What are the Strongest Statistical Predictors of a Stroke?

FINC 430 Research Project

By Dalia Poblano

Professor Vijay Kumar

December 03, 2025

Introduction

According to the Center for Disease Control and Prevention, “Every 40 seconds, someone in the United States has a stroke. Every 3 minutes and 14 seconds, someone dies of stroke in this country.” Needless to say, strokes can affect anyone and, “it is a leading cause of death for Americans”(Center for Disease Control and Prevention). However, there are many lifestyle habits that someone can introduce in their life in order to minimize the chances of experiencing a stroke. Some of these habits include: eliminating smoking or vaping, eating healthy foods, being physically active, staying at a healthy weight, limiting alcohol intake, being cautious of your blood pressure, reducing stress levels, keeping up with medical checkups, and getting adequate sleep (American Stroke Association).

While experts urge society to follow these tips in order to avoid strokes and stroke complications, many people still overlook this guidance and continue engaging in unhealthy behaviors. Understanding which factors most strongly increase the likelihood of a stroke may motivate individuals to take this information more seriously. That is why I chose to explore a stroke-prediction dataset for my research project, aiming to answer the question: *What are the strongest statistical predictors of a stroke?* Like me, many people rely on statistics and evidence-based information, and I hope the findings from this project can encourage the community to stay aware of these factors and help reduce the number of strokes.

Material and Methods

Data: Click Here!

The dataset I used for this project came from Kaggle, a reliable source that provides people with various datasets for data analytics and projects. This dataset is composed of data from many participants, some who have experienced a stroke and some who haven't, and 12 columns. The column names are **id, gender, age, hypertension, heart_disease, ever_married, work_type, Residence_type, avg_glucose_level, bmi, smoking_status, and stroke**.

Numerical Values: id, age, avg_glucose_level, and bmi

Categorical Values: gender, hypertension, heart_disease, ever_married, work_type, Residence_type, smoking_status, and stroke

*Note: Although hypertension, heart_disease, and stroke are categorical values (yes/no questions), the dataset values include 0 or 1 for these columns, meaning 0 for no and 1 for yes.

Focus Variables:

To answer the research question, I included all variables except id, Residence_type, and work_type. The variable id was removed because it is only a unique identifier and provides no predictive or clinical information. The variables Residence_type and work_type were excluded because they are not biological or medical risk factors for stroke and showed limited relevance during preliminary exploration. Removing them allows the analysis to focus on clinically meaningful predictors.

Techniques Used:

- Data Importing and Cleaning
- Data Wrangling
- Exploratory Data Analysis
- Statistical Models
- Machine Learning Models
- Model Interpretability
- DataVisualization

Data Prep:

Before starting to work with the data, I checked the data types that I would be working with and if there were any missing values in the data. I discovered that the variable “bmi” had 201 missing values.

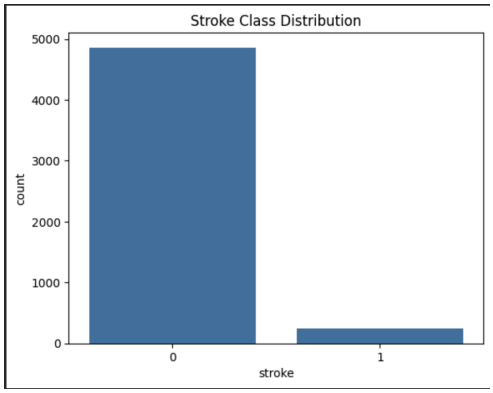
```
Missing values:
 id                   0
gender                0
age                   0
hypertension          0
heart_disease         0
ever_married          0
work_type             0
Residence_type        0
avg_glucose_level     0
bmi                 201
smoking_status        0
stroke                0
dtype: int64
```

Due to this reason I decided to impute missing bmi values to make the data more clean and usable. I also dropped the columns that I wouldn’t be using for my research such as id, work_type, and Residence_type. Lastly, I encoded my categorical variables such as gender, ever_married, and smoking_status, for better usage.

