Stroke Analysis

Capstone Project Report

December 18, 2022

Priya Johny

Ms. Data Science

TABLE OF CONTENT

ABSTRACT

In various fields around the world, machine learning is used. There are no exceptions in the healthcare sector. Machine learning has a crucial role in diagnosing conditions including Stroke, high blood pressure, heart disease, and more. If foreseen far in advance, such information can offer intuitions essential to doctors, who can modify their diagnosis and approach per patient.

Visualization has helped people to view, interact with, and better understand data in several meaningful ways. Data visualization is essential for data cleaning, exploring data structure, detecting outliers and unusual groups, and identifying trends and clusters to a great expend. It helps businesses, stakeholders, and gain an enormous amount of knowledge on how they can make the best of their data for the company's growth.

In this project, we conduct an analysis of patients' data and perform machine learning modeling to predict the lifestyle of stroke-affected patients. We further analyze the data to find through research that lifestyles result in about 87%, according to the Centers for disease control and prevention which led to the ischemic dataset analysis.

INTRODUCTION

The World Health Organization estimates that Stroke causes millions of deaths worldwide each year. One of the leading causes of morbidity and mortality among the global population is Stroke. One of the most crucial topics in the data analysis area is Stroke prediction. Numerous studies have been carried out to identify the most important risk factors and to precisely estimate the overall risk. This disease is also referred to as a silent killer because it causes a person to pass away without any evident signs. In high-risk individuals, an early diagnosis of the disease is crucial for helping them decide whether to change their lifestyle, which lowers the consequences.

There is not much research employing machine learning (ML) to predict stroke at the professional level, even though ML models have been found to perform better than clinical risk estimates when compared to statistical methods. This study employed machine learning techniques on a dataset gathered during a health assessment survey of patients from various age groups. We are predicting whether a person is having a chance of Stroke or not using different features like age, BMI, smoking, etc.

Ischemic stroke comprises of according to CDS is one of the two major types of stroke occurrence in patients. CDC defines it as “An ischemic stroke that occurs when blood clots or other particles block the blood vessels to the brain. Fatty deposits called plaque can also cause blockages by building up in the blood vessels.”

Abbreviations: BMI: Body Mass Index; OSHPD: Office of Statewide Health Planning and Development

PURPOSE

The purpose of our project is to analyze lifestyle features with respect to Stroke using a dataset obtained through medical examinations in hospitals. According to earlier research, the variability of stroke is a substantial and independent risk factor for many other diseases. We further analyze the ischemic stroke dataset to classify hospitals as better, as expected, or worse with respect to hospitals in California.

DATASET

Dataset I: In this case study is from Kaggle named as “Stroke Prediction dataset” and contains medical information of patients. The dataset contains 5110 patient records and 11 features. We are using “Stroke” as our target variable. It is a Binary class classification problem, where the value of the class is 0 and 1 (0 means the patient will not have a stroke and 1 means the patient will have a stroke). Dataset is imbalanced and consists of 10 attributes other than the target attribute.

Dataset II: In this case study is from California Health and Human Services Agency named “Ischemic Stroke 30-Day Mortality and 30-Day Readmission Rates and Quality Ratings for CA Hospitals”. The dataset contains risk-adjusted 30-day mortality and 30-day readmission rates, quality ratings, and the number of deaths or readmissions and cases for ischemic stroke patients that are treated in California hospitals, but does not include ischemic stroke treated in outpatient settings.

Problem Statement

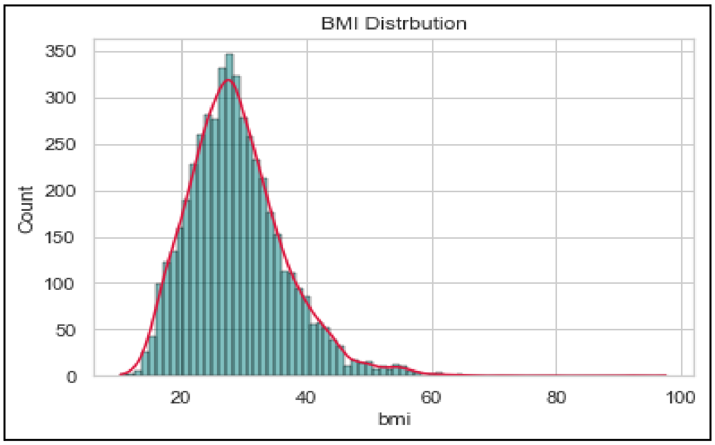
This project, analyzes lifestyle factors such as age, BMI, heart disease, etc, which cause the risk of stroke in patients. Feature importance using some of the ML models is predicted. And hospitals pertaining to the second dataset are classified as better, as expected, or worse through the tableau visualization analysis.

This project will help medical and health professionals to evaluate the patient medical condition more precisely and accurately through visualization.

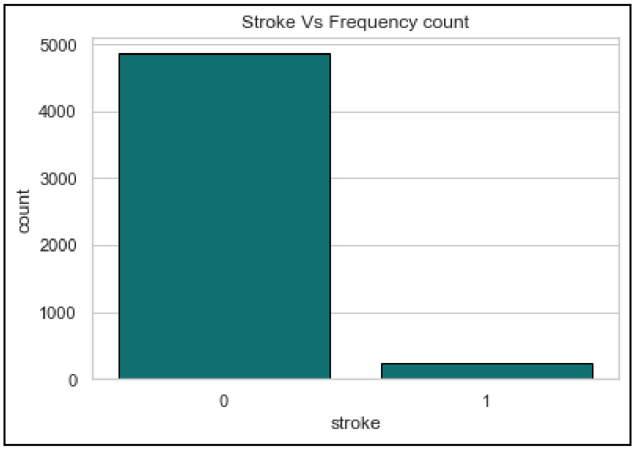
Exploratory Data Analysis

Before moving toward machine learning modeling and data preprocessing, we are doing some exploratory data analysis to fetch insights and information from our data. Following is the useful information which we have got.

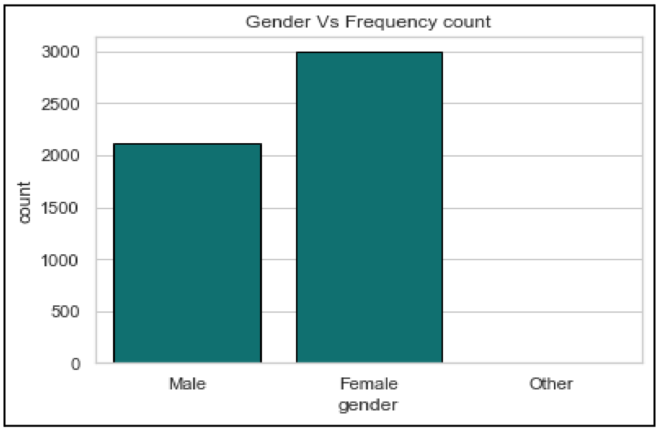
At first, we are plotting the distribution of the BMI column, which is showing that mean of the distribution is around 30. Our distribution is following a bell curve so we can say that most of the people in our dataset have **a** BMI **of** around 30.



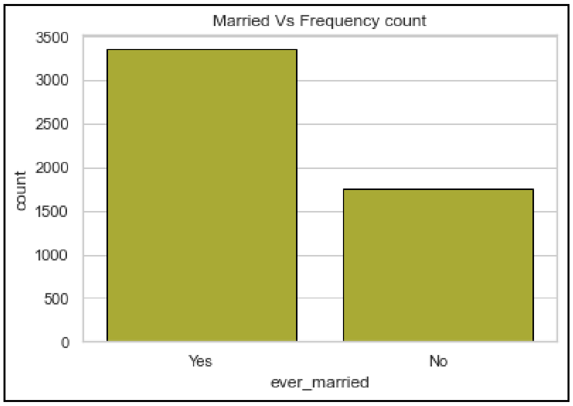
Now we are checking the distribution of our class label “stroke”. We can see that number of patients having stroke problems is deficient compared to the people having no stroke issues. So, our class is highly imbalanced as both classes are not equal.



