Project 3 **Veronika Ebenal, Raghavendra Tikare, Stanislav Klein**

Background and goal of the project

Stroke is a disease that affects the arteries leading to and within the brain. A stroke occurs when a blood vessel that carries oxygen and nutrients to the brain is either blocked by a clot or bursts (or ruptures). When that happens, part of the brain cannot get the blood (and oxygen) it needs, so it and brain cells die. The main risk factor for stroke is high blood pressure. Other risk factors include smoking, obesity and high blood cholesterol. In 2015, stroke was the second most frequent cause of death after coronary artery disease accounting for 6.3 million deaths.

The goal of this project is to build a tool that could help predict the probability of a stroke happening to help doctors take proactive health measures for these patients. By using Apache Spark, a framework for cluster-computing, we can analyze and evaluate large sets of data, which we then have to extract statistical information from. To see if processing the data has an impact on the final results the same has to be done after cleaning the data. Finally, we need to implement three different approaches to reach the goal. A clustering approach to analyze the data's underlying structure, the predictor to compute the probability of a stroke happening and a classifier to classify the patient as being in danger of a stroke.

Description of the data (including a summary of the data statistics)

The data we used was obtained from the McKinsey Stroke Dataset. This dataset contains patient data influencing the risk of a stroke for more than 60.000 patients including their age, average glucose level as well as if they have ever suffered from a heart disease or hypertension before. We gathered some preliminary information about the data. For example, 1.85% of males had strokes, while only 1.68% of females did. 90.86% of the strokes had occurred for people who were over 50 years old.

Preprocessing steps (e.g. cleaning)

The initial data needed to be processed in two ways. First, it had some missing data which needed to be imputed. We imputed it by filling in null data for categorical attributes with "unknown" and for numerical data we filled in the missing values with the mean. We also did one-hot encoding on some attributes, so that it was emphasized to the program the categorical nature of those attributes.

Approaches

CLUSTERING

```
+-----+------+----+------------+-------------+------------+-------------+--------------+-----------------+----+------
|   id|gender| age|hypertension|heart_disease|ever_married|    work_type|Residence_type|avg_glucose_level| bmi| smoki
+-----+------+----+------------+-------------+------------+-------------+--------------+-----------------+----+------
|36306|  Male|80.0|           0|            0|         Yes|      Private|         Urban|            83.84|21.1|former
|61829|Female|74.0|           0|            1|         Yes|Self-employed|         Rural|            179.5|26.0|former
|14152|Female|14.0|           0|            0|          No|     children|         Rural|            95.16|21.2|
|12997|  Male|28.0|           0|            0|          No|      Private|         Urban|            94.76|23.4|
|40801|Female|63.0|           0|            0|         Yes|     Govt_job|         Rural|            83.57|27.6|   nev
| 9348|Female|66.0|           1|            0|         Yes|      Private|         Urban|           219.98|32.2|   nev
|51550|Female|49.0|           0|            0|         Yes|Self-employed|         Rural|            74.03|25.1|
|60512|  Male|46.0|           0|            0|         Yes|     Govt_job|         Urban|            120.8|32.5|   nev
|31309|Female|75.0|           0|            0|         Yes|Self-employed|         Rural|            78.71|28.0|   nev
|39199|  Male|75.0|           0|            0|         Yes|Self-employed|         Urban|             77.2|25.7|
|15160|Female|17.0|           0|            0|          No|      Private|         Rural|            78.16|21.9|
|21705|Female|10.0|           0|            0|          No|     children|         Urban|           107.23|19.4|
|19042|Female|47.0|           0|            0|         Yes|      Private|         Rural|             91.6|26.7|   nev
|12249|Female|42.0|           0|            0|         Yes|      Private|         Urban|            83.05|32.3|
|33104|Female|67.0|           0|            0|         Yes|     Govt_job|         Urban|            236.6|24.2|   nev
|55264|Female|52.0|           0|            0|          No|Self-employed|         Urban|           109.49|24.5|   nev
|29445|  Male|73.0|           0|            0|         Yes|Self-employed|         Rural|           109.66|40.0|
|49013|Female|19.0|           0|            0|          No|      Private|         Rural|            88.51|22.1|
|  276|  Male|15.0|           0|            0|          No|     children|         Rural|           101.36|22.3|
|47721|Female|37.0|           0|            0|         Yes|     Govt_job|         Urban|           165.44|36.1|former
+-----+------+----+------------+-------------+------------+-------------+--------------+-----------------+----+------
only showing top 20 rows
```

In clustering, data points are grouped together due to having similar properties. Anything not in a cluster is an outlier. This is a type of unsupervised learning. This is mostly done to analyze the underlying structure of the data

