



# Predicting Heart Disease

---

STEPHEN LANIER

## PROBLEM STATEMENT

---

- 13 predictors
- Recall vs precision
- Combining models

Features: age, sex, angina, resting blood pressure, serum cholesterol, fasting blood sugar, resting EKG results, max heart rate, exercise-induced angina, ST depression, slope of ST segment, fluoroscopy, thallium heart scan.

Medical situation: favor recall over precision, but for now, since no class imbalance, just look at accuracy.

## BUSINESS VALUE

"Heart disease is the **leading cause of death** for men, women, and people of most racial and ethnic groups in the United States."

"Heart disease costs the United States about **\$219 billion** each year from 2014 to 2015. This includes the cost of health care services, medicines, and lost productivity due to death."

Citation: [CDC](#)

<https://www.cdc.gov/heartdisease/facts.htm#:~:text=Heart%20disease%20is%20the%20leading,1%20in%20every%204%20deaths.>

## METHODOLOGY

01

Refine  
data

02

Tune  
models

03

Combine  
strongest  
models

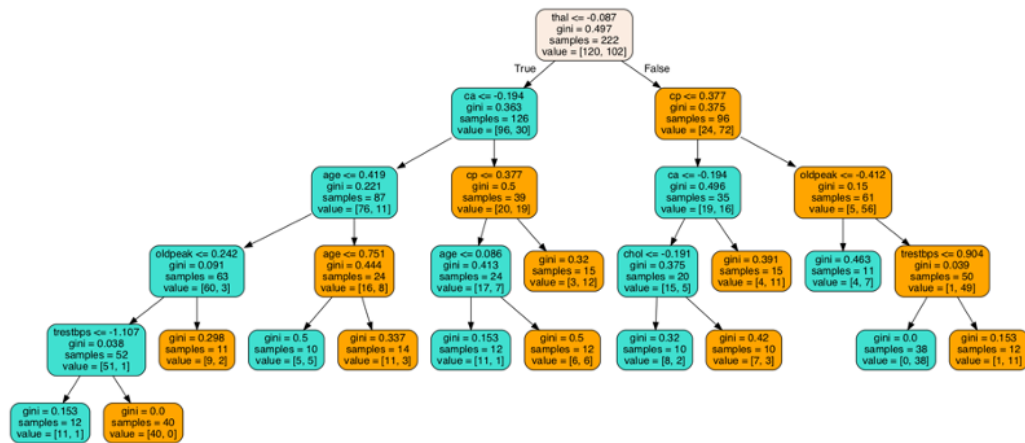
## FINDINGS

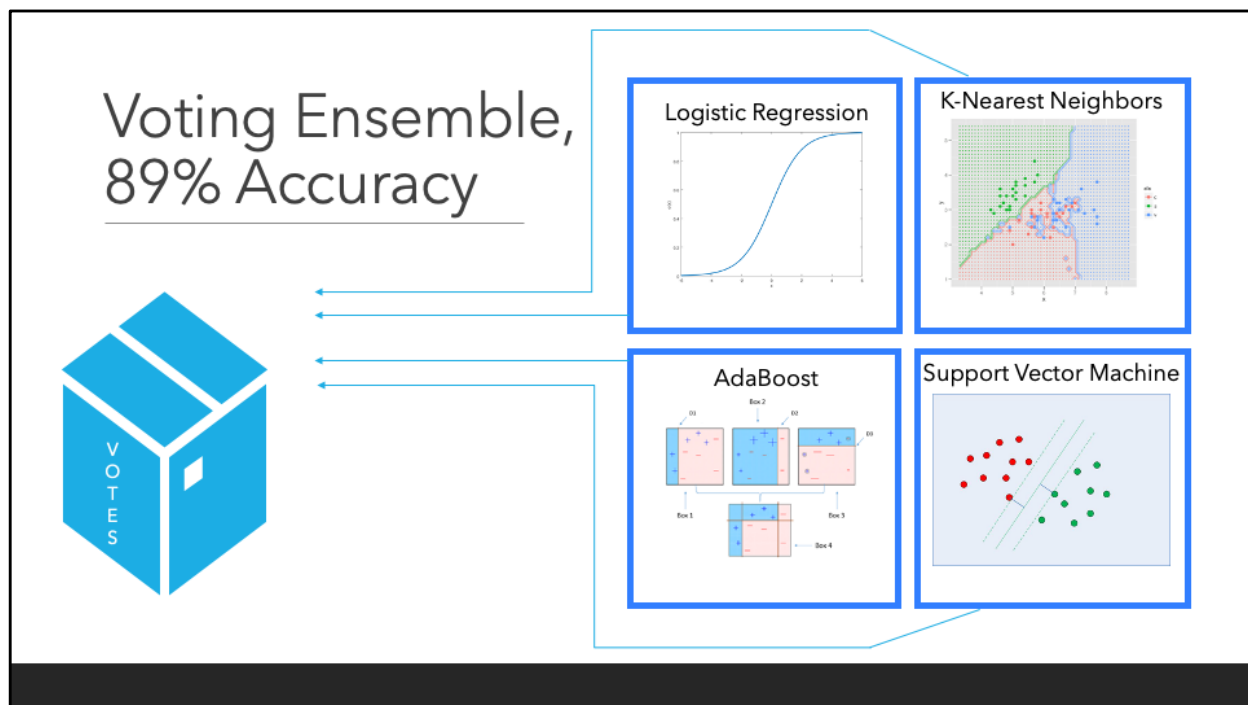
---

1. **7** important predictors
2. Best single models: **88%** accuracy
3. Combined methods: **89%** accuracy

angina, thallium heart scan, fluoroscopy, ST depression, age, cholesterol, resting blood pressure

# Decision Tree, 87% Accuracy





1. Models trained independently
2. Each model votes
3. Votes tallied, prediction made

## CONCLUSIONS

---

Able to predict heart disease from 7 measurements with ~90% accuracy.



