leaf node accounts for 33.9% of our sample size.

## Logistic Regression Results

Appendix 3 depicts the attributes weight plot of the logistic regression model. Given the plot we observe that age has the highest weight at 1.318, followed by heart\_disease=true at 0.701, hypertension=true at 0.436, and avg\_glucose\_level at 0.202. This means that age, heart disease, hypertension and average glucose level are the most important variables when using a logistic regression to predict the probability that a patient will have a stroke. Moreover, smoking\_status=smokes, smoking\_status=formerly\_smoked, and residency\_type=urban have moderate to low predictive power for stroke with weights of 0.159, 0.08, and 0.051 respectively. Finally, heart\_disease=false, and hypertension=false have no predictive power for stroke with 0 weight.

Warning: Removed collinear columns [work\_type = Gov\_Job.true, smoking\_status = smokes.true]

Attribute	Coefficient	Std. Coefficient	Std. Error	z-Value	p-Value
hypertension.true	0.431	0.431	0.026	16.497	0
heart_disease.true	0.646	0.646	0.033	19.505	0
gender = Female.true	-0.036	-0.036	0.021	-1.692	0.091
ever_married = Yes.true	-0.031	-0.031	0.034	-0.902	0.367
work_type = Private.true	0.152	0.152	0.031	4.963	0.000
work_type = Self-employed.true	0.182	0.182	0.035	5.243	0.000
work_type = Gov_Job.true	0	0	?	?	?
Residence_type = Urban.true	0.106	0.106	0.021	5.135	0.000
smoking_status = never smoked.true	-0.388	-0.388	0.027	-14.301	0
smoking_status = formerly smoked.true	-0.221	-0.221	0.029	-7.567	0.000
smoking_status = smokes.true	0	0	?	?	?
age	0.074	1.313	0.001	91.447	0
avg_glucose_level	0.004	0.214	0.000	19.899	0
bmi	-0.000	-0.002	0.002	-0.198	0.843
Intercept	-4.957	-0.146	1.691	-2.931	0.003

-4.957123704795426

Figure 6: Logistic Regression model

Figure 6 shows the deployment model of the logistic regression. We observe that among the independent variables, heart\_disease.true and hypertension.true have the largest coefficients at 0.646, and 0.431 respectively. In addition, these two attributes are statistically significant at the 0.05 level with a p\_value of zero. This means that people who had heart disease or hypertension in the past have a higher probability of having a stroke compared to people who did not have heart disease or hypertension. The difference between those who had heart disease and those who had not is 0.646, which is significant at the 5% level. Likewise, the difference between those who had hypertension and those who did not is 0.431 which is also positive and significant at the 5% level. Moreover, age and avg\_glucose\_level also have positive and significant coefficients at the 5% level, meaning that age and avg\_glucose\_level are also significant predictors of stroke. Older people tend to have a higher risk of getting stroke compared to younger people. More tellingly, getting older by one additional year increases your odds of stroke by 0.074, all other things being equal. Also, the higher your avg\_glucose\_level the higher the chance that you might get a stroke. If you are average glucose level increases by 1 unit, your odds of getting a stroke also increase by 0.004, all other things being equal.

Furthermore, Residence\_type=Urban.true has a coefficient of 0.106, which is significant at the 5% level with a p\_value of zero. This means that residence\_type=Urban.true is a significant predictor of stroke. Thus, people living in cities have a higher likelihood of getting stroke compared to those who live in rural areas. More tellingly, the positive coefficient of 0.106 represents the odds differences between those who live in urban settings and those who live in a rural area. Likewise, the attribute work\_type=private.true has a positive and significant coefficient at the 5% level with a p\_value of zero. This means that work\_type=private.true is a significant predictor of stroke. Strongly inferring that people working in the private sector have higher odds of developing a stroke compared to people who have a government job. The coefficient of 0.152 represents the odds differences between those who work in the private sector and those who work for the government all other things being equal. Next, the attribute work\_type=self-employed.true also has a positive coefficient of 0.182, which is statistically significant at 5% level. This coefficient represents the difference in the odds of getting a stroke between people who work for themselves and people who work for the government. This means

that the likelihood of getting a stroke is higher for self-employed people compared to those who work for the government, controlling for other factors.

Another interesting output of the logistic regression is that the attribute `gender=Female.true` has a negative coefficient of -0.036, meaning that males have higher odds of getting a stroke when compared to females. The coefficient -0.036 means that the odds of a female developing a stroke is 0.036 less than that of a male getting a stroke, controlling for other factors. However, this coefficient is not significant at the 5% level with a `p_value` of 0.091, meaning that there might not be a significant difference between males and females when it comes to the risk of getting a stroke. Likewise, there might not be a significant difference between married people and those who have never been married when it comes to the odds of getting a stroke since the coefficient of -0.031 is not statistically significant at the 5% level. The same applies to `bmi` that has a coefficient of -0.000, which is not significant at the 5% level with a `p_value` very high at 0.843. This means that `bmi` has no predictive power for the odds that someone will get a stroke. Indicating that people with a low `bmi` have the same odds of getting a stroke as those with a high `bmi`, controlling for other factors.

Finally, the attributes `smoking_status=never smoked.true` and `smoking_status=formerly smoked.true` have coefficients of -0.388 and -0.221 respectively, which are significant at the 5% level, meaning that smoking status has a significant predictive power of stroke. Moreover, the coefficient -0.388 means that a patient who never smoked has a lower likelihood of getting a stroke compared to those who currently smoke. The difference of the odds of getting a stroke between those who have never smoked and those who currently smoke is -0.388. Likewise, the patients who formerly smoked also have lower odds of getting a stroke compared to those who currently smoke. The difference in their odds is -0.221, all other things being equal. It is also worth mentioning that the intercept has no importance in this case.