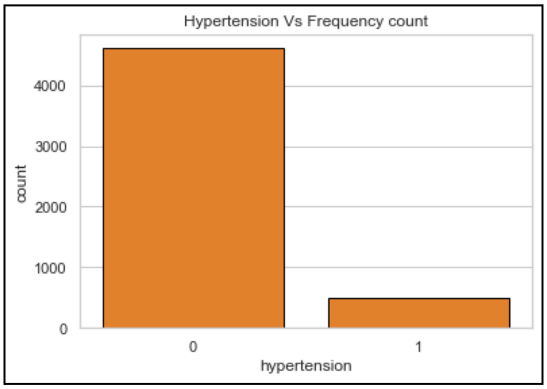
Most of the people in our dataset are females followed by males. Only 2 genders are present in our dataset. The number of females is around 3000, whereas males are around 2100.



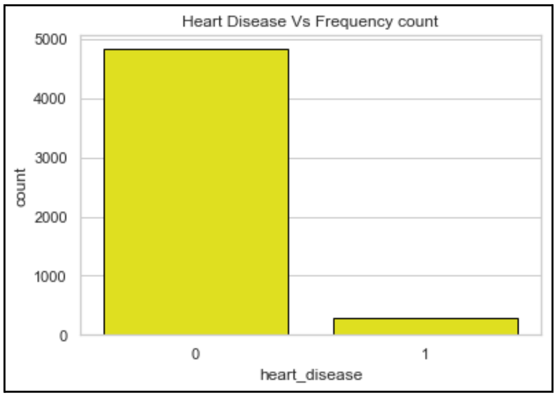
The marital status of most of the people in our dataset is married, as we can see in the bar graph below. The numbers of married people are around 3400 which is almost double the number of unmarried people.



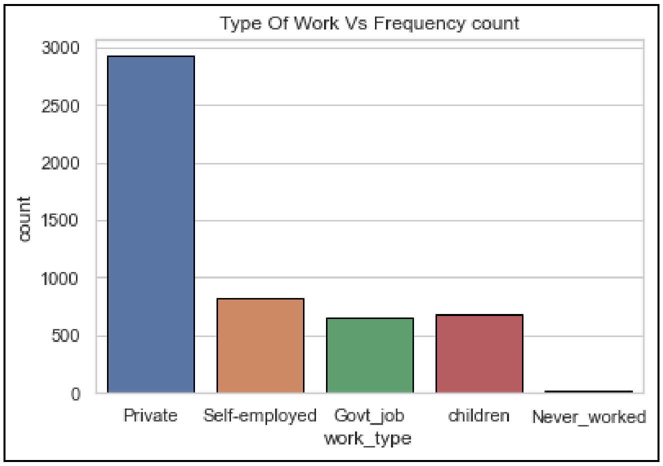
The majority of people in our dataset don't have any hypertension issues. Only a few people are having this issue.



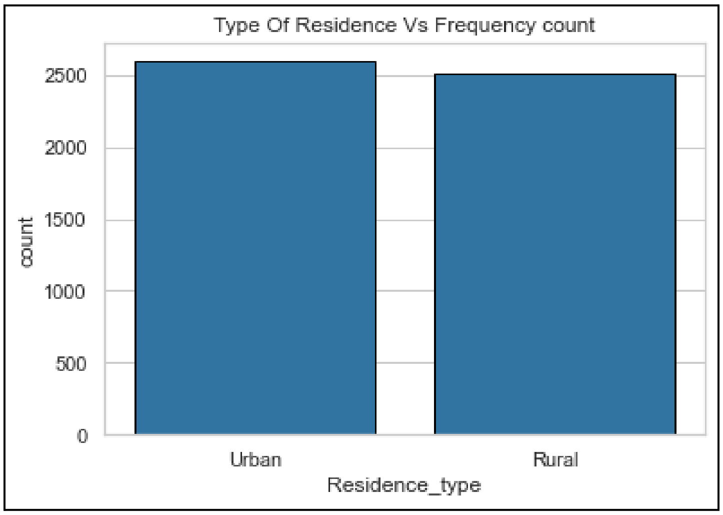
Very few people in our dataset are facing heart disease, which is another major disease and the leading cause of death worldwide.



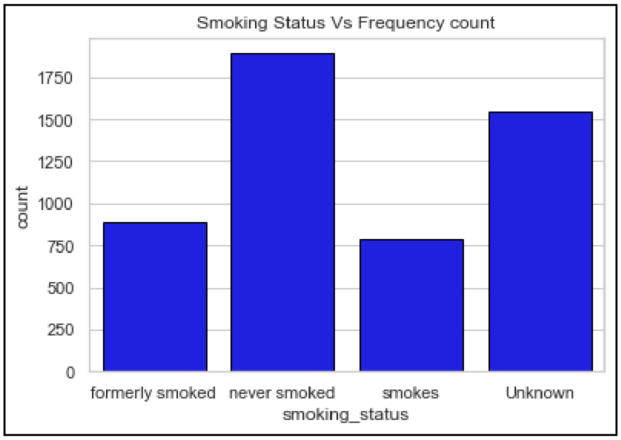
Most people are private employees in some organizations, and some of them are Government workers and self-employed. Around 750 children are also presented in the data.



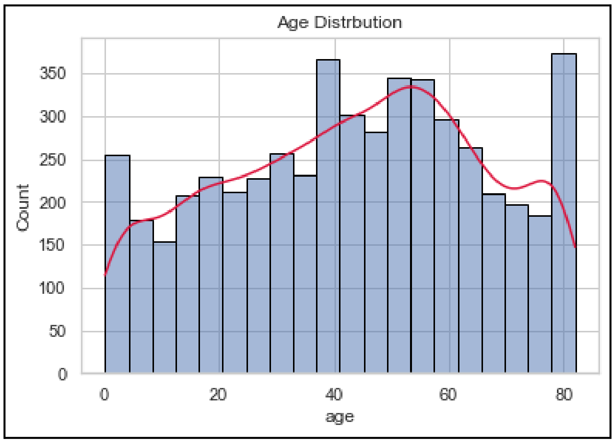
People from both rural areas and urban areas are present in the dataset. Almost equal numbers of people are from rural and urban backgrounds.



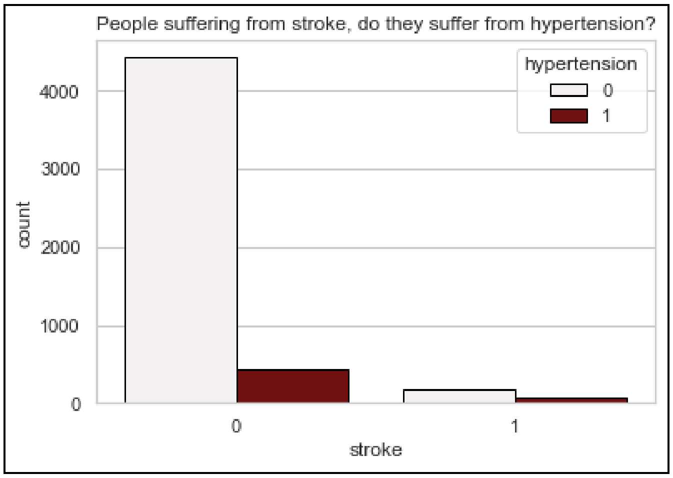
Most people are the ones who never smoked in their life. While there is a subsequent amount of people whose status is unknown whether they smoke or not. Whereas around 750 people are smokers.



