CS 725
From Theory to Application: Project Roadmap

Anuj Attri (23M0808) Arnav Attri (23M0811)

Satheeshkumar M (23D0198) Piyush Paliwal (30005245)

October 31, 2023

Project Title

Strokes Uncovered: *Machine Learning Exploration and Predictive Insights for Stroke Prevention*

In response to the TA's valuable feedback*: In response to the valuable feedback and guidance provided, we have made revisions to the project's abstract, as detailed below. The* **revised abstract** *now explicitly highlights our use of a Deep Learning Model, specifically a Convolutional Neural Network (****CNN****), for analyzing brain CT scan images.*

Abstract

Stroke, a cerebrovascular disease , is the second leading cause of death worldwide, carrying significant health and economic implications. The project, titled "Strokes Uncovered," presents an in-depth exploration of machine learning techniques applied to the prediction and prevention of strokes, with a focus on addressing the *feedback* received from our ***TA***.

We acknowledge the significance of traditional machine learning models in predicting strokes based on structured data from **CSV files**. However, we recognize the opportunity to advance our research by incorporating *deep learning techniques*, specifically *Convolutional Neural Networks (**CNN**)*, which were recently introduced by our *professor* in the class and we employ that with particular emphasis on analyzing Brain CT scan images.

Our enhanced project scope will not only encompass the *traditional*

ML models for stroke prediction but also extend into the realm of medical imaging analysis. By utilizing ***CNN***, we intend to harness the power of *image data* to enhance our predictive models. This innovative approach seeks to leverage the information extracted from ***Brain CT scan*** images to gain a deeper understanding of the intricate relationships between stroke risk factors and the physiological characteristics of the brain.

This expansion into *deep learning and medical imaging analysis* will introduce a novel dimension to our project, differentiating it from the traditional CSV-based dataset approaches. It will allow us to explore *uncharted territory*, pushing the boundaries of stroke prediction research while demonstrating originality and relevance to the class.

I*n summary**, "Strokes Uncovered" continues to fuse machine learning techniques with exploratory data analysis, but now, it embraces the challenge of incorporating *deep learning models* in the form of CNN. This step forward not only demonstrates our *adaptability and responsiveness* to *feedback* given by the ***TA but also highlights our commitment to breaking ground for new approaches in healthcare and risk assessment.

1 Relevance of Our Project

To further underscore the *importance of our project*, we reference recent high-profile news articles:

- In a recent article by ***The Washington Post***, published on **October 19, 2023**, experts predict a potential ***50***% increase in stroke-related deaths by the year 2050. This alarming projection highlights the urgency of developing effective strategies for stroke prevention[1].

- ***The New York Times***, in an **October 18, 2023** report, shed light on the higher incidence and severity of strokes in women. This emphasizes the need for targeted prevention efforts to address this gender-specific health concern[2].

- Additionally, ***Forbes***' recent piece on **October 29, 2023**, emphasizes the stark reality that, on World Stroke Day, a staggering 33,425 individuals are estimated to suffer from strokes, with only 1,000 of them receiving life-saving treatments. This underscores the critical need for improved stroke prediction and intervention strategies[3].

These *high-profile news articles* serve to highlight the pressing global concerns

[1] `https://www.washingtonpost.com/health/2023/10/19/stroke-deaths-could-double-2050/`

[2] `https://www.nytimes.com/2023/10/18/well/live/stroke-prevention-risk-women.html`

[3] `https://tinyurl.com/37pmsbn8`

related to strokes and reinforce the importance of our project's focus on stroke prediction and prevention.

2 Solution Blueprint: An Insight into the Proposed Solutions

1. ***Laying the Foundation***: Configuring the Environment and Libraries.
2. ***Data Narratives***: Unveiling Insights through EDA.
3. ***The Art of Data Cleaning***: Best Practices in Preprocessing.
4. ***Balancing Act***: Visualizing Data Imbalance and Sampling Techniques.
5. ***Beyond Accuracy***: In-Depth Analysis of Machine Learning Model Evaluations.
6. ***Deploying the Stroke Model***: A User-Friendly Web Application.
7. ***Mind Matters***: Holistic Approaches to Stroke Prevention and Well-being.

