

Final Project Requirements for Graduate Students

ICS 635

Spring 2023

UHM Computer Science Career Fair

Final event of Hawai'i Tech Days of Spring

Are you a student in the field of computer science? Interested in opportunities in software engineering, data science, and more?

Attend this career fair for a great opportunity to network, learn about job or internship openings, and make connections with industry professionals!



ACM at Manoa
Hosted



ICS Department
at UHM
Sponsored



Register here! Or go to:
bit.ly/icscareerfair

Friday, March 31st
2:00pm – 4:00pm

3rd Floor of Pacific Ocean Science &
Technology (POST) Building

TALK STORY WITH TECH PROFESSIONALS

Meet over 30 software engineers, company founders, product managers, and cybersecurity experts from local companies like First Hawaiian Bank, to international companies like Google.



TUESDAY
MAR 21

5:00 PM - 6:30 PM

📍 University Lab School
Multipurpose Room

- Open to all UH students
- 15-minute welcome reception followed by a 1-hour 15-minute mixer
- Doors open at 4:30 PM
- Light refreshments will be served

While supplies last

Brought to you by



In collaboration with:



builders | vc



Sign up at:

pace.shidler.hawaii.edu/tech

HW2 Top-Performing Models

1. MLP Regressor (a type of simple neural network) with 2 hidden layers
2. Gradient Boosting (a type of tree boosting) (tie)
2. Voting Regressor using Linear Regression, Random Forest, and Gradient Boosting (tie)
- 3 and 4. Extra Trees Regression and Support Vector Regression

HW2 Top-Performing Models

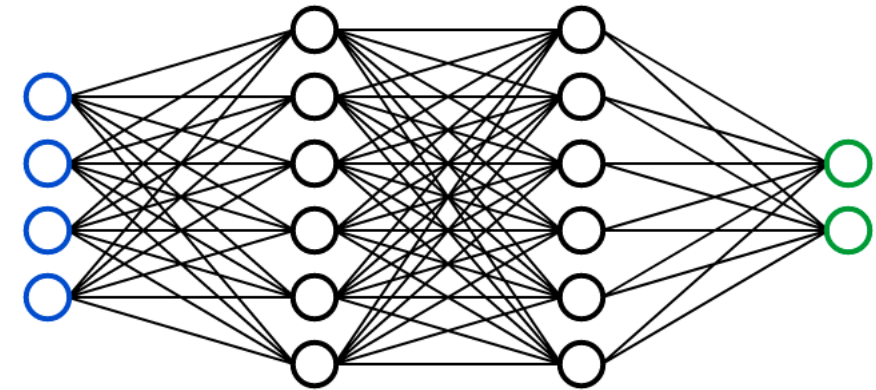
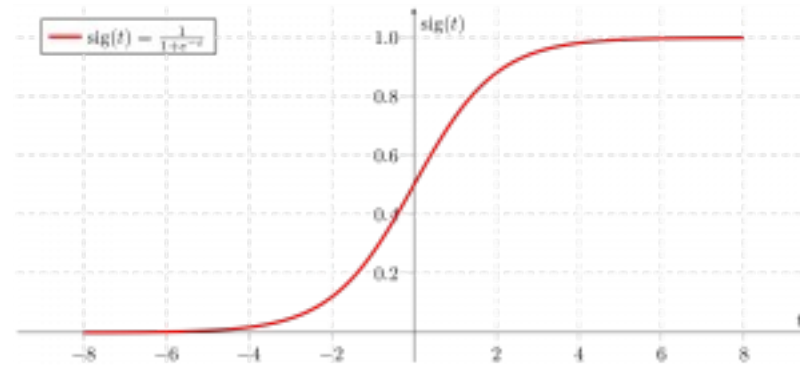
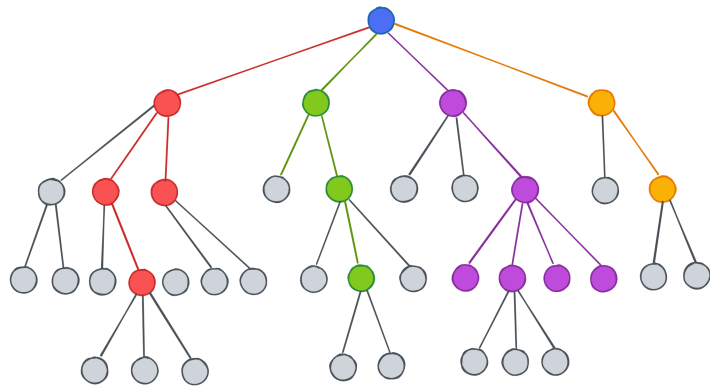
1. MLP Regressor (a type of simple neural network) with 2 hidden layers
2. Gradient Boosting (a type of tree boosting) (tie)
2. Voting Regressor using Linear Regression, Random Forest, and Gradient Boosting (tie)
- 3 and 4. Extra Trees Regression and Support Vector Regression

