


# Cardiovascular Disease Prediction



Bhavesh Shah, Griffin Coccari, Jamie Chen  
DSCI 303 Final Presentation



# Table of Contents

---



## Project Overview

- Why the problem matters
- Prior work
- Our contribution

- EDA

- Our data - features/observations
  - (#, distributions)

- Our Strategy

- Clustering

- K-Means
- GMM

- Classification

- Linear Regression
- KNN
- SVM
- Random Forest
- Neural Nets

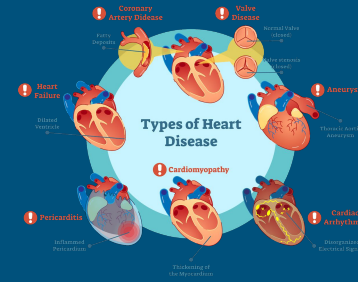
- Discussion - how to interpret results

- Algorithm performance
- Comparison with prior work

- Conclusion

- Quick Demo

# Why the Problem Matters



- Cardiovascular diseases (CVDs) are the leading cause of death globally
- ~18 million people died from CVDs in 2019 (32% of total deaths, global)
- However, according to WHO, most CVDs can be prevented by addressing behavioural risk factors such as tobacco use, unhealthy diet and obesity, physical inactivity, and excessive alcohol

# Our Goal

---

- To examine behavioral and medical factors widely thought to have some correlation with CVDs
- To create a tool to help doctors predict a patient's likelihood of getting CVDs to prescribe the necessary treatment/medication earlier on



# Prior Work

---

## Machine Learning–Driven Models to Predict Prognostic Outcomes in Patients Hospitalized With Heart Failure Using Electronic Health Records: Retrospective Study

- **Goal:** Predict 1-year in-hospital mortality, use of positive inotropic agents, and 1-year all-cause readmission rate
- **Method:** **Decision tree** of mortality risk (after consideration of logistic regression, support vector machine, artificial neural network, random forest, and extreme gradient boosting models)
- **Data:** real-world electronic health records

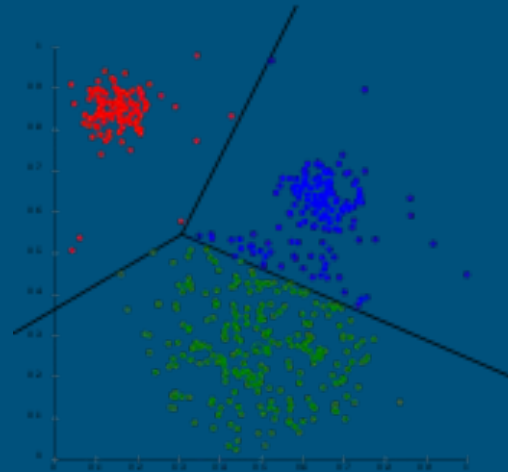
## Using Deep Learning to Identify High-Risk Patients with Heart Failure with Reduced Ejection Fraction | Published in Journal of Health Economics and Outcomes Research

- **Goal:** Predict hospitalizations, worsening HF events, and 30-day and 90-day readmissions in patients with heart failure with reduced ejection fraction (HFrEF)
- **Data:** Adult HFrEF patients from IBM® MarketScan® Commercial and Medicare Supplement databases (2015-2017)
- **Method:** **Sequential model architecture (DL)** based on bi-directional long short-term memory (Bi-LSTM) layers (also tested traditional ML models such as logistic regression, random forest, and eXtreme Gradient Boosting (XGBoost))
- **Results:** For all outcomes assessed, the DL approach outperformed traditional machine learning models

# Our Contribution

---

- How our work stands out?  
→ Our unique **clustering** → **classification strategy**
- Advantages/Motivation:
  - Meaningful clusters highly interpretable
  - More specific models to induce higher classification accuracies



# EDA



# Data Overview

## Dataset: Cardiovascular Disease dataset

- Clean - no missing values, no duplicate observations

Source: <https://www.kaggle.com/sulianova/cardiovascular-disease-dataset>

## Snapshot of our data:

1x ID      11x Features      1x Target

	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
0	0	18393	2	168	62.0	110	80	1	1	0	0	1	0
1	1	20228	1	156	85.0	140	90	3	1	0	0	1	1
2	2	18857	1	165	64.0	130	70	3	1	0	0	0	1
3	3	17623	2	169	82.0	150	100	1	1	0	0	1	1
4	4	17474	1	156	56.0	100	60	1	1	0	0	0	0