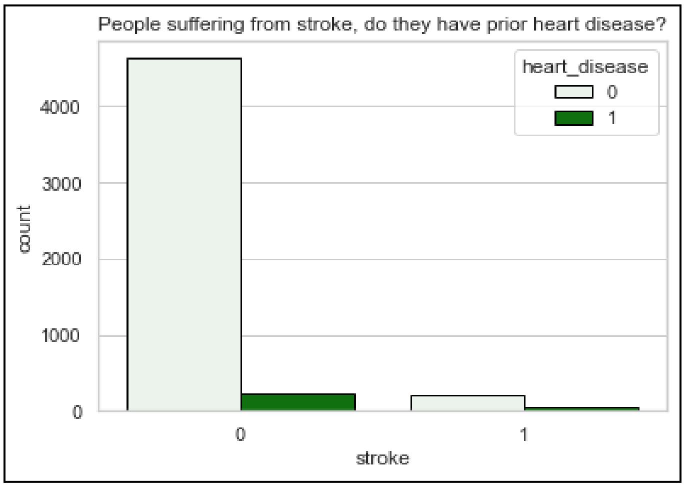
People from every age group are added to the dataset. We can see the age distribution below. Most of the people are having age around 55 years.



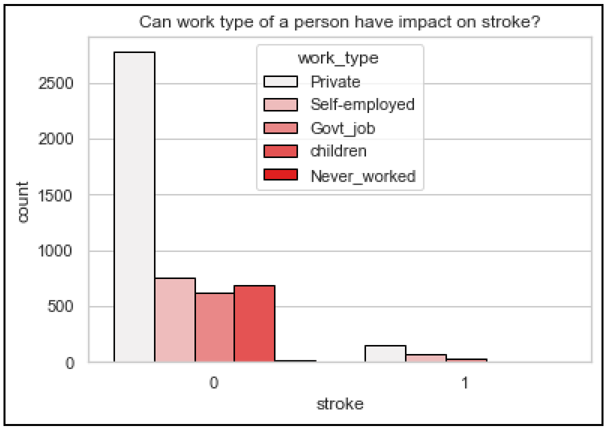
People who are having hypertension issues and who don’t have hypertension issues, both can face a stroke attack. The chances of stroke among people having this hypertension issue are a little high.



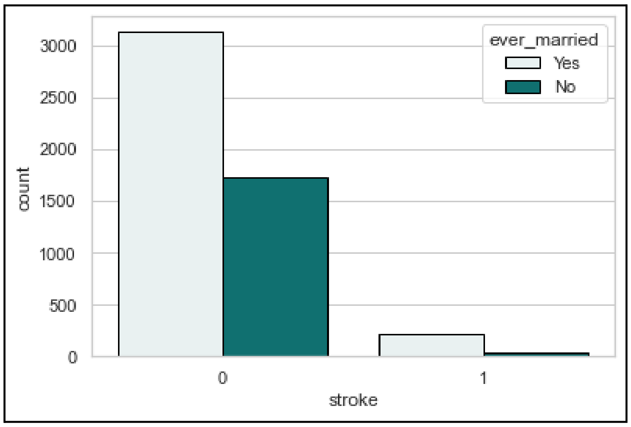
People who are having heart issues and who don't have a heart issues, both can face a stroke attack. The chances of stroke among people having this hypertension issue are a little high.



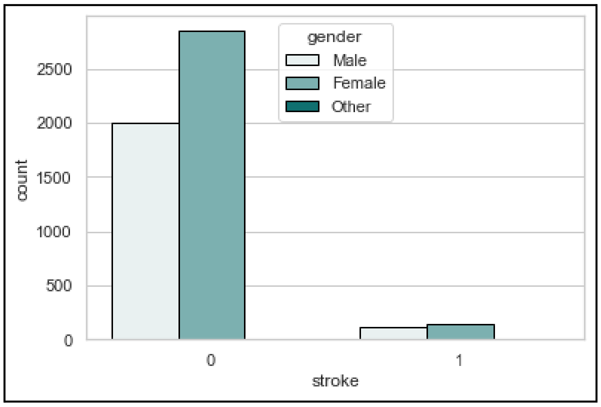
Children and people who never worked have almost zero chances of stroke attack as we can see in the graph below. From this, we can say that people with more hectic jobs are having more chances of stroke attacks.



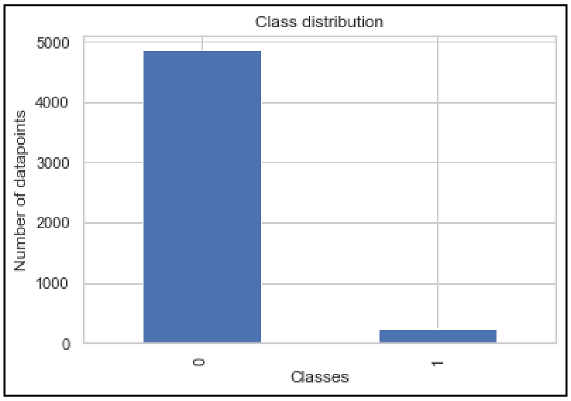
People who are married are having a high chance of stroke attacks as compared to unmarried people.



Males are having a higher chance of stroke attacks as compared to women in marital status.