## Performance

Decision Tree	Logistic Regression
Sensitivity 64.6%	Sensitivity 77.61%
Specificity 80.25%	Specificity 71.83%
AUC 0.795	AUC 0.819
Error 27.58 %	Error 25.28 %

Table 3: Performance (Note: see Appendix 4 for the performance screenshots.)

## Prediction

Given the performance visualized in Table 3, we decided to use logistic regression to make our predictions as it is the best performing model. The logistic regression model outperformed the decision tree model in sensitivity, AUC (area under the curve), and prediction error while the decision tree performed better in specificity only. Therefore, we chose the logistic regression model since we are most interested in finding the people who are most likely to get a stroke in the future. The lower prediction error, the higher sensitivity score, and the largest AUC score make the logistic model a clear winner over the decision tree.

## Prediction Results

We applied our logistic regression model on the prediction data set and collected the results depicted in Appendix 5 below. We predicted that out of the 12,242 patients, 3,552 have a high likelihood of having a stroke, which is 29% of the prediction sample size (see Figure 8). For instance, the patient with ID number 36306 who is a male, has never had hypertension, never had a heart disease, who is or has been married, has a BMI of 21.1, has an average glucose level of 83.8, is 80 years old, has formerly smoked, lives in an urban area, and works in the private sector, the model predicted that this patient will have a stroke with a probability as high as 80%. A second illustration could be the patient with the id number 61822. This patient has a BMI of 26, an average glucose level of 179.5, is a 74-year-old female, has formerly smoked, lives in a rural area, she is self-employed, is married or formerly married, and has heart disease, but has never had hypertension. The model predicted that this patient has a 90% chance of having a stroke.

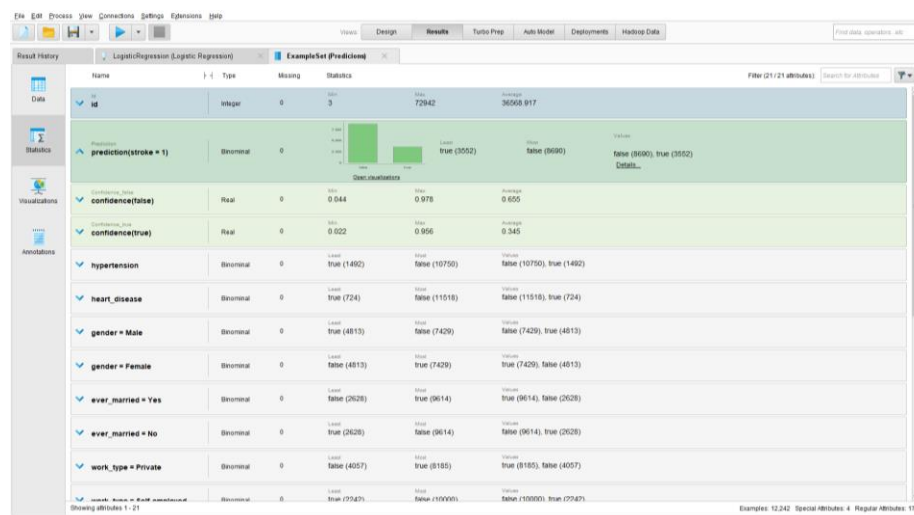


Figure 8: Logistic Regression prediction example set

## Findings

We found that hypertension, heart disease, work type, residence type, smoking status, age, and average glucose level have a positive and statistically significant predictive power over a patient's likelihood in developing a stroke in the future. For instance, we found that all other things being equal, patients who previously had hypertension have an advanced risk of developing stroke when compared to individuals who have never had hypertension. Likewise, patients who have heart disease have an increased risk of stroke when compared to those who have never developed heart disease. Also, the higher your age, the higher your probability of developing a stroke. Similarly, the higher your average glucose level, the higher your risk is for developing stroke, provided circumstances remain the same. Indicative that keeping a patient's average glucose level lower can consequently help decrease their likelihood of developing a stroke. Likewise, people who have never smoked or formerly smoked have a lesser probability of developing a stroke when compared to people who currently smoke. Indicating that if the circumstances remain unchanged, smoking cessation can decrease an individual's risk for

developing a stroke. Even to a lesser extent, living in a rural area can lessen your chances of developing a stroke all other things being equal. Likewise, working for the government is a safer way to lower your risk for developing a stroke when compared to people who work in the private sector or those that are self-employed, all other things being equal. In contrast, BMI has zero predictive power for stroke, while gender and marital status have a predictive power that is not significant at the 5% level. This means that the difference in the odds of developing a stroke between male and females, or between married people and people who were never married, happens by chance all other factors being constant.

## **V. Discussion**

According to Almadani and Alshammari (2018), predictive modeling for disease risk using data mining techniques is feasible because there is a lot of data generated in the healthcare industry. Supervised data mining methods can be used to predict the risk for a target (disease) such as stroke. These models can then be run on the patients of interest. The algorithm determines which patients have attributes that match the patterns of the training set patients who either suffered a stroke or not and can then make a prediction (Almadani & Alshammari, 2018). Logistic regression is appropriate to use for this analysis since the target variable “stroke” is binary. A patient either has a stroke or not. The results from our analysis showed that the logistic regression model had a higher sensitivity (77.61% vs 64.6%) in comparison to the decision tree model. Providers can then be assured as to the accuracy of our model, better enabling its deployment on a website or through a mobile application. Our analysis determined the following controllable risk factors were significant predictors: hypertension, smoking status, heart disease and average glucose level. Our model will then identify those at-risk individuals relative to their controllable risk factors, flagging them as someone who needs to be followed up on for early intervention (Amini et al, 2013). In medical diagnostic systems, even a small difference in accuracy is important since the correct prediction of the risk of serious disease such as stroke can make a significant change in a person's life.