3 Expedition: Unveiling Key Repositories

In this section, we provide ***links to the codebases*** for the various *machine learning models* and ***tools*** used in our project. These *codebases* have been instrumental in implementing our *solution* approach.

3.1 *Logistic Regression*

- *Codebase*: `https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html` (sklearn.linear_model.LogisticRegression)

3.2 *K-Nearest Neighbors (KNN)*

- *Codebase*: `https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html` (sklearn.neighbors.KNeighborsClassifier)

3.3 *Decision Tree*

- *Codebase*: `https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html` (sklearn.tree.DecisionTreeClassifier)

3.4 *Random Forest*

- *Codebase*: `https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html` (sklearn.ensemble.RandomForestClassifier)

3.5 *XGBoost*

- *Codebase*: `https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html` (sklearn.ensemble.GradientBoostingClassifier)

3.6 *Stacking Model*

- *Codebase*: `https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.StackingClassifier.html` (sklearn.ensemble.StackingClassifier)

3.7 *Convolutional Neural Network (CNN) with TensorFlow*

- *Codebase*: `https://www.tensorflow.org/tutorials/images/cnn` (TensorFlow Core)

These links will *re-direct* to the relevant *code repositories* for each ML model. We will explore these repositories to find code examples and resources for implementing the specified models and tools in our *Jupyter Notebook* using *Python*.

4 *A* Tale of Two Datasets

4.1 *First Dataset* - Patient Information

This dataset is publicly available on Kaggle has more than *5000 tuples* and contains the following *attributes* in a ***CSV*** file:

1. **ID**: A unique identifier for each patient.
2. **Gender**: Categories include "Male," "Female," or "Other."
3. **Age**: The patient's age.
4. **Hypertension**: A binary indicator where 0 means the patient doesn't have hypertension, and 1 means they have hypertension.
5. **Heart Disease**: A binary indicator where 0 indicates the absence of any heart diseases, and 1 indicates the presence of a heart disease.
6. **Marital Status**: Two categories, "No" and "Yes," representing whether the patient is married or not.

7. **Work Type**: Categories include "Children," "Government Job," "Never Worked," "Private," or "Self-employed," denoting the patient's occupation.

8. **Residence Type**: Binary choice between "Rural" and "Urban" for the patient's place of residence.

9. **Average Glucose Level**: The average glucose level in the patient's blood.

10. **Body Mass Index (BMI)**: A measure of the patient's body mass.

11. **Smoking Status**: Categories include "Formerly Smoked," "Never Smoked," "Smokes," or "Unknown."

12. **Stroke**: A binary indicator with **1** representing that the patient had a *stroke* and **0** indicating *no stroke* history.

4.2 *Second Dataset* - Brain Stroke CT Image Dataset

This publicly available dataset on Kaggle contains two classes of images:

1. **Normal Brain Images**: It includes *1551 images* representing normal brain scans.

2. **Stroke Images**: This category consists of *950 images* depicting *brain scans* of patients who have suffered a stroke.

5 *U*nder the Hood: Model Implementation

5.1 Introduction

Within the *framework of our project*, the *solution* approach laid out plays a pivotal role in driving our project forward. In this *section*, we will dive deep into the details of our carefully crafted solution approach, highlighting how it addresses critical *project goals* and leads our journey toward *stroke prediction and prevention.*

5.2 Data Collection

In this *phase*, we meticulously collected and prepared two essential *publically* available datasets: *patient information* and *brain CT images.* The f*irst dataset* comprises comprehensive *patient records*, while the *second dataset contains brain CT scans of patients*, categorized into normal and stroke images. This robust foundation equips us for *in-depth* data analysis and model development.

5.3 Exploratory Data Analysis *(EDA)*

In this section, *we conduct exploratory data analysis (EDA)* to gain insights into various aspects of our dataset. We utilize ***Plotly Express (px) and Seaborn*** to create a range of ***visualizations*** that provide us with valuable *insights* into our dataset, enhancing our understanding of *key* factors related to stroke prediction.