Number of observations: 70000

7000 observations



# Feature Overview

#		<u>Unique Values</u>	Dtype
---			----
0	id	70000	int64
1	age	8076	float64
2	gender	2	int64
3	height	109	int64
4	weight	287	float64
5	ap_hi	153	int64
6	ap_lo	157	int64
7	cholesterol	3	int64
8	gluc	3	int64
9	smoke	2	int64
10	alco	2	int64
11	active	2	int64
12	cardio	2	int64

## Objective features:

1. **Age** (days)
2. **Gender** (1=female, 2=male)
3. **Height** (cm)
4. **Weight** (kg)

## Examination features:

5. **Systolic blood pressure** (mmHg)
6. **Diastolic blood pressure/ ap\_lo** (mmHg)
7. **Cholesterol** (1=norm; 2=above norm; 3=well above norm)
8. **Glucose** (1=norm; 2=above norm; 3=well above norm)

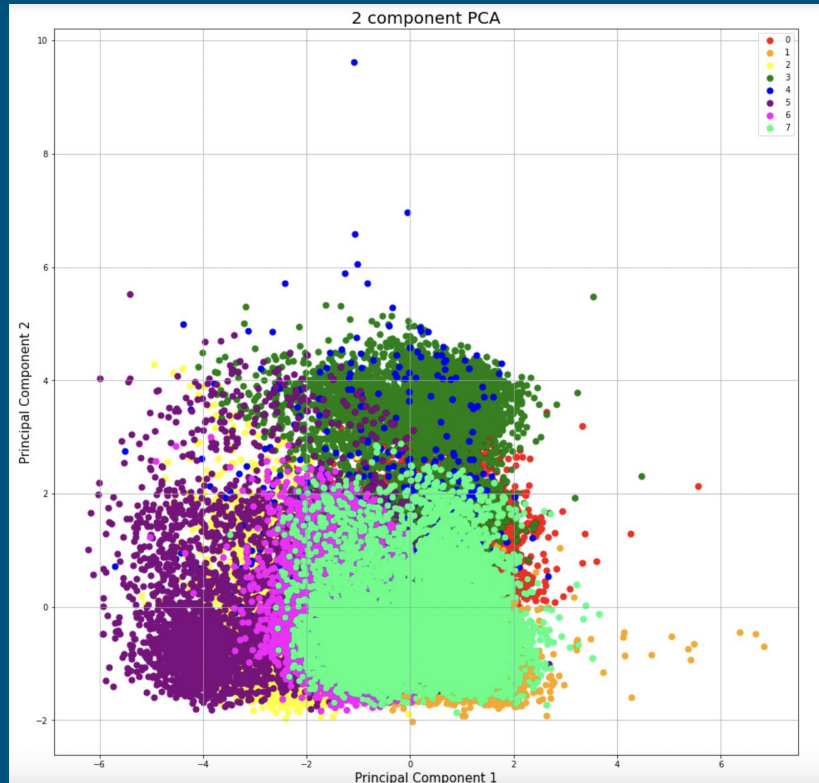
## Subjective features:

9. **Smoking** (0=does not smoke; 1=smokes)
10. **Alcohol Intake** (0=does not drink; 1=frequent drinker)
11. **Physical Activity** (0=not very active; 1=active)

# PCA

- Motivation: to reduce dimensionality of our data and visualize our clusters
- Dimensionality:
  - Total % data represented by principal components not very high
  - Plan to train more complex models (e.g.) neural nets requiring more features→ decide to max features and not reduce dimension
- Cluster Visualization:
  - Some overlap between clusters, but mostly separated → good sign for our clustering strategy on this dataset

```
Data Represented for Principal Component 1: 0.17645335485507652
Data Represented for Principal Component 1: 0.14341115133457946
Total Data Represented: 0.319864506189656
```