Note: Some students implemented larger neural networks with many layers. These did very well on the test set that we provided but severely overfitted, failing to generalize to the held out test set.

HW3 Questions?

I am travelling during Spring Break and will be unable to answer questions during email during this time.

Final Project Type 1: Build a series of ML models with a dataset of your choosing using **at least 3 of the techniques we learned about in class.**

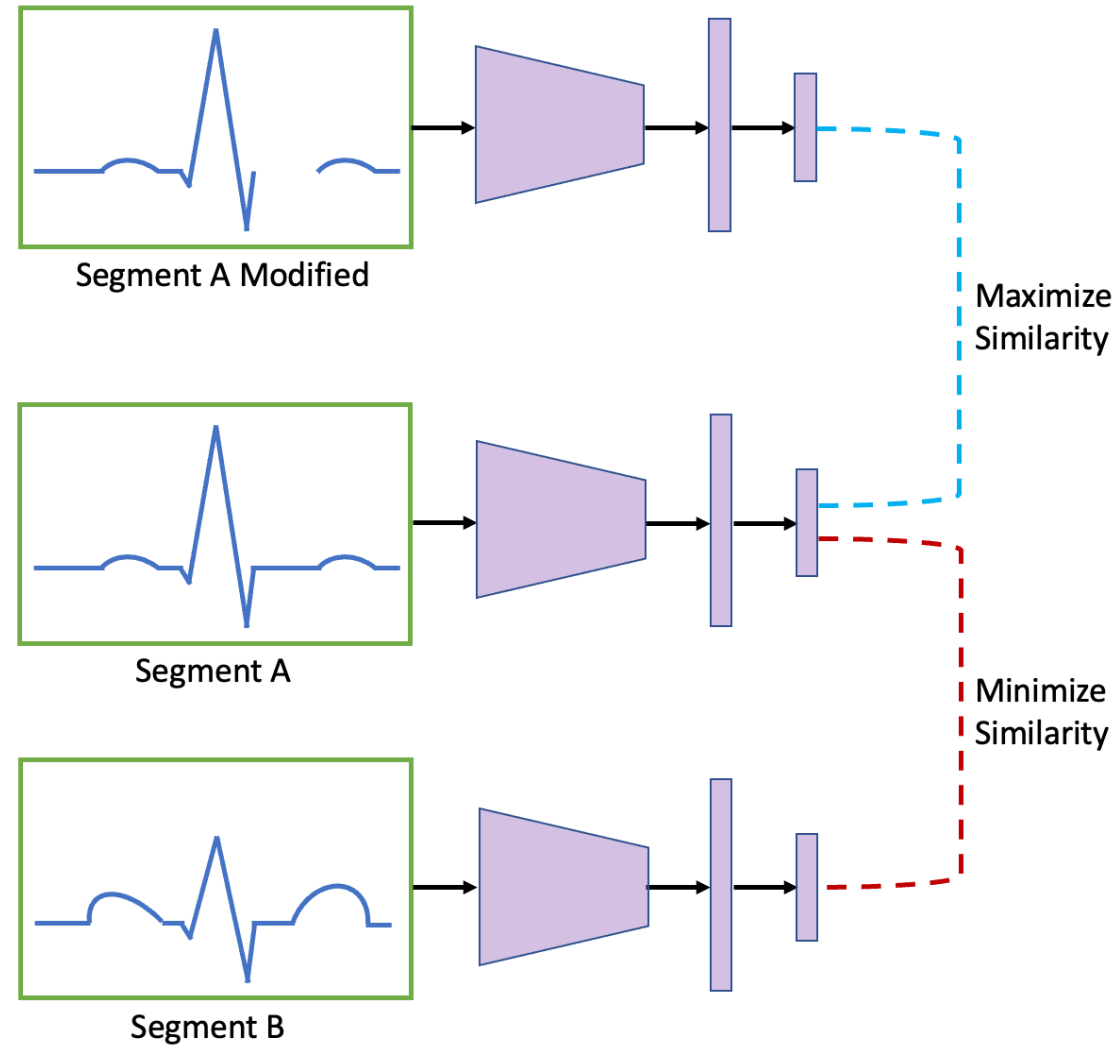


Final Project Type 1: Apply Existing Methods to a New Dataset

Find an interesting, non-mainstream dataset, and tell a coherent story using at least 3 of the methods that we covered in class.

Date	Topic
Tues Jan 10	Overview of Machine Learning (ML) (Notes)
Thur Jan 12	Linear Regression and Intro to Loss Functions (Notes)
Tues Jan 17	Logistic Regression and ML Evaluation Part 1 (Notes)
Thur Jan 19	ML Evaluation Part 2 and Calculus Review (Notes)
Tues Jan 24	Real-World Coding: Python Coding and Libraries (Video + Code on Laulima)
Thur Jan 26	Real-World Coding: Car Price Prediction (Video + Code on Laulima)
Tues Jan 31	Real-World Coding: Breast Cancer and Sentiment Prediction (Video + Code on Laulima)
Thur Feb 02	Gradient Descent (Notes)
Tues Feb 07	Regularization and Probability Review (Notes)
Thur Feb 09	Maximum Likelihood, KNN, and Naive Bayes Intro (Notes)
Tues Feb 14	Naive Bayes and Decision Trees (Notes)
Thur Feb 16	Ensemble Learning (Notes)
Tues Feb 21	Linear Algebra Review and SVMs (Notes)
Thur Feb 23	Kernels and Clustering (Notes)
Tues Feb 28	Midterm Review (Notes)
Thur Mar 02	Feature Selection and Engineering (Notes)
Tues Mar 07	Final project overview and class activity: ICS/DATA 435 students in class only
Thur Mar 09	Final project overview and class activity: ICS 635 students in class only
Tues Mar 21	
Thur Mar 23	
Tues Mar 28	Deep Learning
Thur Mar 30	
Tues Apr 04	Reinforcement Learning
Thur Apr 06	
Tues Apr 11	Practical Topics
Thur Apr 13	
Tues Apr 18	
Thur Apr 20	