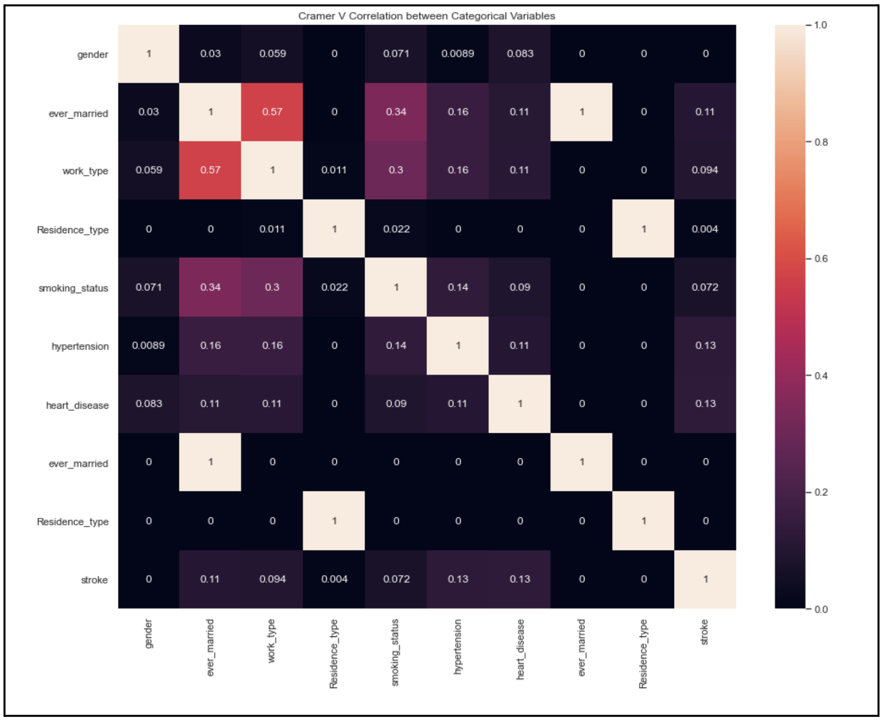


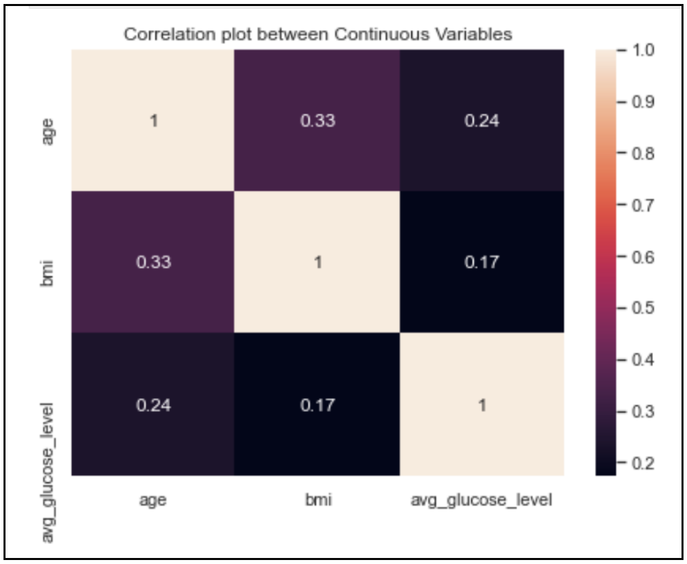
Our class label is an imbalanced column. Only a few cases of strokes are added to the data. We can see the distribution of the class below.



Correlation Matrix:-

Here we get to see the correlation coefficient between variables with the help of the Chi-Square test. Chi-square tests involve checking if observed frequencies in one or more categories match expected frequencies. It is used for goodness of fit for single measurement variables.





There is no significant correlation with respect to stroke as nothing exceeds 0.70. Work type and smoking status show the most significant correlation. We see a stronger correlation between ever-married and work type.

Data Preprocessing

Before using the data in machine learning algorithms, preparing the data is a crucial step. Data preprocessing is a data mining technique used to turn raw data into a format that is both practical and effective.

For our dataset, we also must perform some data pre-processing. First, we load our dataset using pandas and check the shape of our dataset which is 5110 and 11 columns. Dataset is normalized and contains both integer and categorical values. Data has some null values in the BMI column which we replaced by adding an average of values. After that, we dropped some unnecessary columns such as index id.

1) One-Hot Encoding: -

One-hot encoding is a process of converting categorical data variables so they can be provided to machine learning algorithms to improve predictions. One hot encoding is a crucial part of feature engineering for machine learning. our machine learning algorithms only understand numbers, so we must provide them with numbers by converting categorical variables into continuous variables. It Generates different columns and assigns them binary values.

One hot encoding is essential before running a machine learning algorithm on the data set. Some algorithms can understand categorical data directly such as decision trees but most of the supervised learning algorithms cannot operate on categorical data, they require all input variables to be numeric and generate output in numeric value. This technique of transforming columns into binary variables data set is quite famous in Supervised learning algorithms. These binary variables are also known as dummy variables in statistics.

2) Scaling: -

We want to normalize our data, so we are performing standard scaling. So that our data come into one range, and it also improves the performance of the model.

3) Smote Balancing: -

As we have discussed earlier, our dataset is highly imbalanced as most of the records belong to one specific class. So, we must balance our dataset and for that, we are using SMOTE technique to make this dataset balanced.

Train-Test-Split: -

When machine learning algorithms are used to make predictions on data that was not used to train the model, their performance is estimated using the train-test split technique. Here we are splitting our dataset into train dataset and test dataset with the ratio of 80-20% where 80% is our train dataset and 20% is our test dataset.

METHODOLOGY

We are using different machine learning algorithms in this project for Stroke prediction and comparing the results of each algorithm to check out which model is fitted and performing well on the training data set.

First, we are applying our models and check the performance of our models. The model we have used are Decision Tree, XGBoost, Random Forest, K- Nearest Neighbor, AdaBoost, and Logistic regression.

MODEL EVALUATION

We are evaluating the performance of our models using accuracy, confusion matrix, AUC graph, ROC_AUC score, sensitivity, specificity, miss rate, precision, and recall values. Before we move forward to model implementation, it is necessary to know the following terms.

1) Precision: -

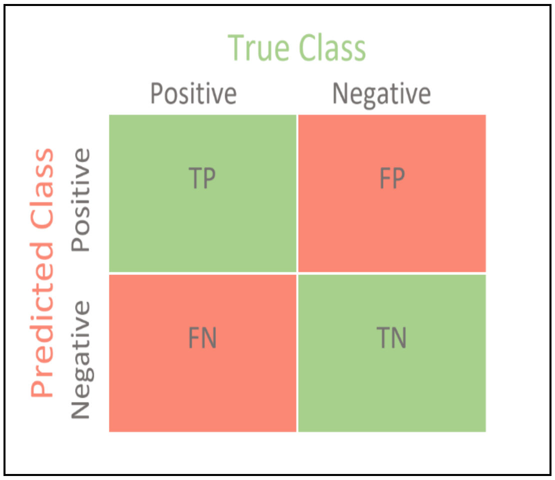
   The proportion of True Positives to All Positives is known as precision. For our problem statement, that would be the measure of cases that we correctly identify having a disease out of all the cases having it.

2) Recall: -

   The recall is the number of our model rightly identifying True Positives values. Thus, for all the cases who have the disease, recall tells us how many we correctly identified as having a disease.

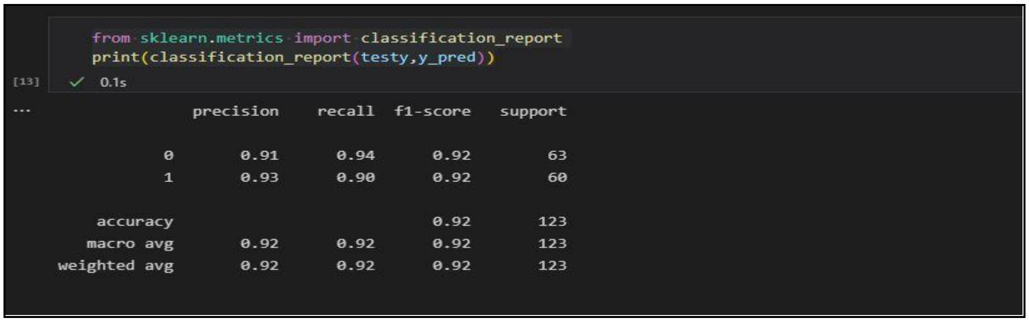
3) Confusion Matrix: -

A classification problem's prediction outcomes are compiled in a confusion matrix. Count values are used to describe the number of accurate and inaccurate predictions for each class. This is the confusion matrix.



4) Classification Report: -

The accuracy of predictions made by a classification algorithm is evaluated using a classification report. How many of the forecasts came true, and how many were wrong? More specifically, the metrics of a categorization report are predicted using True Positives, False Positives, True Negatives, and False Negatives.



```
from sklearn.metrics import classification_report
print(classification_report(testy,y_pred))
```

[13] ✓ 0.1s

```
              precision    recall  f1-score   support

           0       0.91      0.94      0.92        63
           1       0.93      0.90      0.92        60

    accuracy                           0.92       123
   macro avg       0.92      0.92      0.92       123
weighted avg       0.92      0.92      0.92       123
```

5) F-1 Score: -

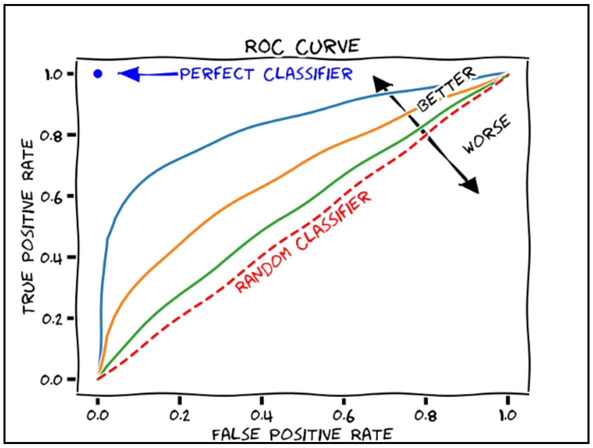
The accuracy of a model on a dataset is gauged by the F-score, also known as the F1-score. It is employed to assess binary categorization schemes that divide examples into "positive" and "negative" categories.