# Strategy

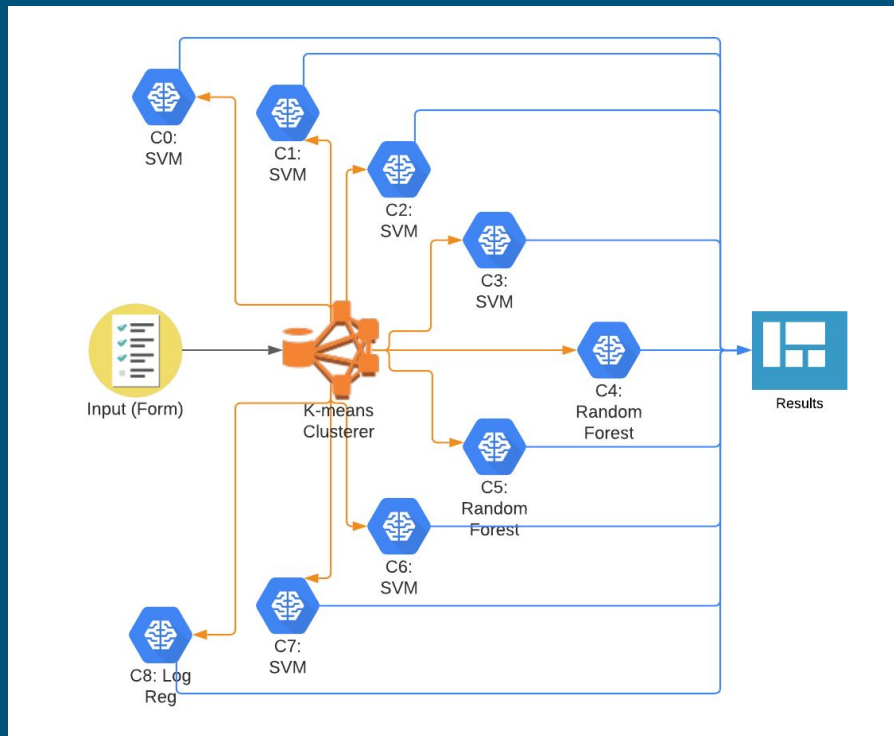
---

# Our Strategy

- Develop clusterer that sorts patients into 1 of 9 categories based on their health data characteristics.
- Test multiple classification models for each cluster to find best model per cluster.

## Benefits + Reasoning:

- More personalized models for patients instead of running general model across entire dataset
- Clusters could have **different feature importances** as well as different optimal models.
  - For ex., a cluster filled with older people might have cholesterol as the most important feature when predicting.



*Finalized Product Architecture*

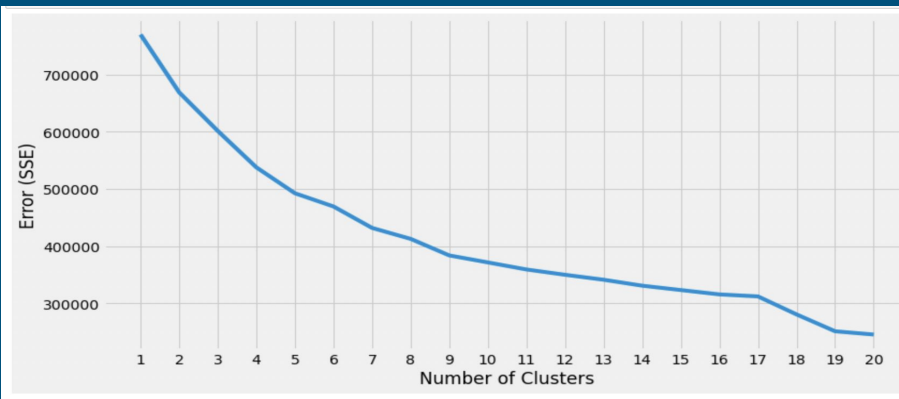
# Models

---

# Cluster Models

## K-means

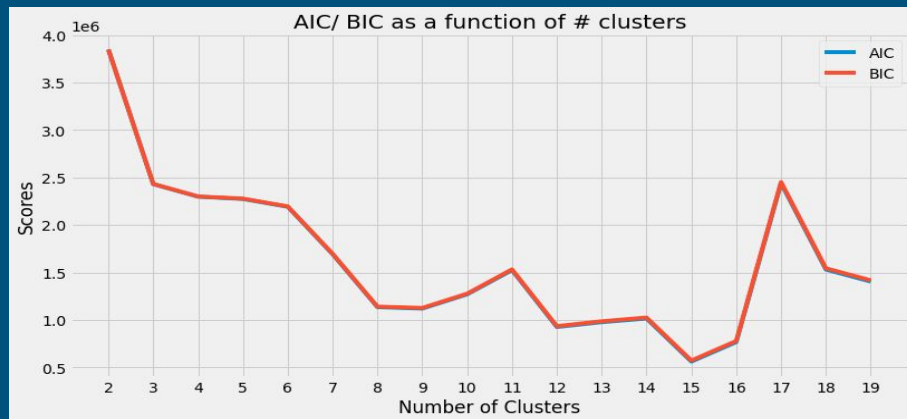
- Cluster patients into like-groups to develop more specific/personalized models
- Elbow method
  - → ideal number of clusters = 9