## **VI. Limitations/Areas for Future Improvement**

Our research was originally limited due to the imbalanced data received from Kaggle. To improve the number of records with the target variable "stroke", we used oversampling techniques in Excel. The details were provided in the 'Data Preparation' section. The team found that only 1.8% of the 'stroke' attribute had the value 1 and that there was as much as 31% of the values missing for the attribute 'smoking\_status'. Both issues, along with the fact we are using secondary data, would have impacted the accuracy of the prediction. Even though we did not control for ethnicity in this research, other studies show that stroke is more common among black patients (Howard, 2013). We encourage further research on predictive modeling for stroke risk. We suggest using primary data that takes ethnicity into account. This information will provide a broader picture of patients at risk for stroke and allow for better predictive models to be created. We also suggest carrying out longitudinal studies on the research subjects so data can be observed over the course of several years to assess the actual outcome of earlier prediction. In addition, advanced techniques such as Support Vector Machine, Neural Network, Bayesian Classifier, Stochastic Gradient boosting and Penalized Logistic Regression can also be applied to stroke risk prediction in order to improve accuracy.

## References

- Agarwal, S. (2018, April 16). Healthcare Dataset Stroke Data. Retrieved from <https://www.kaggle.com/asaumya/healthcare-dataset-stroke-data>
- Almadani, Ohoud & Alshammari, Riyad. (2018). Prediction of Stroke using Data Mining Classification Techniques. *International Journal of Advanced Computer Science and Applications(ijacsa)*, 9(1), 2018.
- American Stroke Association. (2020). Stroke Risk Factors. Retrieved from <https://www.stroke.org/en/about-stroke/stroke-risk-factors>
- American Stroke Association. (2018, October 10). Stroke Risk Factors You Can Control, Treat and Improve. Retrieved from <https://www.stroke.org/en/about-stroke/stroke-risk-factors/stroke-risk-factors-you-can-control-treat-and-improve>
- Amini, Leila & Azarpazhouh, Reza & Farzadfar, Mohammad & Mousavi, Sayed & Jazaieri, Farahnaz & Khorvash, Fariborz & Norouzi, Rasul & Toghianfar, Nafiseh. (2013). Prediction and Control of Stroke by Data Mining. *International journal of preventive medicine*. 4. 245-249.
- Benjamin, E. J., Balaha, M. J., Chiuve, S. E., et. al. (2017). *American Heart Association*, 135(10), 146–603. Retrieved from <https://www.ahajournals.org/doi/epub/10.1161/CIR.0000000000000485>
- Centers for Disease Control and Prevention. (2020, January 31). Stroke Facts. Retrieved from <https://www.cdc.gov/stroke/facts.htm>
- Ferriero, Donna M., et al. "Management of Stroke in Neonates and Children: A Scientific Statement From the American Heart Association/American Stroke Association." *Management of Stroke in Neonates and Children: A Scientific Statement From the American Heart Association/American Stroke Association*, vol. 50, no. 3, 2019, doi:10.1161/str.0000000000000183.
- Hankey, Graeme J. "Stroke." *The Lancet* 389.10069 (2017): 641-54. Web
- Howard, V. J. (2013). Reasons Underlying Racial Differences in Stroke Incidence and Mortality. *Stroke*, 44(6, Supplement 1). doi: 10.1161/strokeaha.111.000691
- Longman, C., & Jackson, J. (2017, May 12). One in three stroke survivors have another attack – the NHS could do more to prevent it. Retrieved from <https://theconversation.com/one-in-three-stroke-survivors-have-another-attack-the-nhs-could-do-more-to-prevent-it-77196>
- Mayo Clinic. (2020, February 7). Stroke. Retrieved from <https://www.mayoclinic.org/diseases-conditions/stroke/diagnosis-treatment/drc-20350119>
- Mayo Clinic. (2020). What a Stroke Looks Like. Retrieved from <https://www.mayoclinichealthsystem.org/get-stroke-ready/what-a-stroke-looks-like>

NIH. (n.d.). Stroke. Retrieved from <https://www.nhlbi.nih.gov/health-topics/stroke>

Press, G. (2016, March 23). Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says. Retrieved from <https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#722588936f63>

University of Eastern Finland. (n.d.). Machine learning. Retrieved from <http://www.cs.joensuu.fi/~whamalai/skc/ml.html>