The precision and recall of the model are combined through the F-score.

6) ROC – AUC Curve: -

A measurement tool for binary classification issues is the Receiver Operator Characteristic (ROC) curve. In essence, it separates the "signal" from the "noise" by plotting the TPR against the FPR at different threshold values. The capacity of a classifier to differentiate between classes is measured by the Area Under the Curve (AUC), which is used as a summary of the ROC curve.

The model performs better at differentiating between the positive and negative classes the higher the AUC. The classifier can accurately discriminate between all Positive and Negative class points when AUC = 1. The classifier would be predicting all Negatives as Positives and all Positives as Negatives, however, if the AUC had been 0.



Model Implementation

Now we are implementing our machine learning model after doing all the required preprocessing on the data.

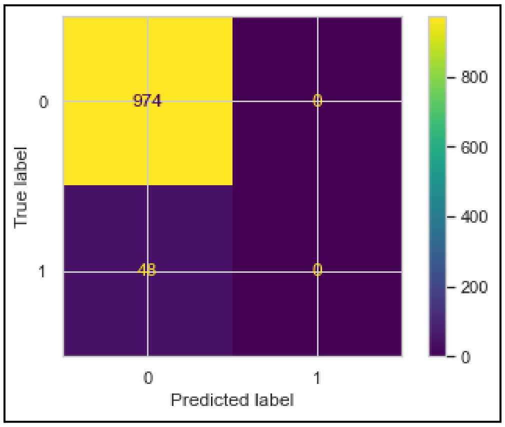
1) Logistic Regression: -

The first model we implement is a classification model called the logistic regression model, which is frequently used to foretell whether a given instance belongs to a particular class or not. It employs the sigmoid function, which returns a number between 0 and 1, to predict the probability of a specific class. In our situation, logistic regression provides accuracy on our test data set of more than 90%. With an F1 value of 0.97, it provides a precision of 0.95, which is good.

Features	**Without feature selection**
Model	Logistic Regression
Accuracy	95%
F1 Score	0.97
Recall	1.0
Precision	0.95

Confusion Metrics for Logistic Regression: -

Following are the results of the confusion matrix on logistic regression.



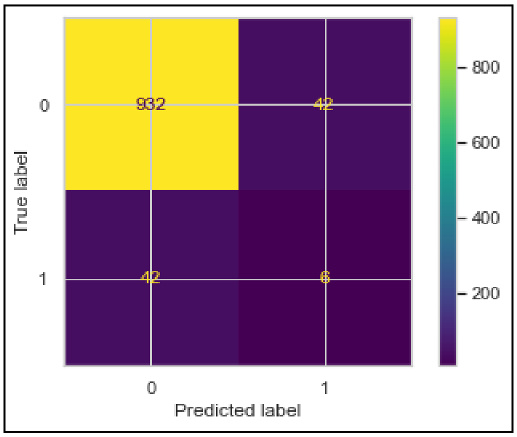
2) Decision Tree: -

The second model we're using on our dataset is a decision tree that uses decision trees to classify results. It is a well-known machine learning algorithm that is used for both classification and regression issues. Although we used a decision tree on several criteria in our research, the outcome of the F1 score prediction was satisfactory. Nevertheless, the accuracy of the decision tree was 91.7% which is a little less than logistic regression.

Features	Results
Model	Decision Tree
Accuracy	91.7%
F1 Score	0.95
Recall	0.95
Precision	0.95

Confusion Metrics for Decision Tree: -

Following are the results of Confusion metrics on the Decision Tree. Here we can observe that our model is correctly predicting 932 True positives and 6 true negatives. While there are also some false negatives and false positives, which means values that are wrongly predicted by the classifier.



3) K- Nearest Neighbor: -

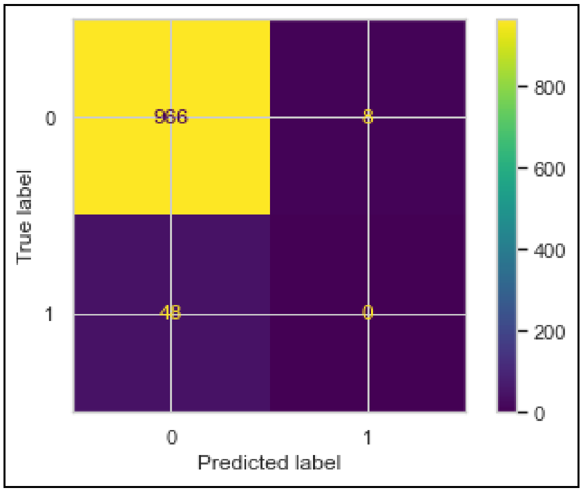
The supervised machine learning approach known as the k-nearest neighbors (KNN) model is straightforward and simple to apply. It can be used to tackle classification and regression issues. KNN searches for nearby nodes and returns results based on them. It determines its closest neighbors by calculating the distance to various neighbors. According to the similarity in a particular group of nearby data points, KNN classifies data points.

We are also using the KNN model in this project; the performance of KNN is also good on this dataset. It's giving an accuracy of around 94%.

Features	**Results**
Model	KNN
Accuracy	94.0%
F1 Score	0.97
Recall	0.99
Precision	0.95

Confusion Metrics for KNN Classifier: -

Following are the results of confusion metrics of KNN Classifier. Here we can observe that our model is correctly predicting 966 True positives.



4) Random Forest Classifier: -

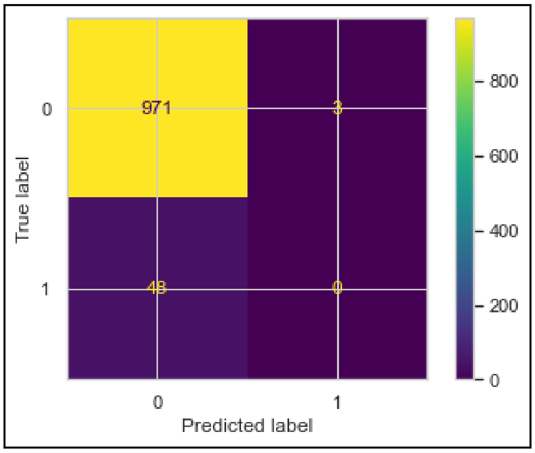
One of the well-known supervised learning methods is random forest. It's an ensemble model that integrates the results of various decision trees to make a learner that is more potent. RF is resistant to overfitting and has a strong noise resistance. The two basic ideas behind Random Forest are bagging and random selection.

In our project, the performance of random forest is quite well. It is giving us an F1 score better than the previous model maybe because it is good at handling a large number of features. It is good with high dimension data as we know that it works with making subsets by replacement. So, maybe this is the reason it is giving us good accuracy and roc score on the training dataset.

Features	**Results**
Model	Random Forest
Accuracy	95.0%
F1 Score	0.97
Recall	0.99
Precision	0.95

Confusion Metrics for Random Forest: -

Following are the confusion metrics of Random Forest Classifier. Here are model is predicting 971 true positive means values which are actually true. Which is actually a very good performance on a binary class classification problem.