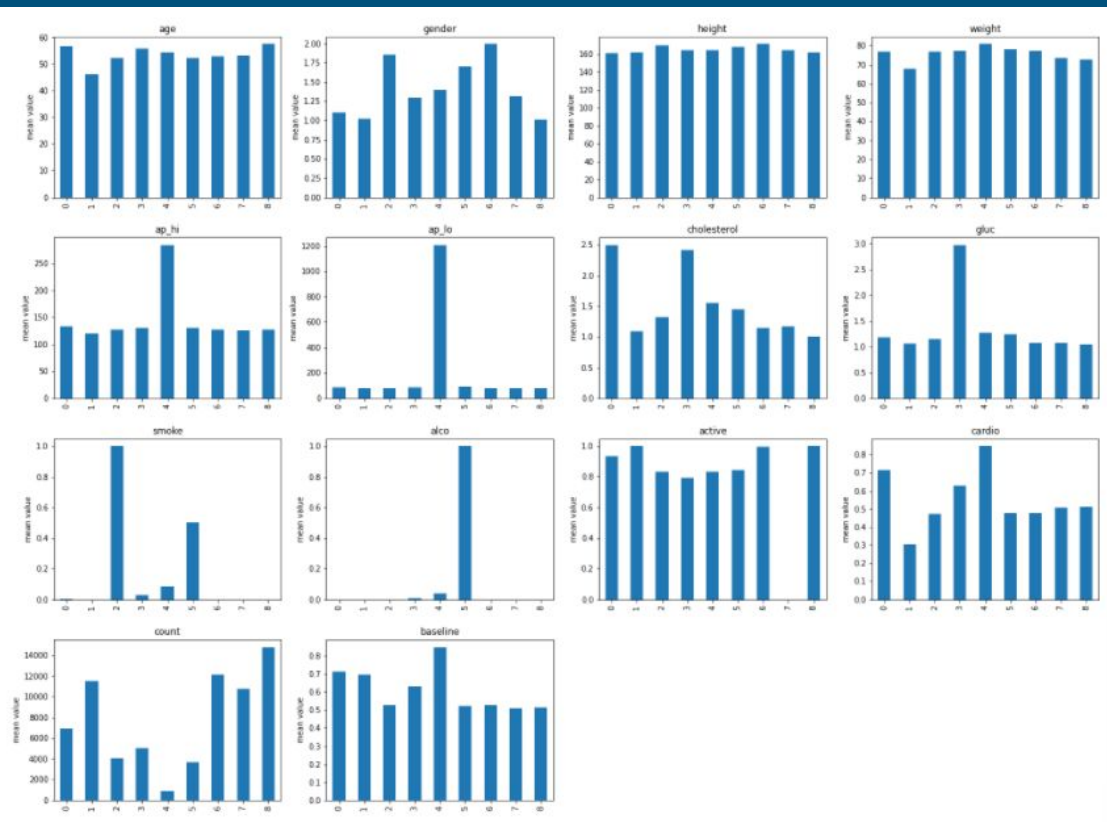
## GMM

- Confirm ideal # of clusters consistent with K-Means
  - → 15 (when minimizing AIC & BIC)
- However, 9 clusters is also top 5 minimizing AIC/BIC and resulted in better interpretable clusters

→ Thus, we decided to go with **9 clusters**



# Clusters Visualized



## ### Key Observations:

- **Cluster 0:** high cholesterol
- **Cluster 1:** lowest cvd%, youngest
- **Cluster 2:** pure smokers (no alcohol)
- **Cluster 3:** high glucose, high cholesterol
- **Cluster 4:** fewest members, highest cvd%, high blood pressure
- **Cluster 5:** alcoholics, some smokers
- **Cluster 6:** more male
- **Cluster 7:** not active at all
- **Cluster 8:** more female