## Appendices

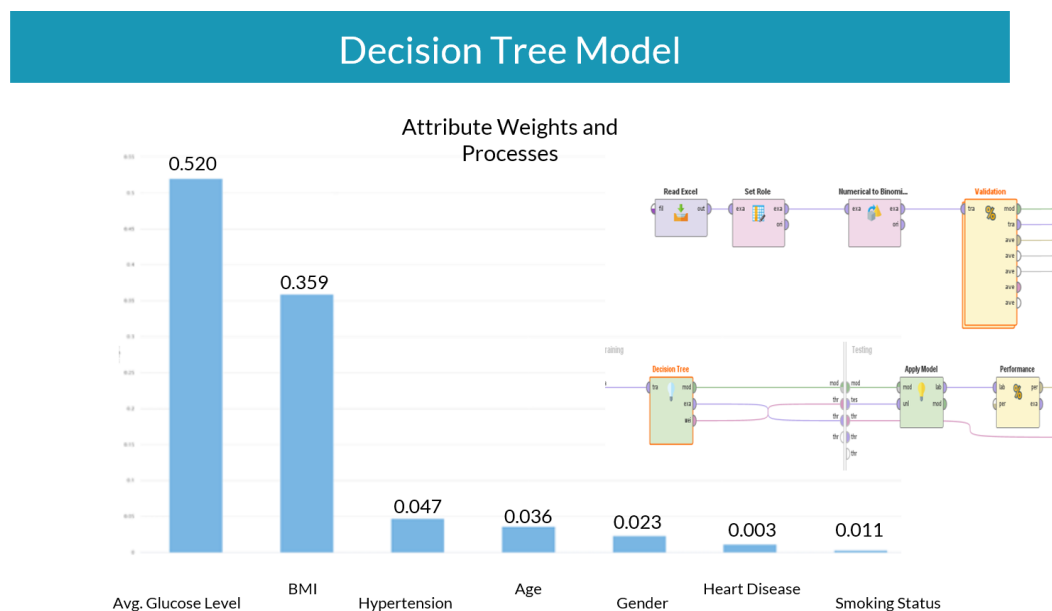
### Appendix 1. Numerical variables descriptive statistics

	<i>age</i> (Training)	<i>age</i> (Validation)	<i>avg_glucose_level</i> (Training)	<i>avg_glucose_level</i> (Validation)	<i>bmi</i> (Training)	<i>bmi</i> (Validation)
<b>Mean</b>	59	49.5	119	108	30	30.35
<b>Median</b>	61	50	97	92.8	30	29.5
<b>Mode</b>	79	78	75	85.7	30	30.35
<b>Std Deviation</b>	18	17.7	55	46.4	7	7.02
<b>Minimum</b>	18	18	55	55	10	12.7
<b>Maximum</b>	82	82	291	275.7	92	88.3

	TRAINING DATASET			VALIDATION DATASET		
Stroke	Yes ( 50%)	No (50%)		?	?	
Hypertension	Yes( 20 %)	No(80%)		Yes (12%)	No(88%)	
Heart_Disease	Yes (14 % )	No (86 %)		Yes (6%)	No (94%)	
Gender	Male( 41%)	Female (59%)		Male (39%)	Female (61%)	
Ever_Married	Yes (84 %)	No (16 %)		Yes (79%)	No (21%)	
Residence_Type	Rural (49.4%)	Urban (50.6%)		Rural (50%)	Urban (50%)	
Work_Type	Private (62%)	Self (25%)	Govt (13%)	Private (67%)	Self (18%)	Govt (15%)
Smoking_Status	Former (29.95% )	Never (48.29%)	Current (21.76%)	Former (26%)	Never (52%)	Current 22%

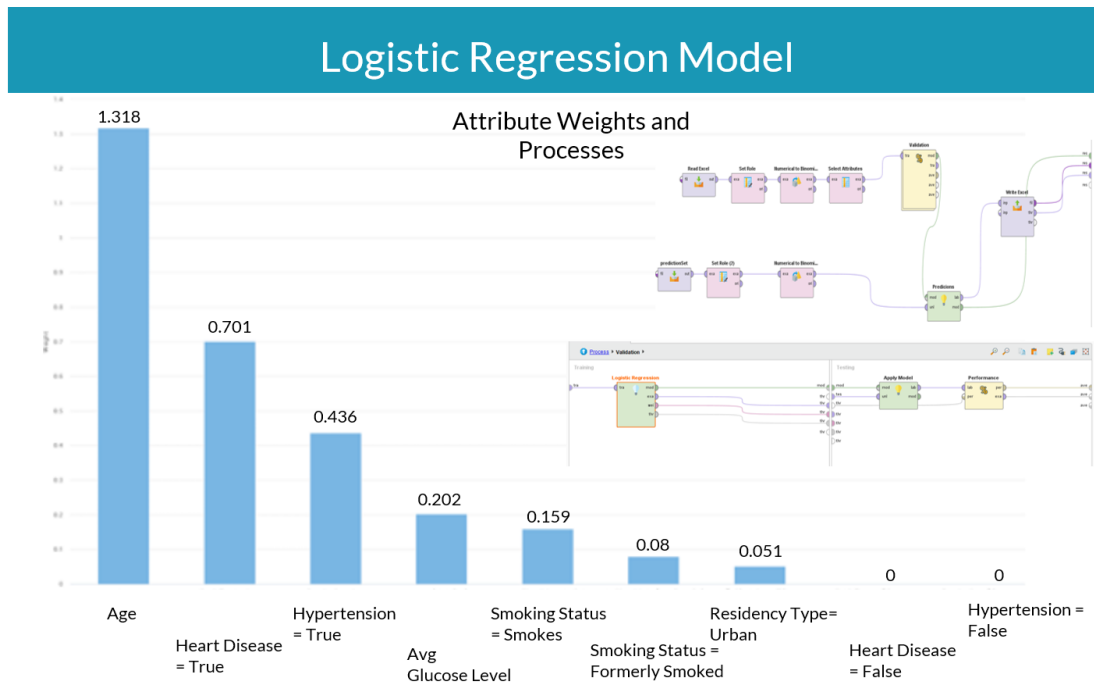
Table 4. Categorical variables descriptive statistics