5) Ada Boost: -

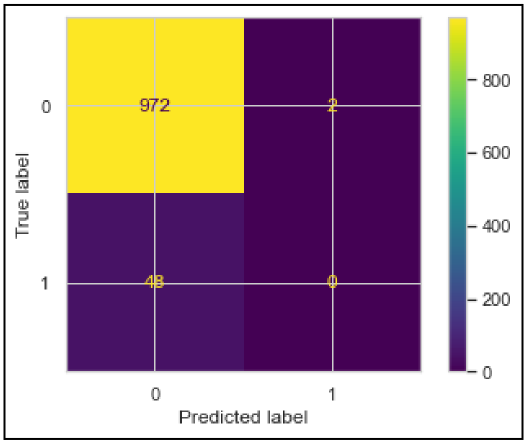
The supervised machine learning algorithm Ada boost is used to solve classification and regression problems. By transforming weak learners into strong learners, we can enhance the model prediction of any given algorithm using this ensemble technique. The weak student is successively corrected by predecessors, becoming a strong learner as a result. When compared to other algorithms, gradient-boosting classifiers are frequently quick and require less storage.

In this project, we also use the AdaBoost classifier and it's giving us optimum results as compared to other algorithms. AdaBoost performs as well as a random forest because it converts and enhances weak learners by reducing its error, but if the data is noisy then the accuracy of the gradient may affect sometime. We implement gradient boosting and it's fitting quite well.

Features	Results
Model	Ada boost
Accuracy	95.1%
F1 Score	0.974
Recall	0.99
Precision	0.95

Confusion Metrics for Ada Boost Classifier: -

In AdaBoost, we can see its rightly predicting 972 value as correct, which is highest by any model till now even more than Random Forest classifier.



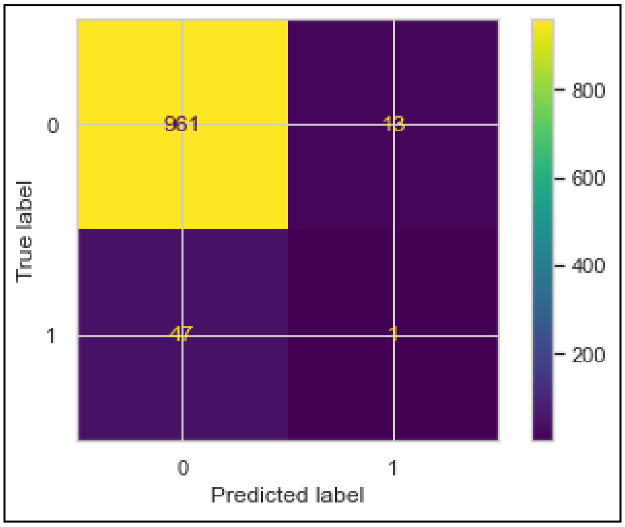
6) XGBOOST: -

XGBOOST is a decision-tree-based ensemble Machine Learning algorithm that uses a framework. In our case study, XGboost is performing well and giving us an accuracy of 94.1% with a good F1 score, precision, and recall, but it is a little less than AdaBoost.

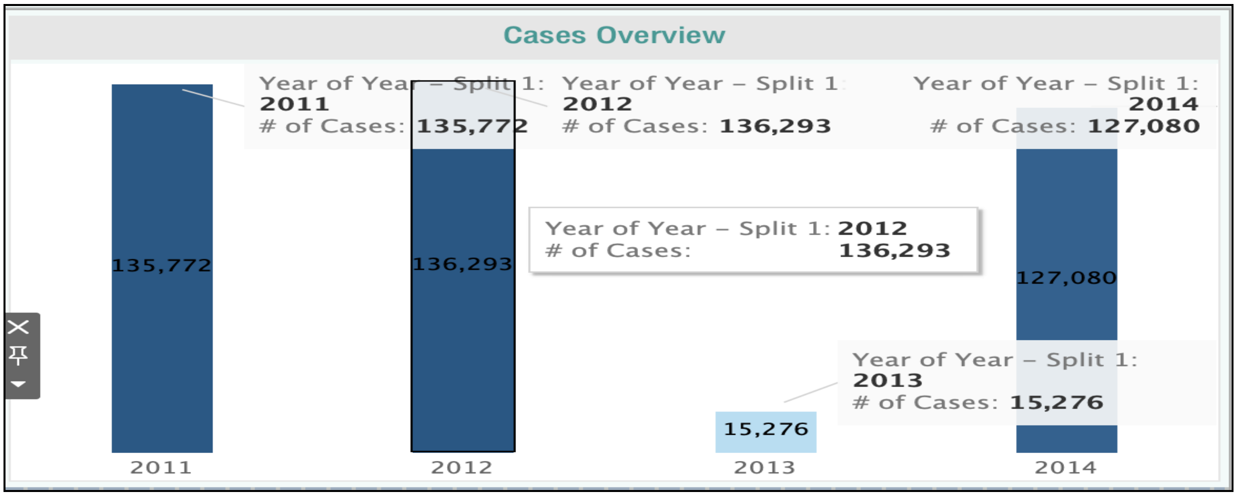
Features	Results
Model	XGboost
Accuracy	94.1%
F1 Score	0.96
Recall	0.98
Precision	0.95

Confusion Metrics for XGBoost: -

Xgboost is predicting 961 values positive which are positive and predicting 47 values as negative which are actually positive. Its prediction is good but a little less than Random Forest and AdaBoost.

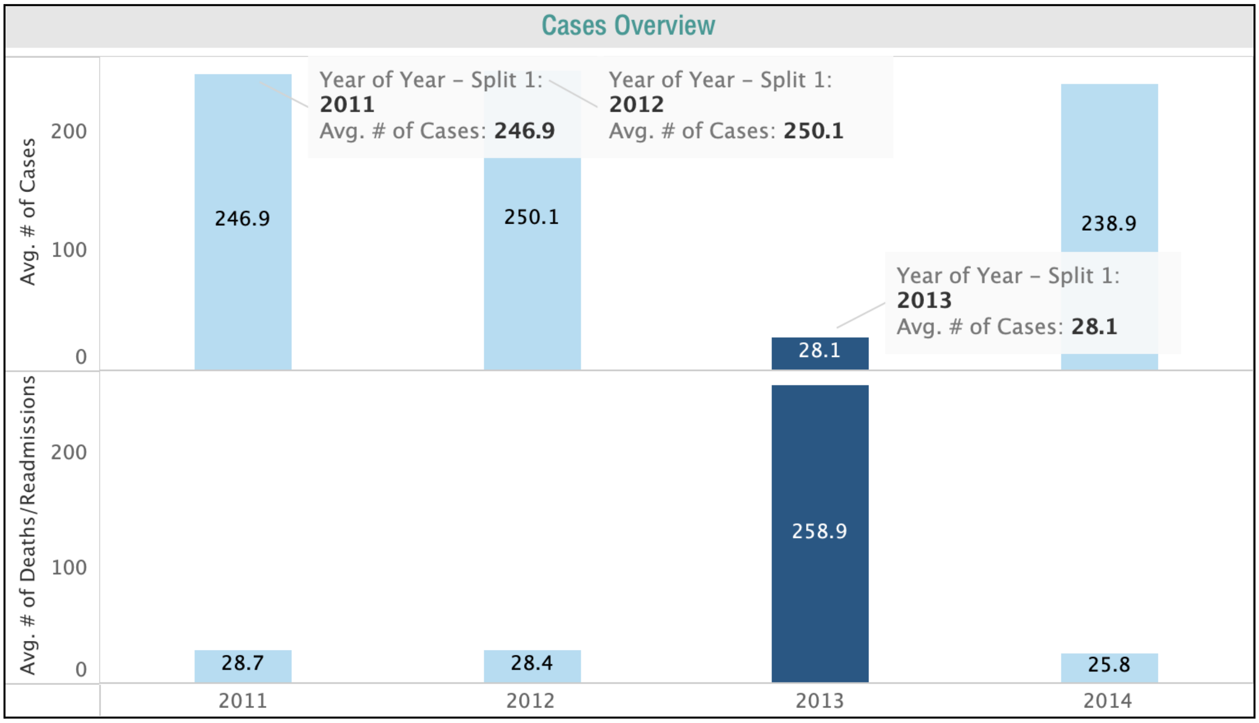


Dataset II:

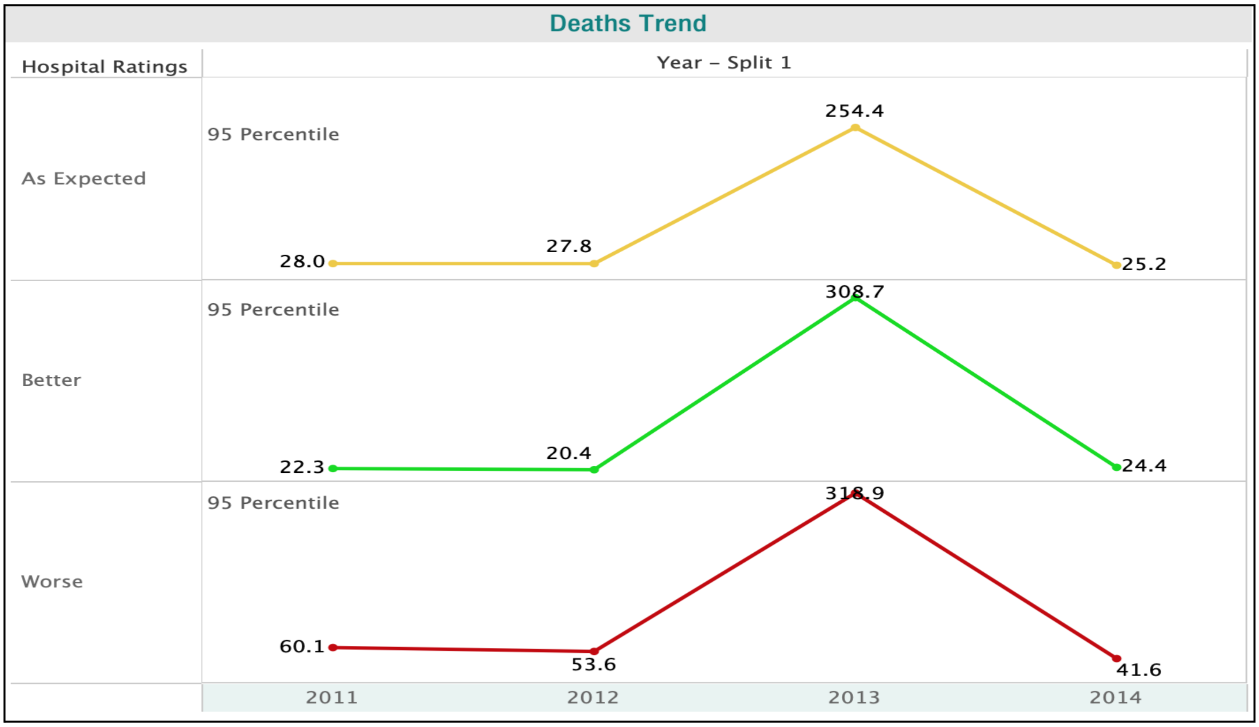


Data for a period of 2011 to 2015 of ischemic stroke 30-day mortality and 30-day readmissions.

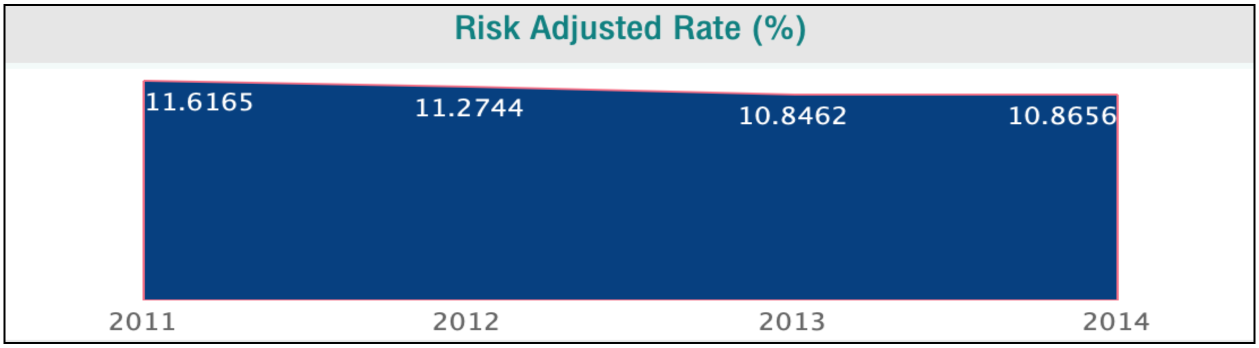
Average of # of Cases for each Year - Split 1 Year. The data is filtered on Hospital Ratings, which keeps As Expected, Better and Worse.



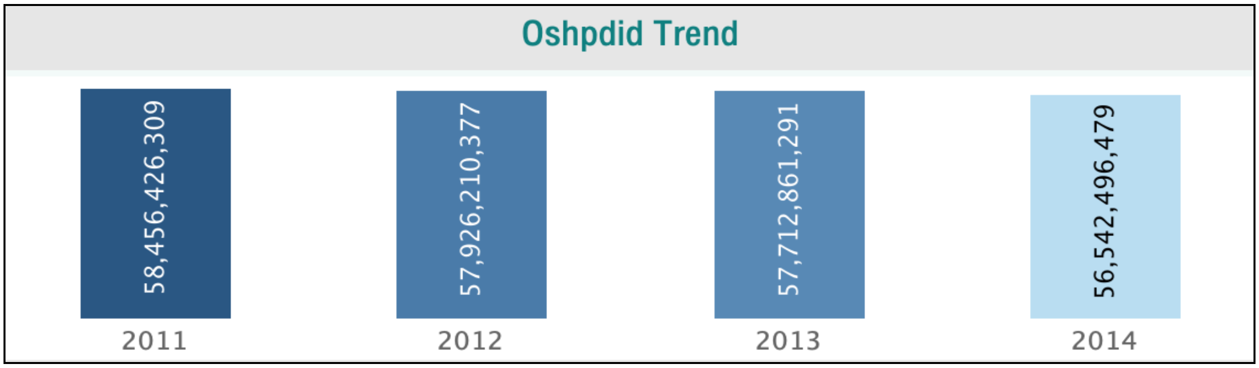
Here, we see the average number of cases was more with higher average death or readmission. Readmission may occur due to poor initial care or subsequent care, early discharge, or insufficient follow-up leading to readmission.