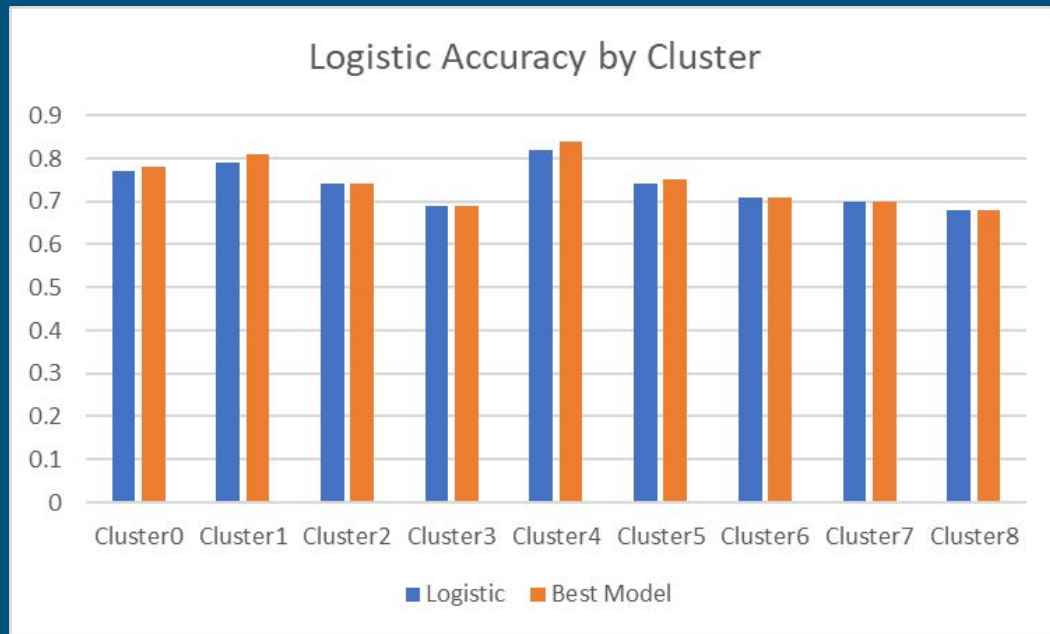
# Logistic Regression

## Justification

- Intuitive and interpretable
- Meaningful coefficients - provides direction of variables
- Good for relatively simple data sets

## Results + Interpretation

- Best model for cluster 8, tied for 2, 3, 6, 7 (with SVM, random forest)





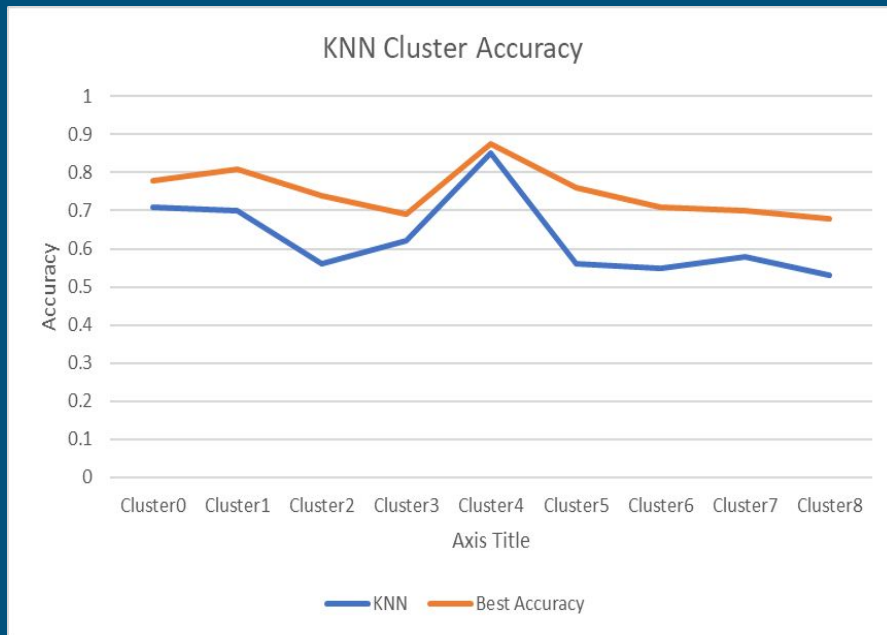
# KNN

## Justification

- Works well when the dataset is interpretable and clean (not missing values, outliers, etc).
- Dataset was not too large → no computational issues when training