### Appendix 2: Decision tree attribute weights

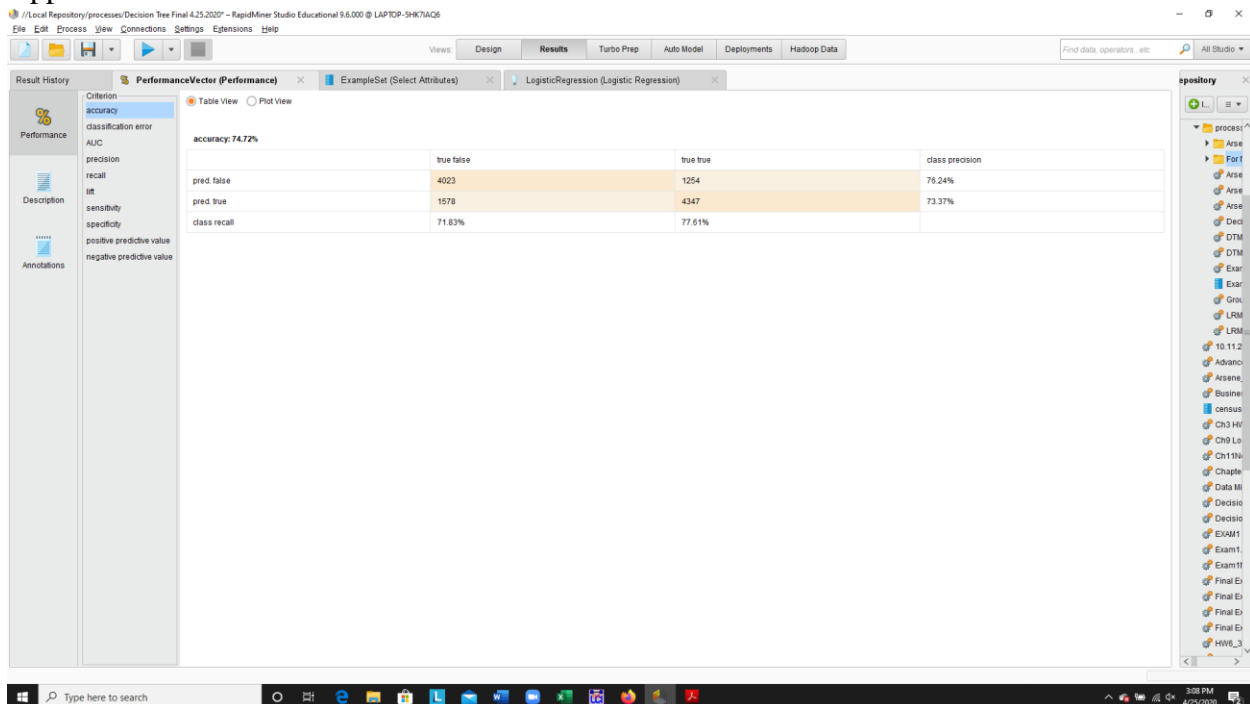


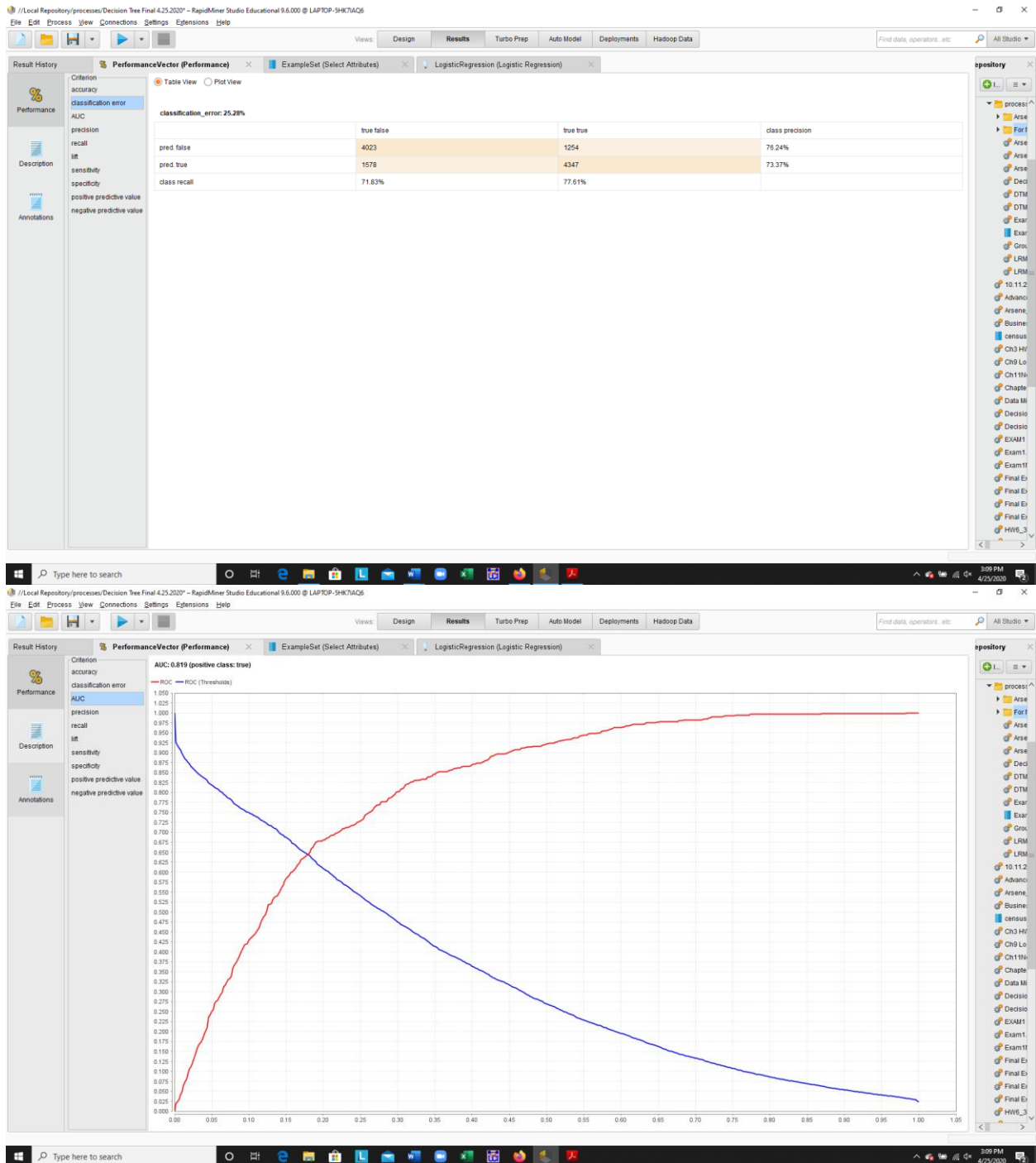


## Appendix 3: Logistic Regression Attribute weights



## Appendix 4: Performance screenshots





## 18

A horizontal Windows taskbar at the bottom of the screen. On the left is the Start button (Windows logo). Next to it is a search bar with the placeholder text "Type here to search". To the right of the search bar are several application icons: Internet Explorer, File Explorer, Microsoft Store, Outlook, Word, PowerPoint, Edge, Chrome, and Firefox. Further right are icons for network status, volume, and battery. On the far right, the system clock shows the time as 3:09 PM and the date as 4/25/2020.

Local Repository/process/Decision Tree Final 4.25.2020 - RapidMiner Studio Educational 5.6.000 @ LAPTOP-3HK7IAQ6

File Edit Process View Connections Settings Extensions Help

Views: Design Results Turbo Prep Auto Model Deployments Hadoop Data

Find data, operators, etc. All Studio

Result History PerformanceVector (Performance) ExampleSet (Select Attributes) LogisticRegression (Logistic Regression)

Criterion: accuracy, classification error, AUC, precision, recall, lift, sensitivity, specificity, positive predictive value, negative predictive value

Performance

Table View Plot View

lift: 146.73% (positive class: true)

	true false	true true	class precision
pred false	4023	1254	76.24%
pred true	1578	4347	73.37%
class recall	71.83%	77.61%	

sensitivity: 77.61% (positive class: true)

	true false	true true	class precision
pred false	4023	1254	76.24%
pred true	1578	4347	73.37%
class recall	71.83%	77.61%	

Repository: process, Arse, For, Arse, Arse, Arse, Dec, DTM, DTM, Exar, Exar, Gro, LRM, LRM, 10.11.2, Advanc, Arsene, Busine, versus, Cn3 HV, Cn3 Lo, Cn1 Hn, Chape, Data M, Decisio, Decisio, Ex/AM1, Exam1, Exam1f, Final E, Final E, Final E, Final E, HW6\_3

Type here to search

Local Repository/process/Decision Tree Final 4.25.2020 - RapidMiner Studio Educational 5.6.000 @ LAPTOP-3HK7IAQ6

File Edit Process View Connections Settings Extensions Help

