Supervised Learning vs Unsupervised Learning

Description of Project

I will be using both supervised and unsupervised learning in this project to show their use case in different projects. For supervised learning I will be using a Heart Disease dataset to train the model and predict who has heart disease and who doesn't. Another supervised learning method I will be using will be a Decision Tree and I will also use Naive Baysian. However, for Unsupervised learning I will be using a different Dataset with patients that have diabetes.

Instructions

Since this is a machine learning project, training the model is a requirement. Because of this, it takes a super long time to train a model (1+ hours) so I already pre trained the model before the class. There will certain areas you can run to show the accuracy of the model and view some visuals and these are the sections you can do it:

- I put ******* on sections you can run the program that will give you the accuracy of that classification and avoid a long runtime. You will see this notation in both the supervised and unsupervised portions of the project.
- For the unsupervised section, you can change the K value to get different number of clusters.
    - 1st: Change the variable "K" in section 4 to any number between 1-7
    - 2nd: Run that cell and the next two to see the scatter plot change.
- This is used to change the number of clusters you want to use in the scatter plot

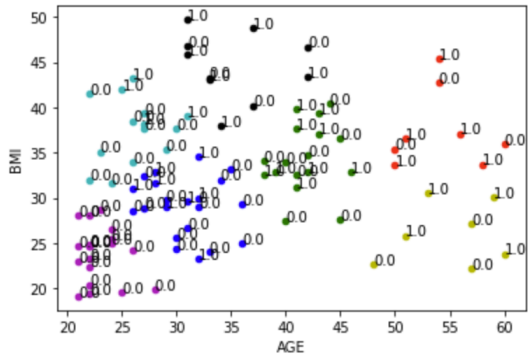
Sample cases/Results

```
[63] x = diabetes_df[["BMI","Age"]]
     k = 7
     clt = KMeans(n_clusters=k)
     clt.fit(x)
     cluster_lables = clt.predict(x)
```

```
[65] colors = ["b","r","g","c","m","y","k"]
     plt.figure(figsize=(16,10))
```

```
     <Figure size 1152x720 with 0 Axes>
     <Figure size 1152x720 with 0 Axes>
```

```
for i, row in diabetes_df.iterrows():
  curr_label = cluster_lables[i]
  curr_color = colors[curr_label]
  plt.scatter(row["Age"],row["BMI"],c=curr_color, s=20)
  plt.text(row["Age"],row["BMI"],row["Outcome"],size=10)
plt.xlabel("AGE")
plt.ylabel("BMI")
plt.show()
```



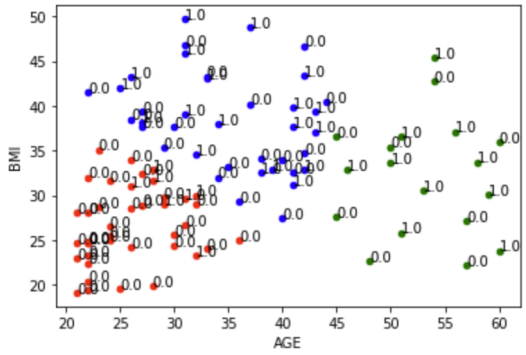
With k = 7 and there being 7 clusters in the scatter plot

```
[67] x = diabetes_df[["BMI","Age"]]
     k = 3
     clt = KMeans(n_clusters=k)
     clt.fit(x)
     cluster_lables = clt.predict(x)
```

```
[68] colors = ["b","r","g","c","m","y","k"]
     plt.figure(figsize=(16,10))
```

```
     <Figure size 1152x720 with 0 Axes>
     <Figure size 1152x720 with 0 Axes>
```

```
[69] for i, row in diabetes_df.iterrows():
       curr_label = cluster_lables[i]
       curr_color = colors[curr_label]
       plt.scatter(row["Age"],row["BMI"],c=curr_color, s=20)
       plt.text(row["Age"],row["BMI"],row["Outcome"],size=10)
     plt.xlabel("AGE")
     plt.ylabel("BMI")
     plt.show()
```



With k = 3 and there being 3 cluster in the scatter plot.

Credits:

https://www.cdc.gov/diabetes/basics/risk-factors.html

https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease (Heart Disease Dataset)

https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset (Diabetes Dataset)