Results with Interpretations

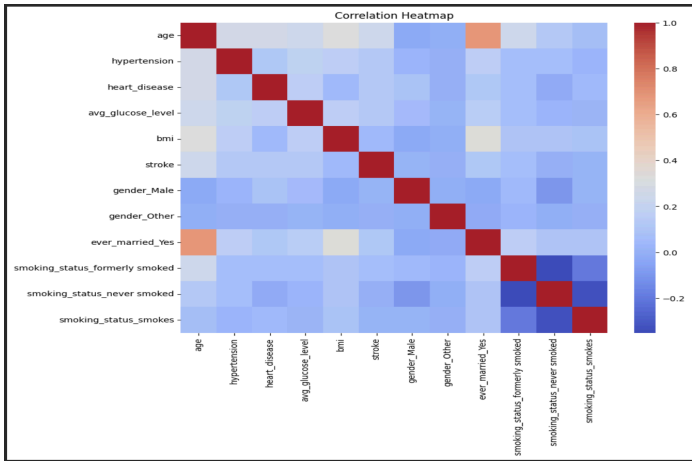
Exploratory Data Analysis:

After cleaning the data from any missing values, I started off by checking if my dataset was imbalanced. This is a crucial step to take because we ideally want reliable model performance and if our dataset was imbalanced then that could lead us to misleading results.



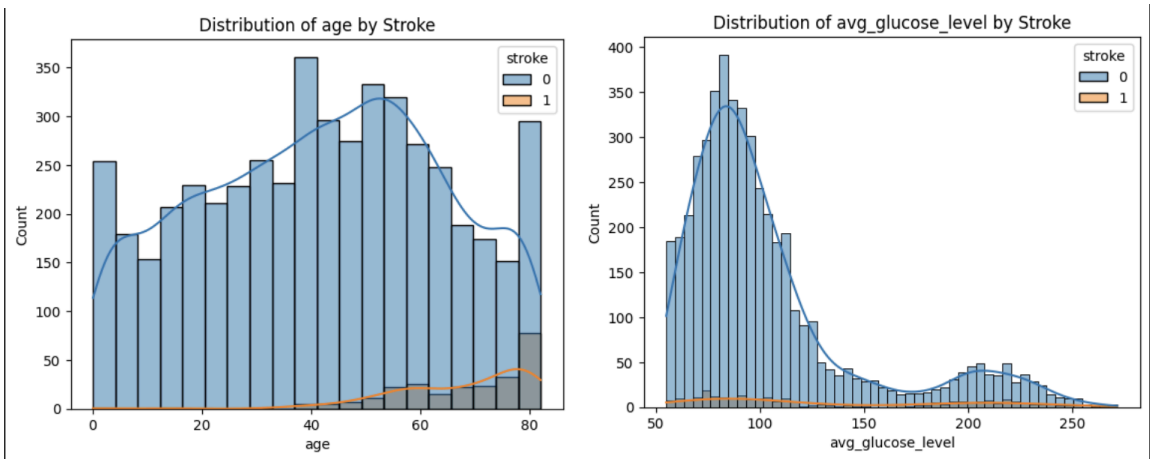
The class distribution shows severe imbalance, with the majority class (no stroke) dominating the dataset. This imbalance can bias machine learning models toward predicting the majority class, so techniques such as class weighting and stratified sampling will be used to ensure fair model evaluation.

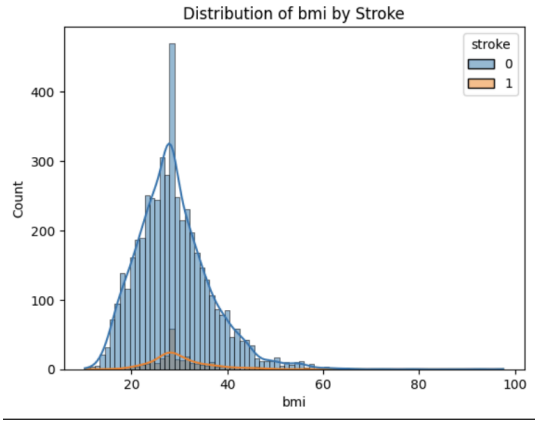
Correlation Heatmap



The correlation heatmap shows that no predictor has a very strong linear relationship with stroke. Age, hypertension, heart disease, and average glucose level show weak–moderate positive correlations with the stroke outcome, while variables such as BMI and smoking status show little linear association. This suggests that stroke risk is influenced by complex, nonlinear relationships rather than simple linear correlations, which justifies the use of machine learning models such as Random Forest, XGBoost, and SVM.