Final Project Type 2: Develop a new machine learning method on any dataset.



Final Project Type 2: Implement a New Machine Learning Method

If you use a dataset which can be imported from sklearn, TensorFlow, or PyTorch, then you must create a new methodology or evaluate an existing methodology which is not implemented by default in sklearn.

(This is the harder option for final projects)

Some possible new innovations:

- Feature Selection method not in sklearn
- Feature Engineering method not in sklearn
- New Loss Functions or Regularization Terms
- Modification of ML algorithms we learned in class
- Active Learning
- Fair Machine Learning
- Explainable Machine Learning

...

You have lots of freedom for your project

- You can choose any topic you want
- You can use ML methods outside of those covered in class if you want as well
- Can always check with Dr. Washington

You may **NOT** copy code from public tutorials. There are a very limited number of datasets which tend to be used for ML tutorials, so we will know if you copy these tutorials.

 [PyCodeMates](https://www.pycodemates.com)
<https://www.pycodemates.com> > ... > Data Science


[Iris Dataset Classification with Python: A Tutorial - PyCodeMates](#)

This **tutorial** will use Python to classify the **Iris dataset** into one of three flower species: Setosa, Versicolor, or Virginica.

 [Kaggle](https://www.kaggle.com)
<https://www.kaggle.com> > code > ash316 > ml-from-s...

[ML from Scratch with IRIS!! - Kaggle](#)

This is a very basic **tutorial** to Machine Learning for complete Beginners using the **Iris Dataset**. You can learn how to implement a machine learning to a ...

 [Towards Data Science](https://towardsdatascience.com)
<https://towardsdatascience.com> > exploring-classifiers-...

[Exploring Classifiers with Python Scikit-learn – Iris Dataset](#)


Jul 13, 2020 – Python Scikit-learn is a great library to build your first classifier. The task is to classify **iris** species and find the most influential ...

