

Project 2

Suhas, Morgan, Bini, Sheryl


STROKE PREDICTION DATASET

- According to the study of World Health Organization (WHO) , stroke is listed as the second leading cause of death globally. It is responsible for approximately 11% of total deaths.
- This dataset is used to predict whether a patient is likely to get stroke based on the parameters like age, gender, various diseases, and smoking status. Each row in the data provides relevant information about the patient.



Our Data

Our data has different information about a patient and whether they had a stroke or not.

- 1) id
 - 2) gender: "Male", "Female" or "Other"
 - 3) age
 - 4) hypertension: 0 patient doesn't have hypertension, 1 patient has hypertension
 - 5) heart_disease: 0 patient doesn't have any heart diseases, 1 patient has a heart disease
 - 6) ever_married: No or Yes
 - 7) work_type: children, Govt_jov, Never_worked, Private or Self-employed
 - 8) Residence_type: Rural or Urban
 - 9) avg_glucose_level: average glucose level in blood
 - 10) bmi: body mass index
 - 11) smoking_status: formerly smoked, never smoked, smokes or Unknown
 - 12) stroke: 1 if the patient had a stroke or 0 if not
- 

Preparing our data

Columns that had binary categorical data was changed to binary integers:

gender: 0: Female 1: Male

ever_married: 0: No 1: Yes

Residence_type: 0: Rural 1: Urban

gender	age	hypertension	heart_disease	ever_married	Residence_type
1	67.0	0	1	1	1
1	80.0	0	1	1	0
0	49.0	0	0	1	1
0	79.0	1	0	1	0
1	81.0	0	0	1	1

Preparing our data cont.

To prepare our categorical data for training we used One-Hot Encoding.

We converted our categorical data into new columns and assigned binary values.

Work type: Never_worked, Govt_job, Private, Self-employed, Children

Smoking status: Unknown, Formerly_smoked, never_smoked, smokes

Govt_job	Never_worked	private	Self-employed	children	formerly_smoked	never_smoked	smokes
0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0
0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0
0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0
0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0
0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0

Details of our dataset

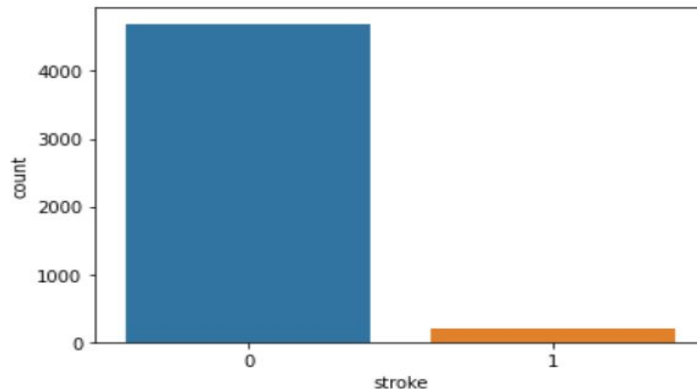
- In our dataset, patients suffer from stroke is 1 and if not is denoted by 0

```
stroke.stroke.value_counts()
```

```
0    4698
```

```
1     209
```

```
Name: stroke, dtype: int64
```



- The data shows 209 positives with 4698 negative. The reason for the positive stroke patients to be so low will be due to the quite short follow up period rather than a person's life course.

Dependent and Independent Variable

We want to predict whether a patient is likely to have a stroke based on different information about the patient.

Our dependent variable is Stroke, and all other columns, not including ID or Unknown (Smoking Status), are our independent variables.

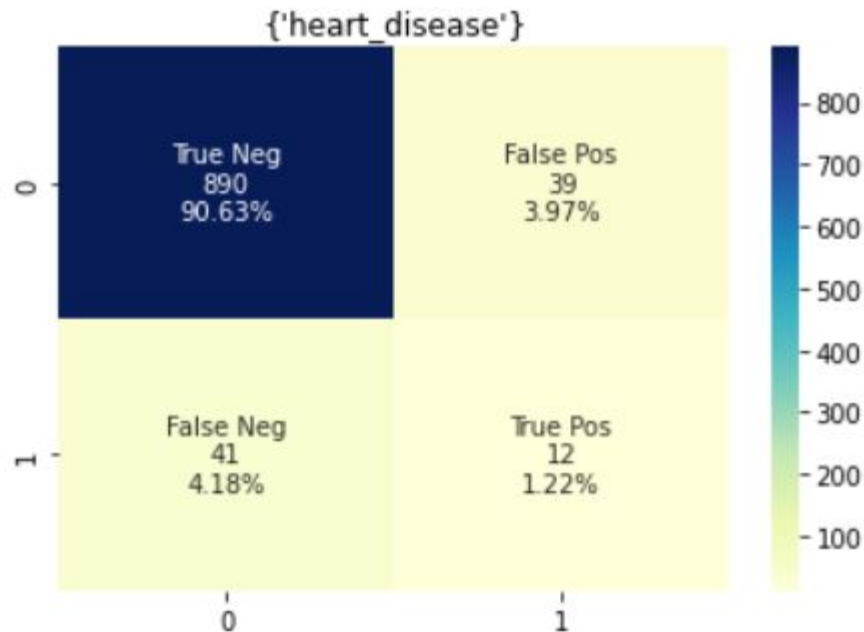
For this data we can predict a stroke with 75.56% accuracy based on the independent variables.