5.4 Data Preprocessing

In the *data preprocessing* phase, we undertake several critical tasks to enhance the quality and suitability of our datasets for model training. These tasks include:

- ***Null Handling*****:** We address *missing data* using imputation strategies to ensure our machine learning models perform optimally.
- ***Outliers Handling*****:** To mitigate the impact of *outliers*, we employ outlier detection and removal techniques in ***BMI column*** to improve the model's resilience to anomalies.
- ***Duplicate Checking*****:** Rigorous *duplicate checking* is performed to maintain data integrity, ensuring that each data point is unique and contributes meaningfully to model training as duplicates in the dataset can ***skew*** the analysis.
- ***Category Data Encoding*****:** We transform *categorical attributes* into *numerical formats* suitable for machine learning by employing encoding techniques such as ***one-hot encoding or label encoding***.

These *preprocessing steps* are integral in *refining* our datasets, facilitating subsequent stages of *feature engineering* and *model training.* They are vital for achieving *reliable* and *accurate* stroke predictions.

6 Model Selection

For *brain stroke prediction*, we selected a diverse *ensemble* of ML models, including traditional models:

- **Logistic Regression:** A fundamental model for *binary classification.*
- **K-Nearest Neighbors (KNN):** Effective for *proximity-based* predictions.
- **Decision Tree:** Useful for *hierarchical data structure* interpretation.

Additionally, we incorporated ***advanced models*** for their ***ensemble capabil-***

ities and performance:

- **Random Forest:** A powerful *ensemble* method for improving accuracy.
- **XGBoost:** Known for its *gradient boosting* techniques.
- **Stacking Model:** Combining *multiple models* for enhanced predictions.

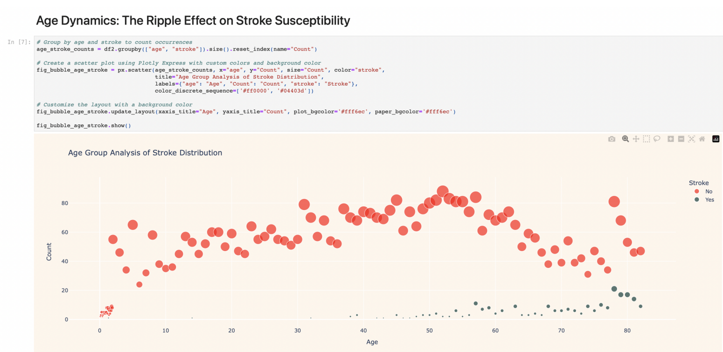
To leverage the potential of ***medical imaging data***, we included a specialized *deep learning model*:

- **Convolutional Neural Network (CNN) with TensorFlow:** Renowned for *image analysis*.

7 Preliminary Results

Up to this point, our focus has been on the E*xploratory Data Analysis (EDA)* phase, the results of which are now available on *our GitHub Repository* in the form of a *Jupyter notebook*[4].

For ***quick reference***, we are providing our work in the form of *snapshots* that can also be reproduced by executing the code from the *GitHub repository*.



[4] https://github.com/IITBCSE/Project

Inhale, Exhale, Analyze: The Smoking-Stroke Pie

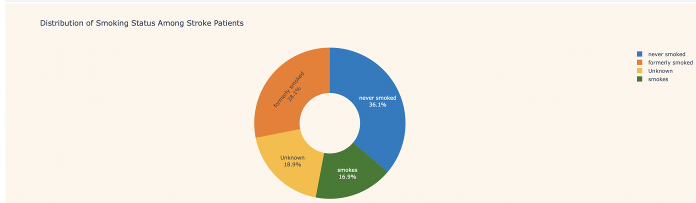
In [19]:
```python
smoking_stroke_counts = df2.groupby(['smoking_status', 'stroke']).size().unstack().fillna(0).reset_index()
custom_colors = ['#007CC3', '#F47A1F', '#FDBB2F', '#377B2B']


fig = px.pie(smoking_stroke_counts, names='smoking_status', values='Yes',
             title='Distribution of Smoking Status Among Stroke Patients',
             color_discrete_sequence=custom_colors)

fig.update_traces(hole=0.4, textinfo='percent+label')


# Customize the layout with a beige background
fig.update_layout({
    "plot_bgcolor": "#fff6ec",
    "paper_bgcolor": "#fff6ec"
})

fig.show()
```