## Future Work

---

- Change metric from accuracy to recall for healthcare setting
- Develop more sophisticated models: clustering, perceptrons, neural networks

## ACKNOWLEDGEMENTS

---

Kaggle for data associated with [Heart Disease Ensemble Classifiers](#)

Thank you!



## Images Used

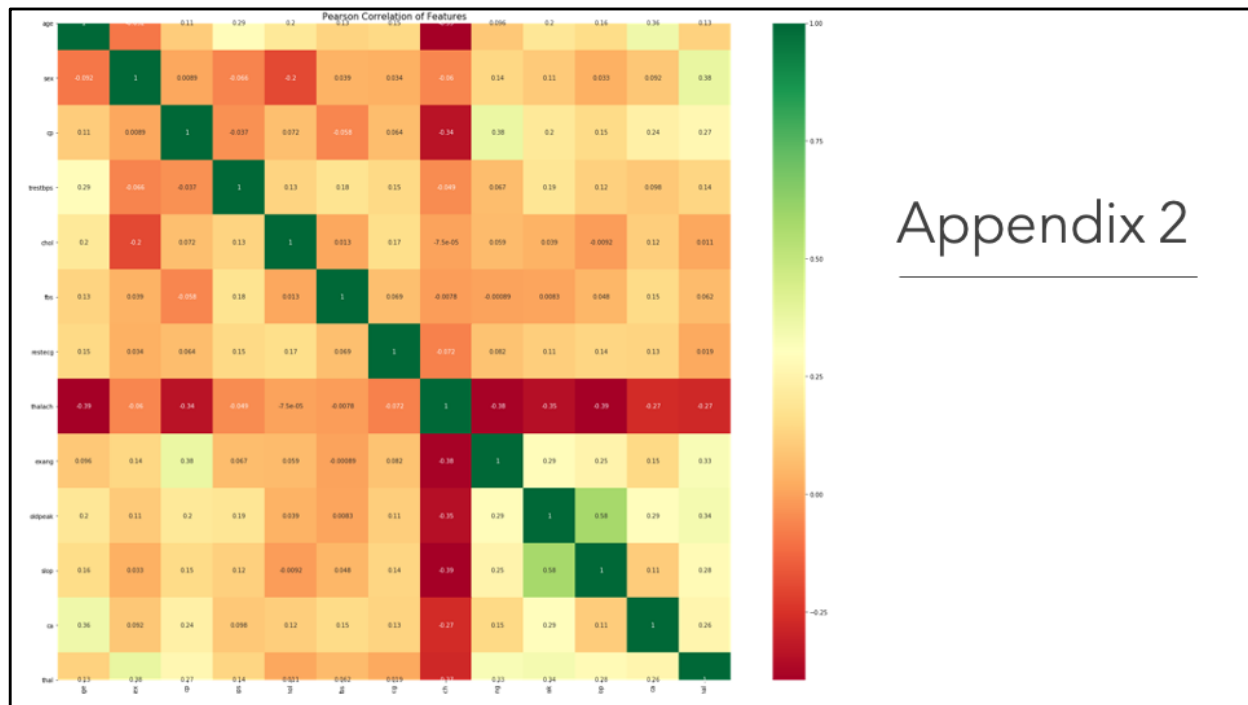
---

- logistic regression: [https://hvidberrrg.github.io/deep\\_learning/activation\\_functions/sigmoid\\_function\\_and\\_derivative.html](https://hvidberrrg.github.io/deep_learning/activation_functions/sigmoid_function_and_derivative.html)
- KNN: [http://dylanwiwad.com/project/predicting\\_car\\_prices/](http://dylanwiwad.com/project/predicting_car_prices/)
- SVM: <https://stackabuse.com/implementing-svm-and-kernel-svm-with-pythons-scikit-learn/>
- AdaBoost: <https://towardsdatascience.com/understanding-adaboost-2f94f22d5bfe>

## Appendix 1

thalach	0.128969
cp	0.122558
thal	0.112145
ca	0.106617
oldpeak	0.104328
age	0.091347
chol	0.083361
trestbps	0.080489
exang	0.064376
slop	0.046107
sex	0.027416
restecg	0.024662
fbs	0.007625

Feature importance from Random Forest. Max heart rate (highly correlate, removed), angina, thallium heart scan, fluoroscopy, ST depression, age, cholesterol, resting blood pressure



## Appendix 2

Heatmap of Pearson correlation coefficient. 'thalach' offends more than any other, was removed. So was 'slop'.

Model	Initial Test Accuracy	Final Test Accuracy
Ensemble Classifier (Soft, Unprocessed)	--	89%
Sigmoid SVC	55%	88%
AdaBoost	72%	87%
Decision Tree	77%	87%
Ensemble Classifier (Soft)	--	87%
Ensemble Classifier (Hard)	--	87%
Logistic Regression	87%	85%
Random Forest	83%	85%
XGBoost	81%	84%
KNN	59%	82%

## Appendix 3

---