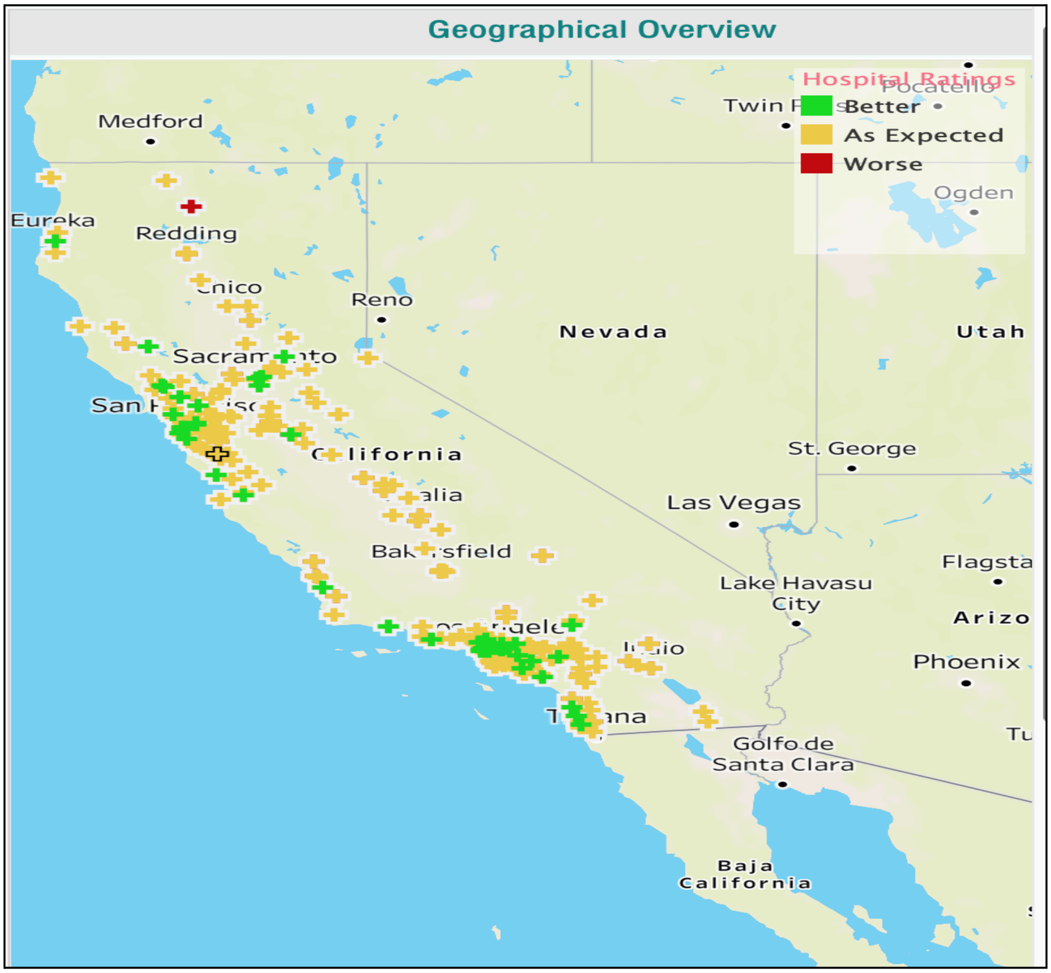
The trend of the average of # of Deaths/Readmissions (actual & forecast) for Year - Split 1 Year broken down by Hospital Ratings. The color shows details about Hospital Ratings. The marks are labeled by the average of # of Deaths/Readmissions (actual & forecast). Details are shown for the Forecast indicator.



The average risk-adjusted rate trend for Year - Split 1 Year. The color shows details about an average of Risk-Adjusted Rate. The marks are labeled by the average of the Risk-Adjusted Rate. Risk-adjusted mortality outcomes are feasible, reliable, and valid measures of patient outcome



Average of Oshpdid for each Year - Split 1 Year. The color shows details about Clusters. The marks are labeled by the average of Oshpdid. Outcome measures appear to be less reliable or not predictable using existing data at OSHPD



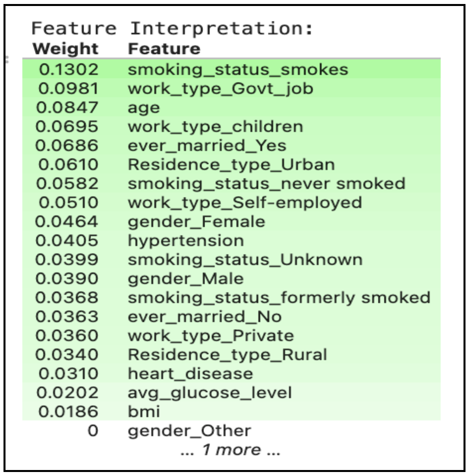
A Hospital’s performance is classified as

- Better: the upper 98% Confidence Interval (CI) of the risk-adjusted rate falls below California's observed rate (10.55% for RAMR, 12.80% for RARR).
- Worse: if the lower 98% CI of the risk-adjusted rate is higher than California's observed rate.
- As Expected: (rating is blank) if California observed rate falls within the 98% CI of the hospital risk-adjusted rate.

The geographical graph is plotted and we can see the regions in which the hospitals are classified as “Better, As Expected and Worse.” The locations are accurate, specifying the California state.

Results and Insights

After applying all the models, we have concluded that the AdaBoost classifier is performing best in terms of accuracy and other metrics. Followed by Random Forest and XGboost.



Feature Interpretation:

Weight	Feature
0.1302	smoking_status_smokes
0.0981	work_type_Govt_job
0.0847	age
0.0695	work_type_children
0.0686	ever_married_Yes
0.0610	Residence_type_Urban
0.0582	smoking_status_never smoked
0.0510	work_type_Self-employed
0.0464	gender_Female
0.0405	hypertension
0.0399	smoking_status_Unknown
0.0390	gender_Male
0.0368	smoking_status_formerly smoked
0.0363	ever_married_No
0.0360	work_type_Private
0.0340	Residence_type_Rural
0.0310	heart_disease
0.0202	avg_glucose_level
0.0186	bmi
0	gender_Other
	… 1 more …

Following are the features which are most impactful on our model. As we can see, “work type government”, age, and unknown smoking status have a lot of impact on our models. The least impactful feature is gender_other because there is no other gender in the dataset. Other than that, BMI also doesn’t have much impact on the model.

Results II:

As expected, in general, 30-day mortality rates are higher than the inpatient death rates. Inpatient mortality rates appear to be biased. Readmission models that do not account for death may generate misleading results since mortality is a more likely event than readmission.

My Tableau Public Dashboard:
https://public.tableau.com/app/profile/priya.johny7090/viz/StrokeSummaryFinale/Main#guest=n&2

CONCLUSION

Most of the models are performing very well in terms of accuracy, precision, recall, etc. our dataset. AdaBoost is performing very well with the best accuracy of 96%. We can also try different techniques such as feature selection and parameter tuning to check out the effect and to increase the accuracy of the models.

Ischemic stroke is a common, severe acute illness with high short-term mortality and prolonged recovery. Patients are seen in the vast majority of acute care hospitals in California that report data to OSHPD.

Thus reporting on stroke outcomes in California hospitals will reflect the care delivered in most California hospitals. Early diagnosis and prevention of hypertension, smoking habits, and living a healthy lifestyle are essential steps to reducing stroke risk and disabilities.

REFERENCES

1. https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset
2. https://www.cdc.gov/stroke/index.htm
3. https://www.who.int/southeastasia/news/detail/28-10-2021-world-stroke-day

4. https://www.cambridge.org/core/journals/international-psychogeriatrics/article/abs/general-and-diseasespecific-risk-factors-for-depression-after-ischemic-stroke-a-twostep-cox-regression-analysis/6686993A3472AB5518857D3CBD6D8755
5. https://www.emro.who.int/health-topics/stroke-cerebrovascular-accident/index.html
6. State of California, Office of Statewide Health Planning and Development. The California Report on Coronary Artery Bypass Graft Surgery 2005-2006 Hospital and Surgeon Data, Sacramento, CA: Office of Statewide Health Planning and Development, March 2009.
7. Wong, K.S., Risk factors for early death in acute ischemic stroke and intracerebral hemorrhage: A prospective hospital-based study in Asia. Asian Acute Stroke Advisory Panel. Stroke, 1999. 30(11): p. 2326-30.
8. https://link.springer.com/article/10.1007/s00415-020-09932-y