Views: Design Results Turbo Prep Auto Model Deployments Hadoop Data

Find data, operators, etc. All Studio

Result History PerformanceVector (Performance) ExampleSet (Select Attributes) LogisticRegression (Logistic Regression)

Criterion: accuracy, classification error, AUC, precision, recall, lift, sensitivity, specificity, positive predictive value, negative predictive value

Performance

Table View Plot View

sensitivity: 77.61% (positive class: true)

	true false	true true	class precision
pred false	4023	1254	76.24%
pred true	1578	4347	73.37%
class recall	71.83%	77.61%	

Repository: process, Arse, For, Arse, Arse, Arse, Dec, DTM, DTM, Exar, Exar, Gro, LRM, LRM, 10.11.2, Advanc, Arsene, Busine, versus, Cn3 HV, Cn3 Lo, Cn1 Hn, Chape, Data M, Decisio, Decisio, Ex/AM1, Exam1, Exam1f, Final E, Final E, Final E, Final E, HW6\_3

Type here to search

309 PM 4/25/2020

Local Repository/processes/Decision Tree Final 4.25.2020 - RapidMiner Studio Educational 5.6.000 @ LAPTOP-3HK7IAQ6

File Edit Process View Connections Settings Extensions Help

Views: Design Results Turbo Prep Auto Model Deployments Hadoop Data

Find data, operators, etc. All Studio

Result History PerformanceVector (Performance) ExampleSet (Select Attributes) LogisticRegression (Logistic Regression)

Criterion accuracy classification error AUC precision recall lift sensitivity specificity positive predictive value negative predictive value

Performance

Table View Plot View

specificity: 71.83% (positive class: true)

	true false	true true	class precision
pred false	4023	1254	76.24%
pred true	1578	4347	73.37%
class recall	71.83%	77.61%	

positive\_predictive\_value: 73.37% (positive class: true)

	true false	true true	class precision
pred false	4023	1254	76.24%
pred true	1578	4347	73.37%
class recall	71.83%	77.61%	

Repository

process  
Arse  
For  
Arse  
Arse  
Arse  
Deci  
DTM  
DTM  
Exar  
Exar  
Gro  
LRM  
LRM  
10.11.2  
Arse  
Arse  
Busine  
versus  
CNS3 HV  
CNS Lo  
CNS HV  
Chaps  
Data MI  
Decisio  
Decisio  
Ex/AMI  
Exam1  
Exam1f  
Final E  
Final E  
Final E  
Final E  
Final E  
HW6\_3