Heart Disease Accuracy

- When it came to using Heart Disease to predict whether one has had a stroke for this data set it was quite accurate with a accuracy of

0.9185336048879837



Heart Disease Explanation

- According to CDC “common heart disorders can increase your risk for stroke”. This is because conditions such as enlarged heart chambers, atrial fibrillation, and heart valve defects “can cause blood clots that may break loose and cause a stroke”.

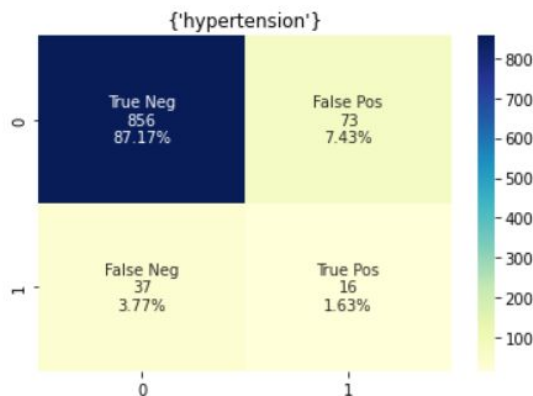


Hypertension

Hypertension is a condition in which someone has long-term high blood pressure.

According to the CDC, "having hypertension puts you at risk for heart disease and stroke."

Using Hypertension to predict a stroke, we got an accuracy of 88.798%.



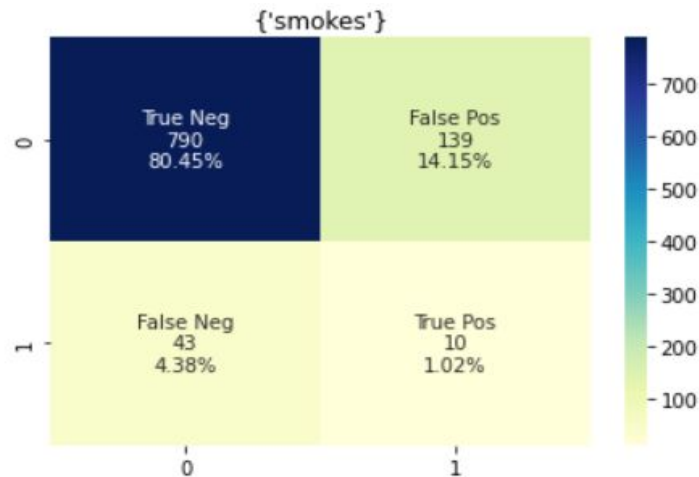
hypertension
Accuracy: 0.8879837067209776

Smokes vs Stroke

- According to JAHA(Journal of American Heart Association) , Smoking is a well-established risk factor of stroke and smoking cessation has been highly recommended for stroke prevention.
- From the smoking_status, the parameter smokes is used to predict stroke, an accuracy of 81.46% is obtained

smokes

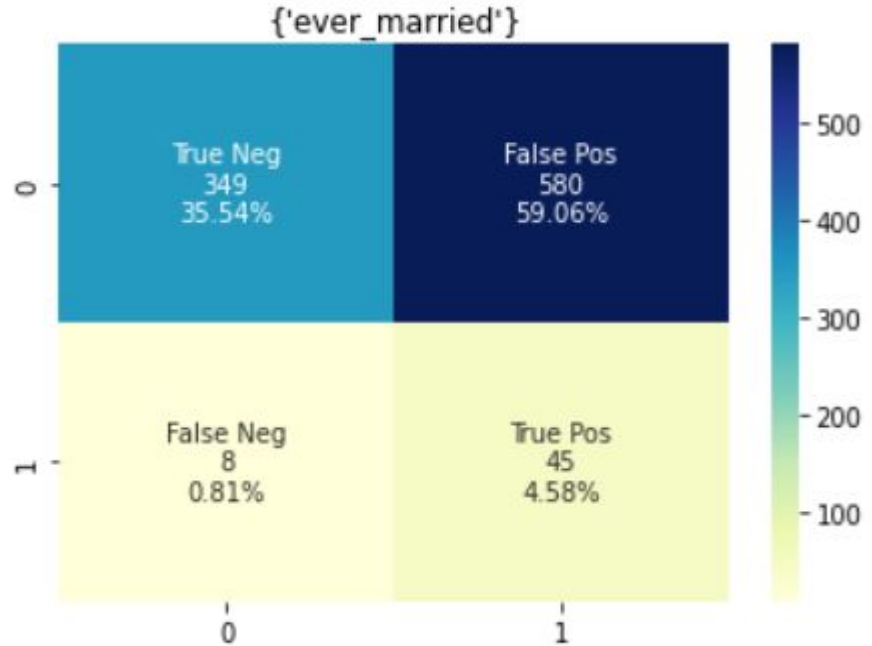
Accuracy: 0.814663951120163



Ever married vs stroke

- The ever married status is used to predict the stroke and an accuracy of 40.122% is obtained