## Results + Interpretation

- Worst model for every cluster (even with grid search hypertuning), but almost equal to the best accuracy model for cluster 4
  - Makes sense since the cluster 4 dataset is the smallest and most imbalanced



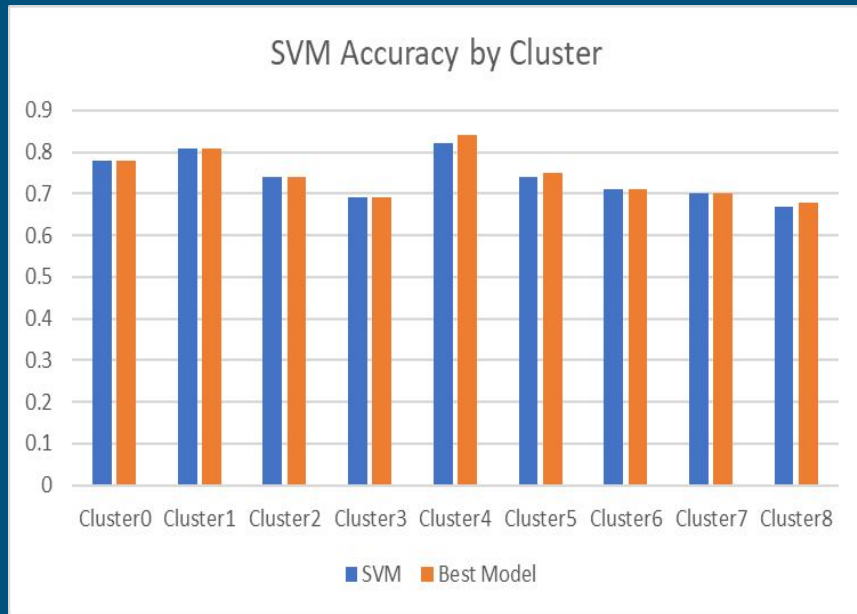
# SVM (with RBF kernel)

## Justification

- Commonly used in previous research
- Our data set is not too large
- Ample observations per class (in comparison to number of features)

## Results + Interpretation

- Scored as well or better than logistic regression and Random Forest
- Best Model for cluster 1
- Underperformed on cluster 4 (most imbalanced cluster)



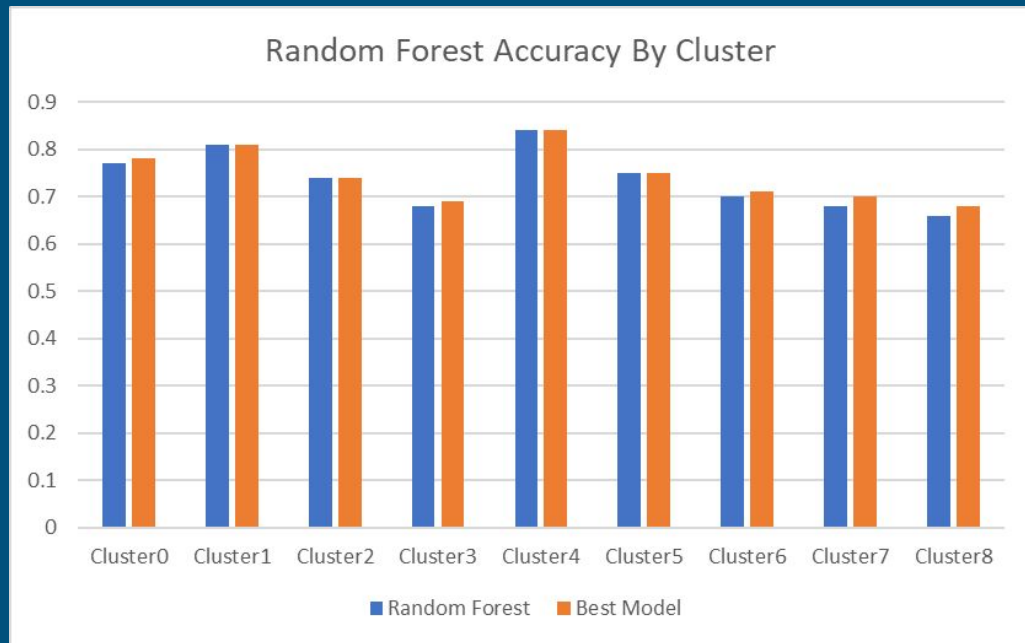
# Random Forests

## Justification

- Increasing # of trees improves overall accuracy while decreasing variance

## Results + Interpretation

- Performed as well as SVM on clusters 1 and 2
- Outperformed SVM on clusters 4 and 5
- Performed well on balanced and imbalanced clusters