Analyzing Distributions





These distributions can help us come up with some initial conclusions about how age, avg_glucose_level, and bmi interacts with stroke predictions.

Initial Conclusions Based on Distributions:

- Age is an important predictor of stroke risk. The older someone is, the higher their chance of experiencing a stroke.
- Higher average glucose levels appear linked to increased stroke risk. Glucose is likely another meaningful predictor.
- BMI may have some influence on stroke risk, but it is not a highly predictive feature on its own.

Statistical Model: Logistic Regression

After running a simple logistic regression model on my dataset, these are the coefficients I received back:

	Feature	Coefficient
7	ever_married_Yes	-0.127743
9	smoking_status_never smoked	-0.043392
6	gender_Other	-0.023229
4	bmi	0.001632
5	gender_Male	0.031328
8	smoking_status_formerly smoked	0.048096
2	heart_disease	0.049218
10	smoking_status_smokes	0.089560
1	hypertension	0.134638
3	avg_glucose_level	0.181822
0	age	1.589709

Based on these results, we can make the following conclusions:

- Age is the most influential factor, meaning as age increases, the odds of having a stroke increase significantly. **(1.589709)**
- Higher glucose levels increase the likelihood of stroke which is consistent with diabetes being a risk factor. **(0.181822)**
- People with hypertension have higher odds of stroke. **(0.134638)**
- Current smokers are more likely to experience a stroke. **(0.089560)**
- Having heart disease increases stroke risk slightly. **(0.0492)**
- Former smokers have a small increased risk, but less than current smokers.**(0.0481)**
- Being male slightly increases stroke risk, though this effect is very small. **(0.0313)**
- BMI does not meaningfully contribute to stroke prediction in your model.**(0.0016)**
- People who have never been smokers have slightly lower stroke risk.**(–0.0433)**
- Married individuals show lower stroke risk in our dataset.**(–0.1277)**

Machine Learning Models:

Aside from the Logistic Regression model, I also decided to use the following Machine Learning models to compare their performances and ultimately help me answer the research question

1. KNN
2. Naive Bayes
3. SVM (Linear and RBF)
4. Decision Tree
5. Random Forest
6. XGBoost

These were the performance results for the models mentioned above:

	Accuracy	Precision	Recall	F1	AUC
Logistic Regression	0.048924	0.048924	1.00	0.093284	0.563981
KNN	0.948141	0.200000	0.02	0.036364	0.626739
Naive Bayes	0.345401	0.068340	0.98	0.127771	0.801296
SVM Linear	0.951076	0.000000	0.00	0.000000	0.515720
SVM RBF	0.951076	0.000000	0.00	0.000000	0.662716
Decision Tree	0.909002	0.169231	0.22	0.191304	0.582222
Random Forest	0.947162	0.000000	0.00	0.000000	0.813683
XGBoost	0.945205	0.312500	0.10	0.151515	0.805885

Due to our imbalanced dataset, our models show high accuracy but very low recall or high recall but low precision. This makes our accuracy scores unreliable or useless. Due to this reason, we will be relying on the F1 and AUC metrics. F1 balances precision and recall, making it a better indicator of how well the model identifies stroke patients. Similarly, AUC shows how well the model separates positive vs negative classes regardless of threshold.

Logistic Regression: The F1 score is 0.093284, meaning it identifies cases but it predicts too many positive cases. The AUC is 0.563981 which means this model did a below average separation of non-stroke and stroke cases.

KNN: The F1 score is 0.036364 which is still very low, indicating that its ability to detect true strokes is very weak. The AUC is 0.626739 which shows a moderate separation performance for non-stroke and stroke cases.

Naive Bayes: The F1 score is 0.127771 which is very high compared to the other models. The AUC is 0.801296 which is also one of the highest AUC scores amongst the other models. This means the naive bayes model did a very good job at separating classes.

SVM Linear/ RBF: Both SVM models performed weakly compared to the other models. For SVM Linear and RBF the F1 scores were both 0s which means these models predicted zero stroke cases. According to their AUC scores, SVM Linear performed barely above random guessing and SVM RBF performed moderately.

Decision Tree: The F1 score is 0.191304 which demonstrates the best balance between identifying stroke cases and limiting false alarms. The AUC score is 0.582222 which indicates below average separation.