Type here to search

Local Repository/processes/Decision Tree Final 4.25.2020 - RapidMiner Studio Educational 5.6.000 @ LAPTOP-3HK7IAQ6

File Edit Process View Connections Settings Extensions Help

Views: Design Results Turbo Prep Auto Model Deployments Hadoop Data

Find data, operators, etc. All Studio

Result History PerformanceVector (Performance) ExampleSet (Select Attributes) LogisticRegression (Logistic Regression)

Criterion accuracy classification error AUC precision recall lift sensitivity specificity positive predictive value negative predictive value

Performance

Table View Plot View

positive\_predictive\_value: 73.37% (positive class: true)

	true false	true true	class precision
pred false	4023	1254	76.24%
pred true	1578	4347	73.37%
class recall	71.83%	77.61%	

Repository

process  
Arse  
For  
Arse  
Arse  
Arse  
Deci  
DTM  
DTM  
Exar  
Exar  
Gro  
LRM  
LRM  
10.11.2  
Arse  
Arse  
Busine  
versus  
CNS3 HV  
CNS Lo  
CNS HV  
Chaps  
Data MI  
Decisio  
Decisio  
Ex/AMI  
Exam1  
Exam1f  
Final E  
Final E  
Final E  
Final E  
Final E  
HW6\_3

Type here to search

3:09 PM 4/25/2020

Local Repository/processes/Decision Tree Final 4.25.2020 - RapidMiner Studio Educational 5.6.000 @ LAPTOP-3HK7IAQ6

File Edit Process View Connections Settings Extensions Help

Views: Design Results Turbo Prep Auto Model Deployments Hadoop Data

Find data, operators, etc. All Studio

Result History PerformanceVector (Performance) ExampleSet (Select Attributes) LogisticRegression (Logistic Regression)

Criterion: accuracy, classification error, AUC, precision, recall, lift, sensitivity, specificity, positive predictive value, negative predictive value

negative\_predictive\_value: 76.24% (positive class: true)

	true false	true true	class precision
pred false	4023	1254	76.24%
pred true	1578	4347	73.37%
class recall	71.83%	77.61%	

repository

processes: Arse, For, Arse, Arse, Arse, Dec, DTM, DTM, Exar, Exar, Gro, LRM, LRM, 10.11.2, Advanc, Arsene, Busine, versus, Cn3 HV, Cn3 Lo, Cn1 Hn, Chape, Data M, Decisio, Decisio, Ex/AM1, Exam1, Exam1f, Final E, Final E, Final E, Final E, HW6\_3

Type here to search

Local Repository/processes/Decision Tree Final 4.25.2020 - RapidMiner Studio Educational 5.6.000 @ LAPTOP-3HK7IAQ6

File Edit Process View Connections Settings Extensions Help

Views: Design Results Turbo Prep Auto Model Deployments Hadoop Data

Find data, operators, etc. All Studio

Result History PerformanceVector (Performance) ExampleSet (Select Attributes) LogisticRegression (Logistic Regression)

Criterion: accuracy, classification error, AUC, precision, recall, lift, sensitivity, specificity, positive predictive value, negative predictive value

negative\_predictive\_value: 76.24% (positive class: true)

	true false	true true	class precision
pred false	4023	1254	76.24%
pred true	1578	4347	73.37%
class recall	71.83%	77.61%	

repository

processes: Arse, For, Arse, Arse, Arse, Dec, DTM, DTM, Exar, Exar, Gro, LRM, LRM, 10.11.2, Advanc, Arsene, Busine, versus, Cn3 HV, Cn3 Lo, Cn1 Hn, Chape, Data M, Decisio, Decisio, Ex/AM1, Exam1, Exam1f, Final E, Final E, Final E, Final E, HW6\_3

Type here to search

Local Repository/processes/Decision Tree Final 4.25.2020 - RapidMiner Studio Educational 5.6.000 @ LAPTOP-3HK7IAQ6

File Edit Process View Connections Settings Extensions Help

Views: Design Results Turbo Prep Auto Model Deployments Hadoop Data

Find data, operators, etc. All Studio

Result History PerformanceVector (Performance) ExampleSet (Select Attributes) LogisticRegression (Logistic Regression)

Criterion: accuracy, classification error, AUC, precision, recall, lift, sensitivity, specificity, positive predictive value, negative predictive value

negative\_predictive\_value: 76.24% (positive class: true)