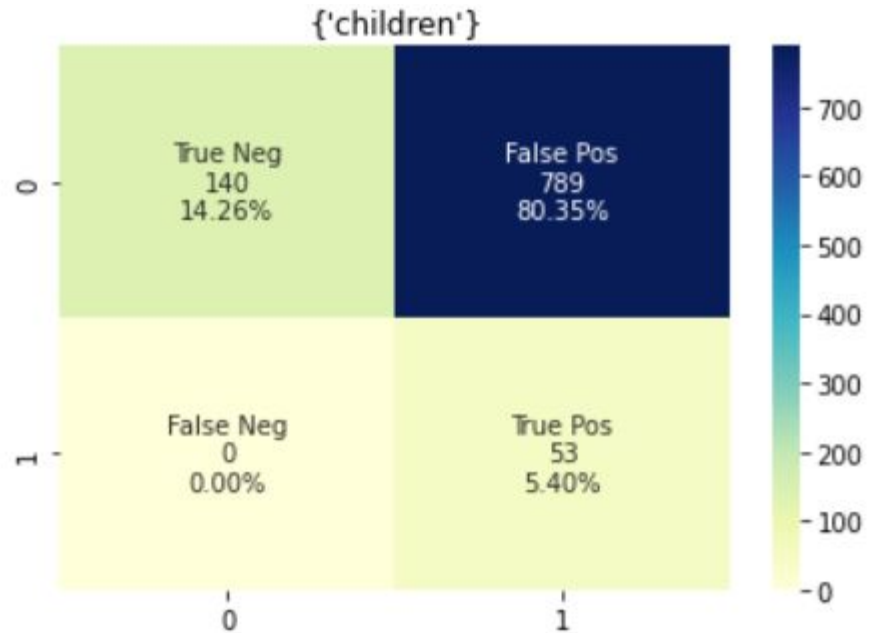
```
ever_married  
Accuracy: 0.40122199592668023
```



Children vs Stroke

- When the work type(children) is used to predict the stroke, an accuracy of 19.65% is obtained.
- Total number of cases = 982
- Accuracy = $(53+140)/982 = 0.1965$

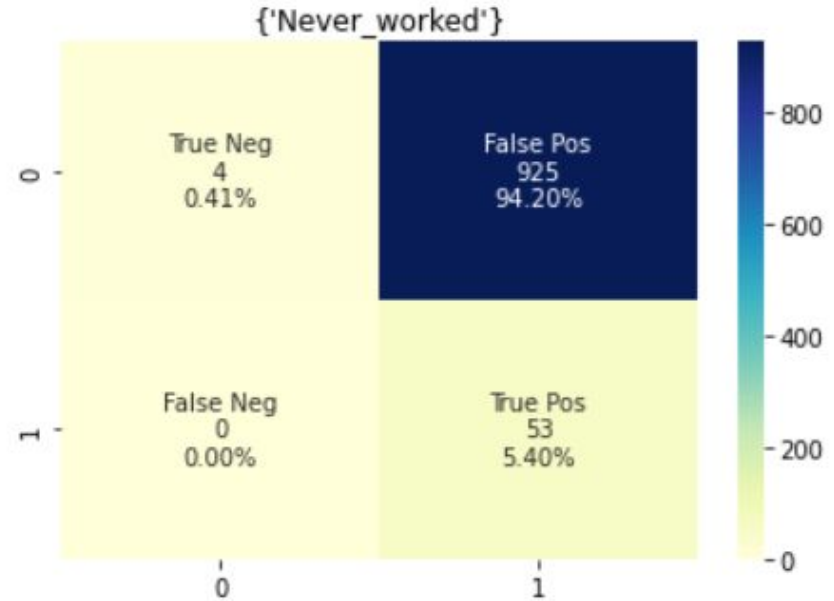
```
children  
Accuracy: 0.19653767820773932
```



Never worked vs Stroke


- When the worktype(never worked) category is used to predict the stroke an accuracy of 5.80% is obtained.
- Accuracy = $(53+4)/982 = 0.0580$

Never_worked
Accuracy: 0.05804480651731161



Conclusion

Hence, from the analysis above we see that:

- There are three independent variables - Heart Disease, Hypertension and Smoking_Status which have a high impact in predicting the dependent variable - Stroke
 - Whereas, there are three independent variables - Ever married, Children and Never worked which have a considerably lower impact in predicting the stroke.
 - Heart Disease has the highest impact and Never Worked(work_type) has the lowest impact in predicting the stroke.
- 



Thank You!