Unsupervised Machine Learning to Investigate Cardiovascular Disease

Introduction & Background:
Cardiovascular disease is one of the leading causes of death in the United States. Development of cardiovascular disease has been correlated with risk factors including obesity, high cholesterol, and hypertension [Tran]. Most of these risk factors are preventable with proper patient education.

Problem Definition:
Identifying specific clinical sub-populations of patients with cardiovascular disease can help researchers develop more personalized treatments and help doctors better predict specific types of patients that will be at risk for developing heart disease.

Methods:
Unsupervised machine learning techniques offer a solution for organizing patients into sub-populations based on their medical history. A previous study has proven clustering algorithms to be successful in identifying clinical sub-populations of Alzheimer's disease patients [Prakash et al]. A similar exploratory method will be applied to study patients with cardiovascular disease. T-Stochastic Neighbor Embedding and Principal Component Analysis will be applied to produce independent dimensionally-reduced versions of the original dataset. K-Means and DBSCAN clustering algorithms will then be separably applied to the original dataset and the reduced dimensionality versions of the dataset. The analysis which produces the most separable clusters will be selected for further statistical analysis to determine the characterizing features of each cluster.

Potential Results and Discussion:
The discovered clinical sub-populations will be presented along with their characterizing features. A review of medical literature will be conducted to contextualize these results with previous findings concerning cardiac disease.

Supervised Machine Learning to Detect and Classify Anti-Vaccine Tweets

Introduction/Background
The ongoing pandemic has yielded not only various vaccines, but also a plethora of anti-vaccination messages and posts all over the internet. The use of social media such as Facebook, Twitter, and Instagram to spread health information (both real and fake!) has contributed to this behaviour due to the prevalence of anti-vaccine content, especially on Twitter. A supervised machine learning technique will be used to classify health and vaccine related tweets as anti-vaccine.

Problem definition: Given a labelled dataset of tweets collected from Twitter during the ongoing pandemic, we wish to learn a machine learning model that can accurately classify a given tweet as either anti-vaccine or not. This can help researchers and scientists to accurately find tweets that may be spreading incorrect and false information about vaccines and would subsequently help in flagging such tweets without any human intervention.

Methods:
(1) **Data Preprocessing:** With around 6,000 tweets being tweeted on twitter every second, we have a lot of data to play around with! We plan on using the Avax Tweets Dataset which contains almost 2 million tweets (collected during the ongoing pandemic, 2020-2021) that are susceptible to the anti-vaccine narrative. There would be a decent amount of pre-processing involved here, such as changing all text to lowercase, omitting hashtags, usernames and the removal of stop-words (words frequently found in text like ‘is’, ‘are’). We would perform lemmatization in order to generate a canonical form of every word in the tweet. And finally, as in any supervised ML task, we would prepare the train/test split of our preprocessed dataset!

(2) **Training ML Models:** We propose to use two ML models for our project: The Support Vector Machine (SVM) and the Naive Bayes Classifier, both of which are classical ML methods that have been extensively used in the literature for text classification tasks. We plan on implementing both of these models for our dataset and use metrics such as F1 Score and Accuracy to evaluate their performance on unseen data. Further, if time permits, we also wish to use a recurrent neural network such as an LSTM and train it on our dataset. This would also require us to featurize the text, that is, create a neural-network-friendly vector representation of the text features.

Potential results and Discussion: We hope to efficiently implement our ML models which would be able to correctly identify anti-vaccination tweets with high accuracy and precision. Subsequently, we would like to compare the performance of our model with other existing ML and DL methods (cite here) and evaluate where our model stands.

At least three references (preferably peer reviewed). You need to properly cite the references on your proposal.

- 1
- 2
- 3

Week	Milestone we hope to accomplish	Primary Responsibilities
Week 7: Oct 1 to Oct 8	Brainstorm on ideas and write the Project Proposal	Everyone!
Week 8: Oct 9 to Oct 15	Dataset Preparation for both tasks	Sean and Abhinav
Week 9: Oct 16 to Oct 22	Analysing supervised and unsupervised methods and data preprocessing	Trey and Shashank
Week 10: Oct 23 to Oct 29	Implementing PCA (unsupervised) and vectorizing/lemmatization (supervised)	McKay and Sean
Week 11: Oct 30 to Nov 05	Implement the ML models - KMeans for unsupervised and Naive Bayes for supervised and obtain results	Abhinav and Trey
Week 12: Nov 06 to Nov 12	Implement the ML models - DBSCAN for unsupervised and SVM for supervised and obtain results	McKay and Shashank
Week 13: Nov 13 to Nov 19	Analyse all results using performance and evaluation metrics	Abhinav and Sean
Week 14: Nov 20 to Nov 26	Further testing and feedback. Attempting to improve model accuracies. Implementing the RNN for the supervised taks (if time permits).	Trey, McKay and Shashank
Week 15: Nov 27 to Dec 03	Work on the Final Presentation!	Everyone!

Add proposed timeline from start to finish and list each project members' responsibilities
Timeline:

- Week 7 (this week) - find primary and backup datasets
- Week 8 - process datasets
- Week 9 - choose supervised and unsupervised methods
- Week 10 - develop methods
- Week 11 - develop methods
- Week 12 - run methods
- Week 13 - analyze results
- Week 14 - draw conclusions
- Week 15 - Final presentation