# CSM148 Final Project

Due June 1st, 2021 at 10am PST via Gradescope

## Introduction:

According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths.
The total cost of stroke in the US was $103.5 billion according to 2016 US dollar values. $68.5 billion or 66% of total cost was accounted for by indirect cost from underemployment and premature death. Age groups 45-64 years accounted for the greatest stroke related direct cost.[1]

## Challenge:

This project is about being able to predict whether a patient is likely to get a stroke based on the input parameters available to us. Use features like gender, age, various medical conditions, and smoking status to build a model that helps you decide if a person is likely to experience a stroke event in the near future.

You will serve as data scientists hired by the UCLA hospital. You will be asked to develop a predictive model for this task and report out your findings to them. This project will include both a structured component, where much like Projects 1 and 2, you will be given a specific set of instructions to complete. There will also be an unstructured contest for you to complete as well where you will be competing against your classmates to achieve the best model.

## Project Overview:

This project will consist of several discrete components. Full credit for Project 3 will require your completion of all 3 components. Specifically you will be asked to produce or submit to the following:

- Report: A report documenting your work on the project and your findings
- Coding Project: Follow the steps detailed below on a Jupyter Notebook
- Kaggle Contest Submission

[1] A contemporary and comprehensive analysis of the costs of stroke in the United States

## Final Deliverables:

- PDF output of Jupyter Notebook (submitted via Gradescope)
- PDF of Final Report (Submitted via Gradescope)
- Kaggle Competition entry submission

## Timeline:

- Project will be released on **MAY 6th**
- Project will be due before the last class of the quarter (Tuesday, Week 10) on **June 1st** before **10am PST**

## Contest:

Once you have effectively trained your model, as a next step you will be asked to participate in a contest among your peers, hosted on Kaggle!
Your outputs will then be compared against the real labels and an accuracy score Generated. This score will be compared against your peers.
Bonus points on the project will be awarded to high-performing individuals.

## Project Requirements:

### Specific Coding Requirements to be completed **in no specific order** :

1. Run some basic statistics on your variables including correlations with labels and report findings - Particularly once you employ PCA and Neural Nets and other 'black box' methods, the descriptive power of any of your features will effectively disappear. Still you want to report out meaningful correlations to the doctors to help them flag key indicators they can employ (this step will also be helpful for you in flagging potential co-linearities).
2. Create a data feature extraction plan and implement a pipeline to execute it - Determine and execute a plan to process your data for modeling and then implement a pipeline to execute it. Specifically:
    a. Determine which fields to retain and which to drop.
    b. For those you retain, determine a categorization strategy.
    c. Determine an imputation strategy (you should choose more than one imputation method depending on the specifics of your data.)
    d. Augment at least one feature, ideally a feature cross, or non-linear transformation.
    e. Determine a strategy for scaling features.
    f. Implement a single pipeline to execute this transformation.

    g. Document your data strategy in your report. Provide an explanation or justification for why you chose the data you did, and also detail any experiments you ran and the results.
3. Implement a basic Logistic Regression - With your newly pipelined data find and interpret important features (e.g. using regression and associated p-values). If there are any collinearities be careful when incorporating them into the regression.
4. Implement Principle Component Analysis (PCA) - Since your resulting dataframe is likely to be high-dimensionality, employ PCA to reduce the complexity of your dataframe.
5. Employ an ensemble method to your classification exercise - Leveraging bagging or equivalent ensemble learning method to generate an optimized classification model.
6. Develop a Neural Net classifier - Modify parameters to optimize outcomes. Report your customized parameter settings in the report.
7. Cross-Validate your training results - Employ K-Fold Cross-validation to your training regimen for both ensemble and NN classifiers. (Optional: employ a stratifiedshufflesplit as well to ensure equitable distribution along a key parameter).
8. Experiment with your own custom models and report out your highest performing model. Submit the model to the class-wide contest. - For this part of the project you have free range to employ any of the tools you've learned in class, along with any additional tools or techniques you research independently.

## Report Requirements:

Each person will be expected to submit a report accompanying their project. There is no specific length or formatting requirement, but is expected to be professionally produced. Points will be deducted for incomplete or unprofessional reports.

The report will be expected to contain the following sections:

1. **Executive Summary**: Single-page high level summation of the work done and key findings.
2. **Background/Introduction**: Use the available information, along with your own (brief… very short) industry research, to better explain the domain challenges.
3. **Methodology**: Incorporate requirements from the coding requirements into a general description of the work that you have done on this project.
4. **Results**: Report out your results from coding requirements.
5. **Discussion**: Provide context to the results you've obtained. Additionally, provide a set of recommendations to UCLA hospital for how to leverage your findings along with next steps for analytic work.
6. **Conclusion**: concisely summarize the work done on the project.

# Contest Submission Requirements:

Once you have trained your own model, as a next step you will be asked to participate in a contest among your peers, hosted on Kaggle.
Go to: https://www.kaggle.com/c/cs148-spring-2021/, register for a kaggle account and join this contest.

Under the data tab access you can find ‘dataset-stroke.csv’ a file with data and a target column ‘stroke’. You can also find ‘dataset-stroke-eval.csv’, a file with 700 additional instances from the dataset with no labels.

Download both files, use the dataset to train your model and then download the eval file and pipeline it and plug it into your model in order to generate the predicted labels.
Output these predictions into a CSV file, using the format described on the contest page, and submit them to the contest. Your outputs will be compared against the real labels and an accuracy score generated. This score will be compared against your peers.
Bonus points on the project will be awarded to high-performing individuals.

## Attribute Information

1) id: unique identifier
2) gender: "Male", "Female" or "Other"
3) age: age of the patient
4) hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension
5) heart_disease: 0 if the patient doesn't have any heart diseases, 1 if the 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