 [thatascience](https://thatascience.com)
<https://thatascience.com> > Machine Learning


[Iris Dataset - A Detailed Tutorial - thatascience](#)

Learn to load **iris dataset** from sklearn datasets with an easy **tutorial**. **Iris dataset** is a part of sklearn library to practice machine learning techniques.


Videos


 [Iris Dataset Analysis \(Classification\) | Machine Learning | Python](#)
YouTube · Hackers Realm
Jun 16, 2020

 11 key moments in this video

 [Python Data Analysis with Iris Dataset | Data Science, plotting ...](#)
YouTube · Joe James
Nov 1, 2021

10 key moments in this video

 [Getting started in scikit-learn with the famous iris dataset](#)
YouTube · Data School
Apr 21, 2015

 10 key moments in this video

[Feedback](#)

[View all](#) →

 [GeeksforGeeks](https://www.geeksforgeeks.org)
<https://www.geeksforgeeks.org> > python-basics-of-pan...


[Python - Basics of Pandas using Iris Dataset - GeeksforGeeks](#)

Jan 10, 2023 – **Iris dataset** is the Hello World for the Data Science, so if you have started your career in Data Science and Machine Learning you will be ...

 [PlainEnglish.io](https://plainenglish.io)
<https://plainenglish.io> > blog > iris-flower-classificatio...

[Iris Flower Classification Step-by-Step Tutorial - PlainEnglish.io](#)

Jan 21, 2021 – Do you want to learn machine learning, but having trouble getting started? · Setting up the Environment. · Loading the **dataset**. · Summarizing the ...

 [scikit-learn](http://scikit-learn.org)
<http://scikit-learn.org> > datasets > plot_iris_dataset

[The Iris Dataset – scikit-learn 1.2.1 documentation](#)

This data sets consists of 3 different types of irises' (Setosa, Versicolour, and Virginica) petal and sepal length, stored in a 150x4 numpy.ndarray The ...

 [Ritchie Ng](http://www.ritchieng.com)
<http://www.ritchieng.com> > machine-learning-iris-data...

[Iris Dataset | Machine Learning, Deep Learning ... - Ritchie Ng](#)

Jan 21, 2023 – Topics¶. About **Iris dataset** · This **tutorial** is derived from Data School's Machine Learning with scikit-learn **tutorial**. I added my own notes so ...

Requirements: Coding

- Project code is provided as both a .ipynb file and a PDF version of the .ipynb file: 1 point
- Project code meets all requirements described above: 1 point
- Code cleanliness, organization, documentation, and comments: 1 point

Requirements: Writing

- At least 1.5 pages single spaced 12pt font (**not including** figures and references): 1 point
- Introduction, Related Work, Methods, Results, Discussion, and References (at least 10 references) sections included: 1 point
- At least 1 Methods figure and 1 Results figure: 1 point
- Professionalism, grammar, and style in the format of a technical ML paper: 1 point
- *The overall writing should be in the style of a research paper.*

Introduction

A few sentences for each of:

- Stating the problem from a societal level
- Stating the problem/gaps from a technical level
- Summarizing what's already been done
- Summarize what is done in this paper and the findings

Related Work

- Your References must be academic (e.g., research papers)
- Don't simply summarize each paper. Rather, synthesize the work and discuss high level themes.

Related Work: Example

The overarching challenge in active learning research is that no single metric-based strategy, whether based on uncertainty or diversity, will work for all types of data. For example, datasets may consist of several repetitive data points, resulting in redundancy when using a static metric such as maximum entropy. Adaptive strategies modeled after reinforcement learning can learn a policy for selecting salient data points, thereby “learning how to learn” [64]. In some cases tailored to CV for human behavior, the active learning system relies on a policy function which is personalized for each human subject [165]. Reinforcement learning for active learning is a relatively understudied area, but initial approaches rely on the formulation of active learning as a Markov Decision Process (MDP). Some approaches model active learning as a streaming process, where individual frames are sent to the system, which then decides whether to label the data point [64, 165]. Other formulations select an individual point from a large pool [120, 200]. The reward of the MDP is usually a function of the increase in classifier performance according to one or more metrics when new data are added [64, 120, 152, 165]. We note that this reward measurement in each state of the MDP is nontrivial, as it requires heavy computation per iteration of active learning to retrain the classifier and measure its performance. The MDP state sometimes does not account for the discrete time inherent of a MDP and is modeled as the point or points being considered or some function (feature representation) of the point [165, 217]. Because uncertainty-based active learning requires a metric for classifier uncertainty, the state can also include the classifier’s parameters or even the prediction itself [152], which is a function of the data point. The MDP action is usually the choice of one or more unlabeled points to label [64, 165].

