

CS 4267

Assignment 3. Due 04/16/2024. 100 points

Use sklearn inbuilt datasets:

```
from sklearn import datasets
```

import some data

```
iris = datasets.load_iris()
```

#retrieve the data

```
x = iris.data
```

```
y = iris.target
```

BOSTON HOUSING Data and K-Means Cluster Analysis

```
import pandas as pd
```

```
import numpy as np
```

```
data_url = "http://lib.stat.cmu.edu/datasets/boston"
```

```
raw_df = pd.read_csv(data_url, sep="\s+", skiprows=22, header=None)
```

```
data = np.hstack([raw_df.values[::2, :], raw_df.values[1::2, :2]])
```

```
target = raw_df.values[1::2, 2]
```

```
from sklearn.datasets import load_breast_cancer
```

```
>>> data = load_breast_cancer()
```

```
>>> list(data.target_names)
```

```
['malignant', 'benign']
```

Purpose

Learn how clustering, ensemble and decision trees works.

Develop established prediction-based algorithm from scratch using python programming language.

Description

- 1) **Create a k-means clustering algorithm from scratch and apply** it on the Boston dataset, create 4 clusters, you may ignore the RAD column because it is categorical or you may use 1-hot encoding to represent it. Visualize your results using scatter plots, you may use t-SNE if you want.
- 2) **Create a decision tree algorithm from scratch** using **Gini Index** or Information gain based on Entropy as your splitting criteria to **classify the Iris datasets**.
- 3) **Apply the logistic regression, Neural network and k-NN algorithms you developed in previous assignments** to create an ensemble classifier for the **breast cancer datasets**. Show your classification with a split of 70% training and 30% testing. Output the classification accuracy.

Deliverables

The following additional comments apply:

A good programming style must be observed.

Your source code must be runnable using Google colab (free cloud-based web service). Using Python libraries for data arrays, splits, and matrices manipulations is fine.

Students should work independently. Each student is responsible for handing in an original program.