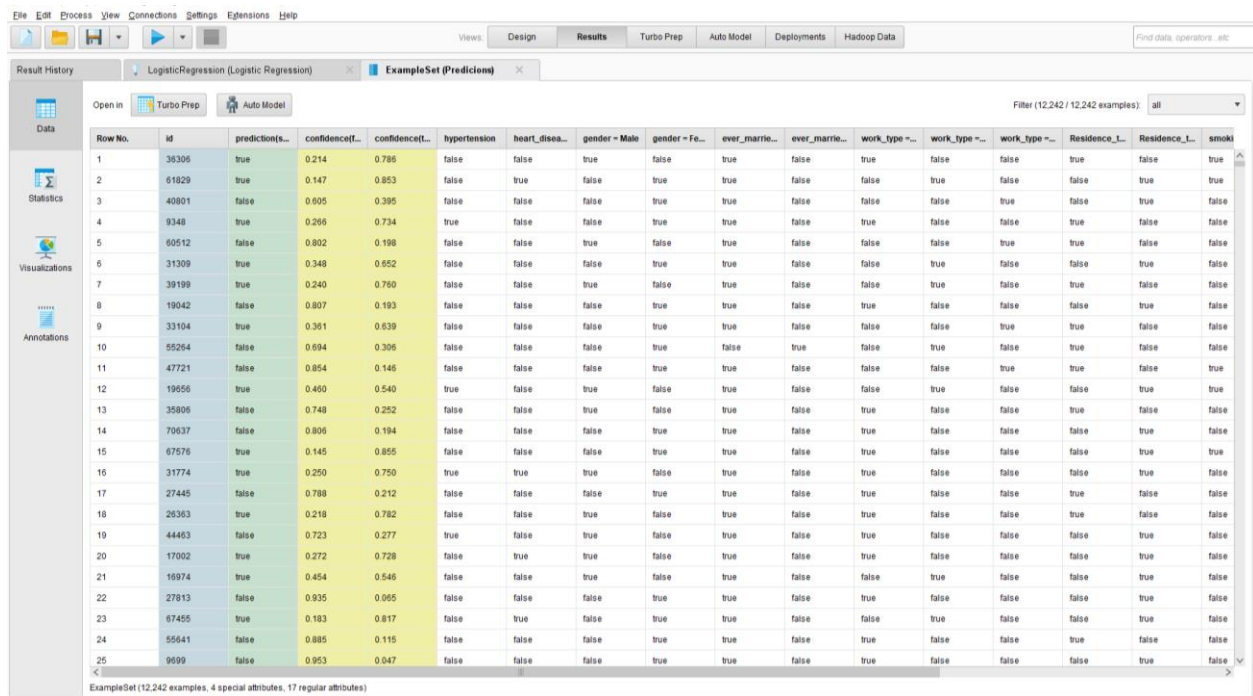
	true false	true true	class precision
pred false	4023	1254	76.24%
pred true	1578	4347	73.37%
class recall	71.83%	77.61%	

repository

processes: Arse, For, Arse, Arse, Arse, Dec, DTM, DTM, Exar, Exar, Gro, LRM, LRM, 10.11.2, Advanc, Arsene, Busine, versus, Cn3 HV, Cn3 Lo, Cn1 Hn, Chape, Data M, Decisio, Decisio, Ex/AM1, Exam1, Exam1f, Final E, Final E, Final E, Final E, HW6\_3

Type here to search

## Appendix 5: Logistic Regression example set



The screenshot displays a data analytics application window. The main area shows a table with 25 rows of data. The columns are: Row No., id, predictions, confidence, confidence, hypertension, heart\_disea..., gender = Male, gender = Fe..., ever\_marrie..., ever\_marrie..., work\_type ~..., work\_type ~..., work\_type ~..., Residence\_L..., Residence\_L..., and smoki. The data is filtered to show 12,242 examples. The interface includes a menu bar (File, Edit, Process, View, Connections, Settings, Extensions, Help) and a toolbar with icons for Data, Statistics, Visualizations, and Annotations. The table is titled 'ExampleSet (Predictions)' and is part of a 'LogisticRegression (Logistic Regression)' model.

Row No.	id	predictions	confidence	confidence	hypertension	heart_disea...	gender = Male	gender = Fe...	ever_marrie...	ever_marrie...	work_type ~...	work_type ~...	work_type ~...	Residence_L...	Residence_L...	smoki
1	36306	true	0.214	0.786	false	false	true	false	true	false	true	false	false	true	false	true
2	61829	true	0.147	0.853	false	true	false	true	true	false	false	true	false	false	true	true
3	40801	false	0.605	0.395	false	false	false	true	true	false	false	false	true	false	true	false
4	9348	true	0.266	0.734	true	false	false	true	true	false	true	false	false	true	false	false
5	60512	false	0.802	0.198	false	false	true	false	true	false	false	false	true	true	false	false
6	31309	true	0.348	0.652	false	false	false	true	true	false	false	true	false	false	true	false
7	39199	true	0.240	0.760	false	false	true	false	true	false	false	true	false	true	false	false
8	19042	false	0.807	0.193	false	false	false	true	true	false	true	false	false	true	true	false
9	33104	true	0.351	0.639	false	false	false	true	true	false	false	false	true	true	false	false
10	55264	false	0.694	0.306	false	false	false	true	false	true	false	true	false	true	false	false
11	47721	false	0.854	0.146	false	false	false	true	true	false	false	false	true	true	false	true
12	19656	true	0.480	0.540	true	false	true	false	true	false	false	true	false	false	true	true
13	35806	false	0.748	0.252	false	false	true	false	true	false	true	false	false	true	false	false
14	70637	false	0.806	0.194	false	false	false	true	true	false	true	false	false	false	true	false
15	67576	true	0.145	0.855	false	false	false	true	true	false	true	false	false	false	true	true
16	31774	true	0.250	0.750	true	true	true	false	true	false	true	false	false	false	true	false
17	27445	false	0.788	0.212	false	false	false	true	true	false	true	false	false	true	false	false
18	26363	true	0.218	0.782	false	false	true	false	true	false	true	false	false	true	false	false
19	44463	false	0.723	0.277	true	false	true	false	true	false	true	false	false	false	true	false
20	17002	true	0.272	0.728	false	true	true	false	true	false	true	false	false	false	true	false
21	16974	true	0.454	0.546	false	false	true	false	true	false	false	true	false	false	true	false
22	27813	false	0.935	0.065	false	false	false	true	true	false	true	false	false	false	true	false
23	67455	true	0.183	0.817	false	true	false	true	true	false	false	true	false	false	true	false
24	55641	false	0.885	0.115	false	false	false	true	true	false	true	false	false	true	false	false
25	9699	false	0.953	0.047	false	false	false	true	true	false	true	false	false	false	true	false