Related Work: Example

Note:
Multiple
sources after
a statement
counts
towards only
1 of the 10
references for
the purposes
of your final
project.

The overarching challenge in active learning research is that no single metric-based strategy, whether based on uncertainty or diversity, will work for all types of data. For example, datasets may consist of several repetitive data points, resulting in redundancy when using a static metric such as maximum entropy. Adaptive strategies modeled after reinforcement learning can learn a policy for selecting salient data points, thereby “learning how to learn” [64]. In some cases tailored to CV for human behavior, the active learning system relies on a policy function which is personalized for each human subject [165]. Reinforcement learning for active learning is a relatively understudied area, but initial approaches rely on the formulation of active learning as a Markov Decision Process (MDP). Some approaches model active learning as a streaming process, where individual frames are sent to the system, which then decides whether to label the data point [64, 165]. Other formulations select an individual point from a large pool [120, 200]. The reward of the MDP is usually a function of the increase in classifier performance according to one or more metrics when new data are added [64, 120, 152, 165]. We note that this reward measurement in each state of the MDP is nontrivial, as it requires heavy computation per iteration of active learning to retrain the classifier and measure its performance. The MDP state sometimes does not account for the discrete time inherent of a MDP and is modeled as the point or points being considered or some function (feature representation) of the point [165, 217]. Because uncertainty-based active learning requires a metric for classifier uncertainty, the state can also include the classifier’s parameters or even the prediction itself [152], which is a function of the data point. The MDP action is usually the choice of one or more unlabeled points to label [64, 165].

Methods

- Dataset description
- Data preprocessing steps
- Machine learning models and methods

Results

- Evaluate your model appropriately for your domain
- Include quantitative comparisons
- Use the evaluation metrics from class or other appropriate metrics

Discussion

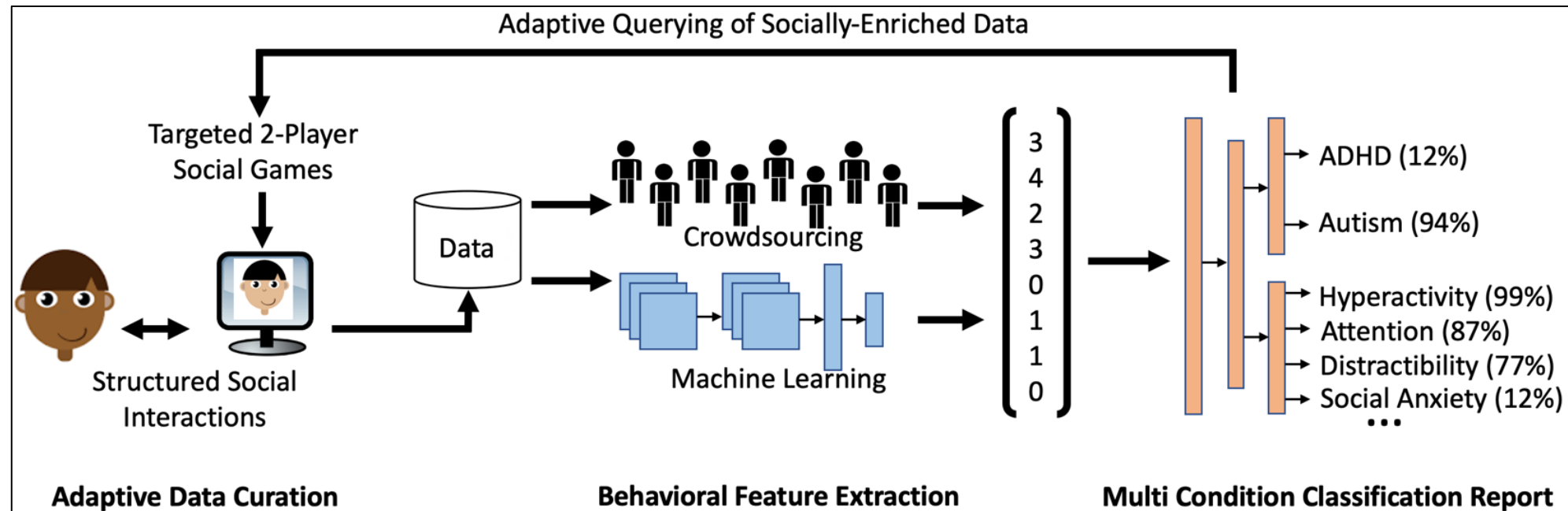
- Summary of your results
- High level implications of your results
- Limitations of your approach
- Opportunities for future work