Random Forest: The F1 score is 0, meaning this model predicted zero stroke cases, resulting in an F1 of zero. The AUC score was 0.813683 which means it has the best ability to distinguish stroke vs non-stroke overall.

XGBoost: The F1 score is 0.151515, indicating a decent overall balance and it is better than most other models. The AUC score is 0.805885 which shows strong and consistent class separation.

Conclusions

Random Forest, Naive Bayes, and XGBoost have the strongest AUC scores, meaning they are best at distinguishing between stroke and non-stroke patients. Decision Tree, XGBoost, and Naive Bayes have the highest F1 scores, meaning they provide the best balance between catching stroke cases and avoiding false positives.

Random Forest Model for Feature Importance:

	Feature	Importance
3	avg_glucose_level	3.169036e-01
4	bmi	2.634034e-01
0	age	2.556072e-01
5	gender_Male	3.059829e-02
1	hypertension	2.756304e-02
9	smoking_status_never smoked	2.351487e-02
2	heart_disease	2.265223e-02
8	smoking_status_formerly smoked	2.057726e-02
7	ever_married_Yes	2.011207e-02
10	smoking_status_smokes	1.906743e-02
6	gender_Other	6.977179e-07

Based on the random forest feature importance results, we can see that avg_glucose_level is the strongest predictor in the model. This makes sense because high glucose levels are known to be linked with metabolic issues and stroke risk. Additionally, bmi has the second largest effect in stroke prediction. This also sounds reliable because higher BMI is associated with obesity, which increases cardiovascular risk. Age is also an important predictor of strokes. Traditional cardiovascular risk factors like hypertension and heart disease were moderately influential. Smoking-related features had small but meaningful contributions. Demographic features like gender and marital status had low impact, with "gender_Other" being almost irrelevant.

XGBoost Model for Feature Importance:

	Feature	Importance
0	age	0.185559
2	heart_disease	0.128700
7	ever_married_Yes	0.128021
1	hypertension	0.111243
4	bmi	0.081208
3	avg_glucose_level	0.078911
10	smoking_status_smokes	0.074723
9	smoking_status_never smoked	0.074143
8	smoking_status_formerly smoked	0.070028
5	gender_Male	0.067464
6	gender_Other	0.000000

The findings from the XGBoost model might look a little different from the rankings of the random forest model because these models learn in different ways. XGBoost builds trees sequentially, each correcting the previous one's mistakes (focusing heavily on hard-to-predict patterns).Random Forest builds many trees independently and averages them (focusing more on overall stability and variance reduction).Overall, features that help reduce error in tricky parts of the data become more important for XGBoost, while Random Forest rewards features that help consistently split data across many trees.

Despite any minor differences, these conclusions remain the same:

- Age is consistently one of the strongest predictors of stroke risk. I can confidently say that age is one of the most influential factors in predicting stroke.
- Avg_glucose_level is a very important predictor. This aligns with clinical expectations.
- BMI, hypertension, and heart_disease are other strong predictors for stroke risk, even if age and avg_glucose_level are stronger predictors.

Conclusion

In conclusion, the strongest statistical predictors of a stroke are age, avg_glucose_level, and bmi. These 3 variables are the strongest predictors but cardiovascular factors like hypertension and heart disease are also influential. I reached these conclusions based on the machine learning models I explored throughout my research project. While I studied six different machine learning models—KNN, Naive Bayes, SVM (Linear/RBF), Decision Tree, Random Forest, XGBoost, and also Logistic Regression—I ultimately decided to use only Random Forest and XGBoost. These two models performed the best and are the most reliable for identifying the strongest predictors. Moreover, the feature importance rankings from both Random Forest and XGBoost were similar, highlighting age, average glucose level, and BMI as the strongest predictors of stroke risk. This consistency across models makes me even more confident in my statistical findings. Lastly, my findings align with medical advice on stroke prevention, which emphasizes monitoring blood glucose levels, maintaining a healthy weight, and being aware that older adults are at higher risk for strokes.

References

"Let's Talk about Lifestyle Changes to Prevent Stroke." *Www.Stroke.Org*, www.stroke.org/en/help-and-support/resource-library/lets-talk-about-stroke/lifestyle-cha

"Stroke Facts." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, www.cdc.gov/stroke/data-research/facts-stats/index.html.