Slice of Health: Hypertension by Smoking Habits

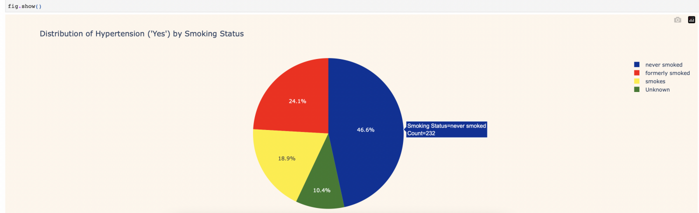
In [24]:
```python
yes_hypertension_df = df2[df2['hypertension'] == 'Yes']
hypertension_by_smoking = yes_hypertension_df['smoking_status'].value_counts().reset_index()
hypertension_by_smoking.columns = ['Smoking_Status', 'Count']

custom_colors = ['#003399', '#ff0000', '#ffec19', '#377B2B']


fig = px.pie(hypertension_by_smoking, names='Smoking_Status', values='Count',
             title="Distribution of Hypertension ('Yes') by Smoking Status",
             color_discrete_sequence=custom_colors,
             labels={'Smoking_Status': 'Smoking Status', 'Count': 'Count'})

# Customize the layout with a beige background
fig.update_layout({
    "plot_bgcolor": "#fff6ec",
    "paper_bgcolor": "#fff6ec"
})


fig.show()
```



Gender and Stroke: A Pie of Insights

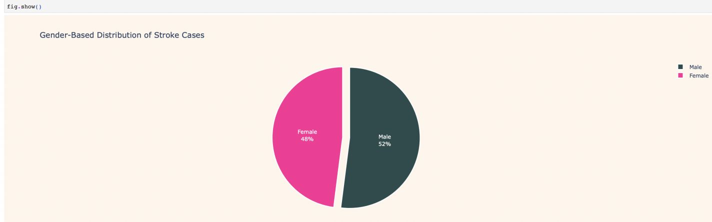
[6]:
```python
gender_stroke_counts = df.groupby(['gender', 'stroke']).size().unstack()
gender_stroke_percentage = (gender_stroke_counts[1] / (gender_stroke_counts[0] + gender_stroke_counts[1])) * 100

fig = px.pie(names=gender_stroke_percentage.index, values=gender_stroke_percentage.values,
             title="Gender-Based Distribution of Stroke Cases",
             color_discrete_sequence=[ '#2a4c4c', '#ff2797' ])

fig.update_traces(textinfo="percent+label", pull=[0.1, 0], marker=dict(line=dict(color="white", width=2)))

# Customize the layout with a beige background
fig.update_layout({
    "plot_bgcolor": "#fff6ec",
    "paper_bgcolor": "#fff6ec"
})

fig.show()
```



8 Semester Roadmap: A Technical Journey Towards Excellence

After successfully navigating through the complex *terrain* of *Exploratory Data Analysis (EDA)*, we're now embarking on an exciting journey to address the *significant challenges* in the field of *stroke prediction and prevention.* In the coming weeks, we'll focus on the following important technical milestones:

1. **Data Preprocessing:** We'll meticulously refine and *optimize our dataset*, ensuring that it's well-prepared for the subsequent modeling phase.

2. **Balancing Data Visualization:** Using advanced *data sampling techniques*, we aim to create a balanced dataset, guaranteeing fairness and accuracy in our analysis.

3. **Modeling and Evaluation:** This is the *high point of our expedition*, where we'll carefully build and fine-tune our *machine learning models* to predict strokes with exceptional precision.

4. **Model Deployment - Building a Web Application:** The *grand finale* of our technical journey will be the creation of an *interactive web application.* This application will allow users to directly experience the outcomes of our hard work and research.