LOGISTIC REGRESSION

```
Logistic Regression Test Area Under ROC:  0.8513157219113173
LR PREDICTIONS
```

	id	features	prediction	probability
0	36306	(0.0, 1.0, 80.0, 0.0, 0.0, 1.0, 0.0, 1.0, 0.0,...	0.0	[0.916457328152081, 0.08338111343218267, 0.000...
1	61829	(1.0, 0.0, 74.0, 0.0, 1.0, 1.0, 0.0, 0.0, 1.0,...	0.0	[0.9212655890612188, 0.07844211598585572, 0.00...
2	14152	(1.0, 0.0, 14.0, 0.0, 0.0, 0.0, 1.0, 0.0, 0.0,...	0.0	[0.99908429170487, 0.0008930211564833068, 2.26...
3	12997	(0.0, 1.0, 28.0, 0.0, 0.0, 0.0, 1.0, 1.0, 0.0,...	0.0	[0.9943532083046926, 0.005596191439829134, 5.0...
4	40801	(1.0, 0.0, 63.0, 0.0, 0.0, 1.0, 0.0, 0.0, 0.0,...	0.0	[0.9810398196130841, 0.018875622214482423, 8.4...

In logistic regression, the algorithm estimates an outcome by choosing parameters that maximize the likelihood of the observed values. This is supervised learning. Logistic regression can predict the probability of a stroke happening to a patient.

DECISION TREE

```
A Decision Tree algorithm had an accuracy of: 98.26%
DECISION TREE PREDICTIONS
```

	id	features	prediction	probability
0	36306	(0.0, 1.0, 80.0, 0.0, 0.0, 1.0, 0.0, 1.0, 0.0,...	0.0	[0.9116809116809117, 0.08831908831908832, 0.0]
1	61829	(1.0, 0.0, 74.0, 0.0, 1.0, 1.0, 0.0, 0.0, 1.0,...	0.0	[0.9444444444444444, 0.05555555555555555, 0.0]
2	14152	(1.0, 0.0, 14.0, 0.0, 0.0, 0.0, 1.0, 0.0, 0.0,...	0.0	[0.991981834953523, 0.008018165046476974, 0.0]
3	12997	(0.0, 1.0, 28.0, 0.0, 0.0, 0.0, 1.0, 1.0, 0.0,...	0.0	[0.991981834953523, 0.008018165046476974, 0.0]
4	40801	(1.0, 0.0, 63.0, 0.0, 0.0, 1.0, 0.0, 0.0, 0.0,...	0.0	[0.991981834953523, 0.008018165046476974, 0.0]

A decision tree classifier is built in a tree structure. It breaks down the dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes for each output, branching off according to possible types, and finally ending in leaf nodes for possible classifications.This

is supervised learning. The decision tree here will classify a patient as “no danger of stroke” vs. “danger of stroke”.

The biggest difference between these approaches is that logistic regression and decision tree are supervised learning, meaning that the program is trained using labeled data and later tested using unlabeled data. Clustering is unsupervised learning, meaning that no training data is used and the program must only work with unlabeled data. Between the supervised learning algorithms, both have benefits and downsides. Logistic regression has low variance, but high bias. Decision trees are arguably the easiest to interpret and work well with categorical data, but are prone to overfitting. For this project, it seems that logistic regression would be the best choice. The output as a probability rather than a simple “chance” or “no chance” of stroke would be more useful in a medical setting. The presence of training data should be made use of, so clustering alone would not be the best approach. And the presence of non-categorical attributes is also well suited to logistic regression.

Effect of imputation (comparison of the predictor trained w/ and w/o imputed data)
We used imputation of the data for replacing missing data with substituted values. Missing data can reduce the statistical power of a study and can produce biased estimates, leading to invalid conclusions. The problem of missing data is relatively common in almost all research and can have a significant effect on the conclusions that can be drawn from the data.
When we carried out the imputation for our dataset and compared the outputs of logistic regression, we overall did not find much difference in the outputted prediction values. The areas under ROC: 0.8182 and 0.8513 for non-imputed and imputed, respectively. As this is a correlated measurement of accuracy, we see that imputed data leads to more accurate results.

Discussion: why is this a typical project for a data-scientist? (Or why not?)
This could definitely be considered a typical project for a data scientist. All of the characteristics of data science discussed in lecture were used in one way or another in this project.
Statistics and mathematics: we gathered data statistics, as well as measured algorithm accuracies
Machine learning: logistic regression and decision tree use machine learning
Programming: the project was actualized using programming in Python.
Data processing: one-hot encoding and imputing of the data is an example of data processing.
Data visualization: in order to better understand the data, we first visualized it with different plots and graphs.
Data knowledge: understanding program outputs requires an understanding of the project background

Data Imputation: cleaning the data with replacing the missing values to gain accuracy and reduce errors.