# Random Forest Cont.

Systolic blood pressure is most important feature for Cluster 5 (characterized by alcoholics and some smokers)

Importance	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8
1st	age : 0.3652944180 730339	age : 0.3077877594 838806	age : 0.2344266841 4674425	age : 0.2952332161 8940124	age : 0.2339381075 7535713	ap_hi : 0.224279955 4220934	age : 0.3326115594 782371	age : 0.3340051390 9833696	age : 0.3652944180 730339
2nd	weight : 0.2094259356 6961135	ap_hi : 0.2264903185 034671	ap_hi : 0.2267028859 4292742	weight : 0.2120213669 6622234	weight : 0.1825400377 8674688	age : 0.2121827807 171412	weight : 0.1893215159 6445285	weight : 0.1979675290 9733503	weight : 0.2094259356 6961135
3rd	height : 0.1762395083 9832322	weight : 0.1752511893 327448	weight : 0.1748720023 410046	height : 0.1832586321 7637725	height : 0.1724207757 9081921	weight : 0.1633329547 9425852	ap_hi : 0.1883660671 8965473	height : 0.1730334619 153418	height : 0.1762395083 9832322
4th	ap_hi : 0.1546546818 1448655	height : 0.1537198659 8259424	height : 0.1464486758 8085506	ap_hi : 0.1194826089 2487463	ap_hi : 0.1601812870 497378	height : 0.1334893229 448047	height : 0.1603499799 4176603	ap_hi : 0.1584944167 3169606	ap_hi : 0.1546546818 1448655
5th	ap_lo : 0.0851605487 150848	ap_lo : 0.1132327502 0295946	ap_lo : 0.1129556547 728865	ap_lo : 0.0708020660 2918615	ap_lo : 0.0974790849 8974903	ap_lo : 0.1214229483 7531606	ap_lo : 0.0971801836 3027738	ap_lo : 0.0899279571 340723	ap_lo : 0.0851605487 150848
6th	gluc : 0.0064097030 55124658	cholesterol : 0.0111889927 54255068	cholesterol : 0.0512970068 29260514	cholesterol : 0.0533604390 7537191	cholesterol : 0.0399258293 3809939	cholesterol : 0.0677372371 3293915	cholesterol : 0.0193733567 3427911	cholesterol : 0.0191807727 5307003	gluc : 0.0064097030 55124658
7th	gender : 0.0022919035 40280786	gluc : 0.0088284257 13093677	gluc : 0.0201374369 5690987	gender : 0.0257207147 98103055	gluc : 0.0354850803 6272542	gluc : 0.0219473680 4478814	gluc : 0.0092800380 05585124	gender : 0.0163468610 6209627	gender : 0.0022919035 40280786
8th	cholesterol : 0.0002717602 71640507	gender : 0.0031899758 81355818	active : 0.0183644383 87448088	active : 0.0247337454 16409228	gender : 0.0252243975 6262045	smoke : 0.0206892475 18152713	gender : 0.0021288575 2737346	gluc : 0.0110438622 08051548	cholesterol : 0.0002717602 71640507
9th	active : 0.0002515404 624143157	active : 0.0003107221 4564910626	gender : 0.0147952147 41963752	smoke : 0.0072865115 268394435	smoke : 0.0217421400 07980384	gender : 0.0179666547 32173443	active : 0.0013884415 28374211	smoke : 0.0	active : 0.0002515404 624143157
10th	smoke : 0.0	smoke : 0.0	smoke : 0.0	gluc : 0.0053693117 17226398	active : 0.0207825166 11383892	active : 0.0169515320 921675	smoke : 0.0	alco : 0.0	smoke : 0.0
11th	alco : 0.0	alco : 0.0	alco : 0.0	alco : 0.0027313871 79988218	alco : 0.0102807420 24780576	alco : 0.0	alco : 0.0	active : 0.0	alco : 0.0

