# CSDS 435 Project 1

by rmc170, wxy320

# Code Running

Create a python venv.

Install dependencies from requirements.txt

Run main.ipynb

# Algorithm Descriptions and Results

## Data Preprocessing

For data preprocessing we used scikit-learn's `StandardScaler` to remove the mean and scale to the unit variance.

## Algorithms

- Nearest Neighbor
  - Implemented using sklearn's KNeighborsClassifier with n_neighbors set to 5.
- Decision Tree
  - Implemented using sklearn's DecisionTreeClassifier with max_depth set to 5.
- Naïve Bayes
  - Implemented using sklearn's GaussianNB with default parameters.
- SVM
  - Implemented using sklearn's SVC with kernel set to 'rbf', C set to 1, and gamma set to 'scale'.
- Neural Network
  - Uses scikit-learn to train a simple classifier. Then evaluates the classifier.
  - For determining hyperparameter we used `GridSearchCV` to evaluate diferent parameters. To see the optomized parameters check them out in `neural_network.py`. For data set 1 the default stores almost optomized so the improvement is not great. In data set 2 the accuracies are about the same but, precision, recall, and F1 are much higher.

## Results

- Here we show the average results on the 10-fold Cross Validation. ### Dataset 1

| Method | Accuracy | Precision | Recall | F1 Measure |
|---|---|---|---|---|
| Nearest Neighbor | 0.966635 | 0.978279 | 0.928055 | 0.950699 |
| Decision Tree | 0.924373 | 0.902202 | 0.890336 | 0.894694 |
| Naïve Bayes | 0.934900 | 0.917871 | 0.904443 | 0.910032 |
| SVM | 0.977162 | 0.973755 | 0.964308 | 0.968596 |
| Neural Network | 0.971930 | 0.981384 | 0.941080 | 0.959696 |

**Dataset 2**

| Method | Accuracy | Precision | Recall | F1 Measure |
|---|---|---|---|---|
| Nearest Neighbor | 0.653978 | 0.514557 | 0.379458 | 0.425621 |
| Decision Tree | 0.670953 | 0.551663 | 0.392321 | 0.450329 |
| Naïve Bayes | 0.707724 | 0.575085 | 0.620702 | 0.591568 |
| SVM | 0.718455 | 0.646300 | 0.437677 | 0.510323 |
| Neural Network | 0.727197 | 0.634824 | 0.538305 | 0.570610 |

# Comparsion of Methods

- From the results on Dataset 1, the SVM model achieved the highest accuracy (0.977162) and F1 Measure (0.968596), closely followed by the Neural Network with an F1 Measure of 0.959696. Nearest Neighbor also performed well, achieving a strong balance between precision and recall. Decision Tree had the lowest accuracy (0.924373) among the models, indicating that it might not generalize as well on this dataset.
- For Dataset 2, Neural Network achieved the highest accuracy (0.727197) and F1 Measure (0.570610), outperforming other models in recall and precision. Naïve Bayes performed well in terms of recall (0.620702), suggesting it handled class imbalances better. SVM, despite performing well on Dataset 1, had a relatively lower recall (0.437677) in Dataset 2. Nearest Neighbor had the weakest recall (0.379458), indicating it might struggle with the dataset's distribution.
- Key Takeaways:
    1. SVM and Neural Networks consistently performed the best across both datasets.
    2. Naïve Bayes showed strong recall in Dataset 2, making it potentially useful for imbalanced datasets.
    3. Decision Trees and Nearest Neighbors performed worse than other models, suggesting they may be more dataset-sensitive.