References

- Can be in any citation format
- Just be consistent

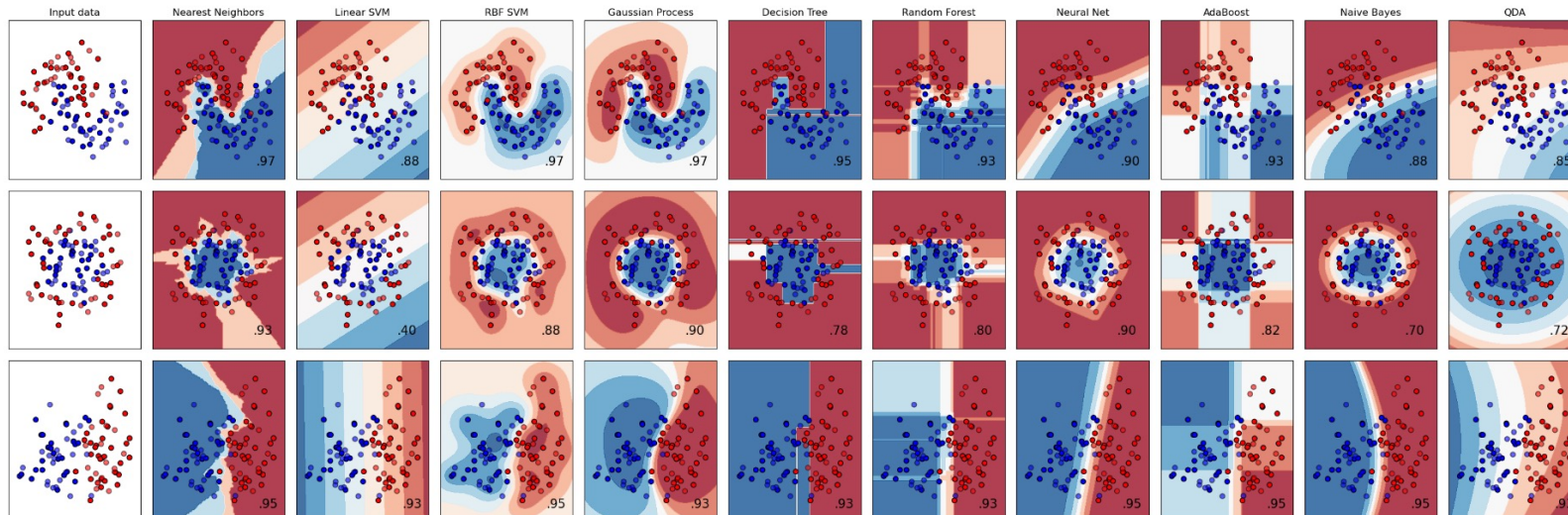
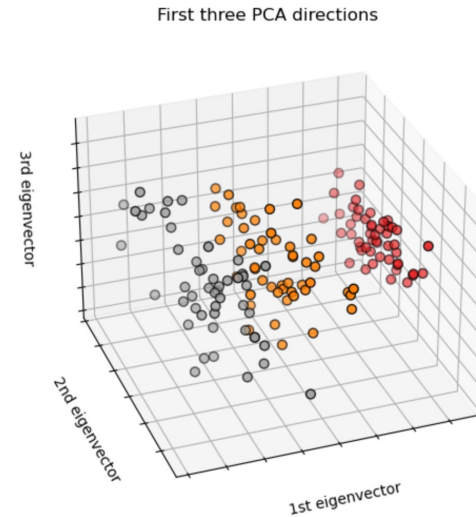
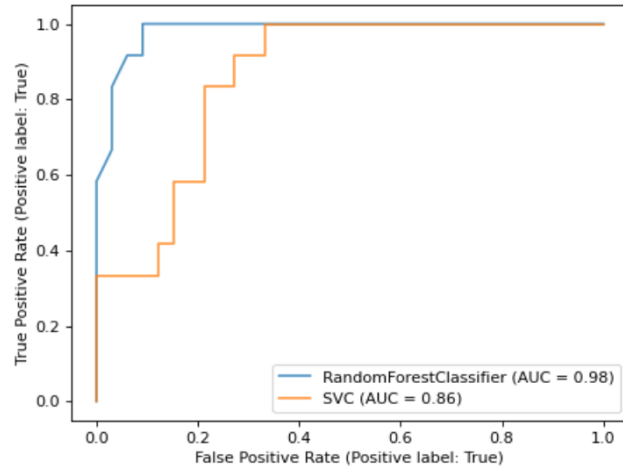
Methods Figure