\*\*\* Different clusters have different feature importances \*\*\*

# Neural Nets

## Justification

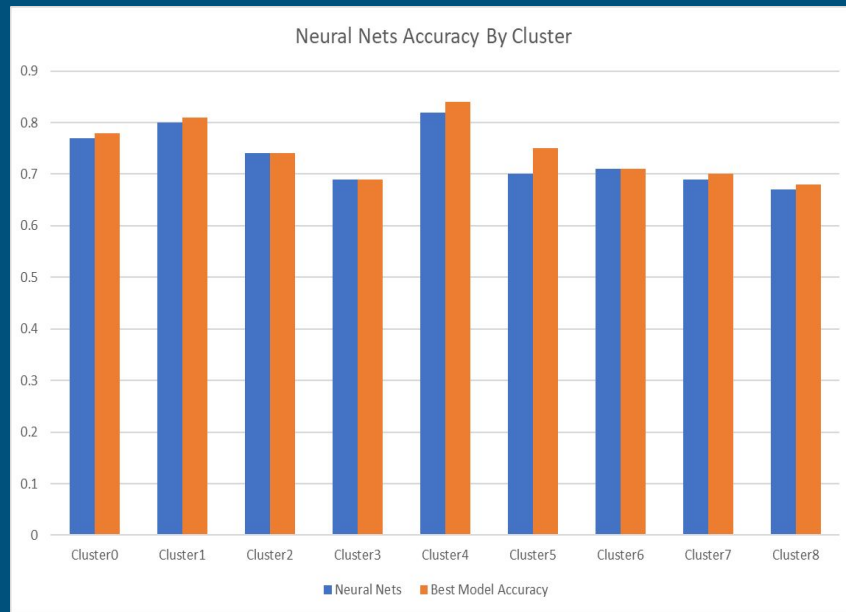
- Non-linear and complex relationships

## Results + Interpretation

- Tied best model accuracies for clusters 2, 3, 6
- For other clusters, accuracy slightly behind best (SVM/random forest)

Note: We are still in the middle of optimizing the NN architecture and are hoping for better accuracies.

*Current architecture includes 2 hidden layers and 64 neurons per hidden layer. Other parameters such as # epochs and batch size vary per cluster).*



# [Discussion] Model Performances (Comparison)

- Other research used data sets with more features (72) and their models had high degrees of separability (AUC). They find that ML outcomes can be accurately predicted with “a large scale of clinical variables”(Haichen)
- Clustering Approach scored slightly better than similar Kaggle projects for the same dataset (low 70%’s):
  - <https://www.kaggle.com/turhancankargin/a-simple-classification-application>
  - <https://www.kaggle.com/abdallahmahmoud/cardiovascular-disease-prediction-73-59-accuracy>

Optimal model per cluster

	Model
Cluster0	SVM
Cluster1	SVM/Random Forest
Cluster2	Logistic/SVM/Random Forest/NN
<b>Cluster3</b>	Logistic/SVM/NN
Cluster4	Random Forest
Cluster5	Random Forest
Cluster6	Logistic/SVM/NN
Cluster7	Logistic/SVM
<b>Cluster8</b>	Logistic

# Demo



# Conclusion

---

- While our ML application may not be production ready, we believe that our approach is still valuable.
- → We were able to determine that clustering does perform better than non-clustered models even if it's only a slight improvement.
  - There is potential to increase the accuracy of these personalized models even more (for ex. changing the NN architecture).
- Eventually, we hope to eventually provide a way to alert patients of their risk of cardiovascular disease with our deployed model



# Works Cited

---

Nou, Alexander. "Logistic Regression Versus Support Vector Machines." *Stockholm Universitet*, Stockholm University, June 2018.

Lv, Haichen. "Driven Models to Predict Prognostic Outcomes in Patients Hospitalized With Heart Failure Using Electronic Health Records: Retrospective Study." *NIH*, April 19, 2021, <https://pubmed.ncbi.nlm.nih.gov/33871375/>.

Van Uden, Cara. "Cardiovascular Disease Diagnosis." *Github*, <https://github.com/caravanuden/cardio?fbclid=IwAR0yx2BEyOXov2qs1x1V20PBpCxiezUcsnNvUGI94XSIBAgTMotiWRWmUGY>.

Mahanta, Jahnavi. "Introduction to Neural Networks, Advantages and Applications." *Towards Data Science*, July 10, 2017, <https://towardsdatascience.com/introduction-to-neural-networks-advantages-and-applications-96851bd1a207>.