- Visually describe what you did
- Example (for a much more involved project):

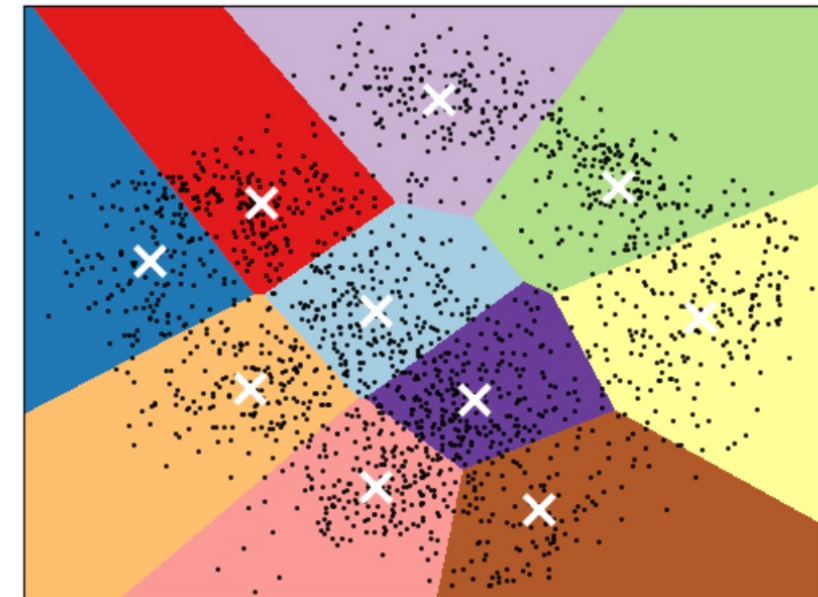


Results Figure(s)

- Show plots, visualizations, and tables describing your results



K-means clustering on the digits dataset (PCA-reduced data)
Centroids are marked with white cross



Technical Writing Style

- See short applied machine learning papers on Google Scholar for examples
 - Short workshop papers at ML conferences
 - Condensed versions of journal papers
- Examples of papers at roughly the level expected for this class:
 - <https://aclanthology.org/2021.smm4h-1.29.pdf>
 - <https://s3.us-east-1.amazonaws.com/climate-change-ai/papers/neurips2021/27/paper.pdf>
 - <https://arxiv.org/pdf/2101.04012.pdf>
 - <https://essamsleiman.com/images/AAAI-23.pdf>
 - <https://arxiv.org/pdf/1811.08592.pdf>

Final Project Video

- Video can summarize any aspect about your project. There are no strict requirements other than communicating what you did.
- Video should be uploaded to YouTube (can be unlisted). Submit the link on Laulima.
- Video should be no longer than 2 minutes.
- Videos will be viewed in class.
- **You must show up to class during the project video presentation days to get the full point for the video, unless you are excused.**

Tues Apr 25 Project Videos Part 1

Thur Apr 27 Project Videos Part 2

Tues May 02 Course Review

[Final Project](#) Presentation Video due

[Homework 5](#) due

Possible Video Content

- Talking through slides describing your project
- Demo video walking through your code and results
- Creative video showing your model in use
- Song / rap about your models
- Poetry about your models
-
- Any combination of the above

Key Takeaways

- Have fun!
- Explore something new
- Use this as an opportunity to build your portfolio and resume
- Highly recommended but optional: post your project on GitHub, describe on your resume, submit report to arxiv

Questions?