Breast Cancer Detection

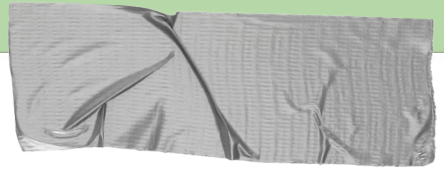
Author: Jie Zhang
Term: Spring 2023
Professor: Karl Ni
Course: CS6220 Data Mining

Why choose this topic

Real life problem

Breast cancer is a **significant public health concern**, affecting millions of women worldwide. Early detection of breast cancer is critical for successful treatment and improving the chances of survival.

A good chance to explore **data mining techniques and ML concepts**



Intro

What is breast cancer?

Why we need early detection?

How to use ML to detection breast cancer in my Project?

(current survey and my approach)

Surprise Fact:

➔ **Rate:**

1 out of 10 women will face breast cancer in the life time.

52%

decline in prostate cancer from 1993-2015

decline in breast cancer in women from 1989-2015

39%

52%

decline in colorectal cancer in both genders from 1970-2015

decline in lung cancer in men from 1990-2015

45%

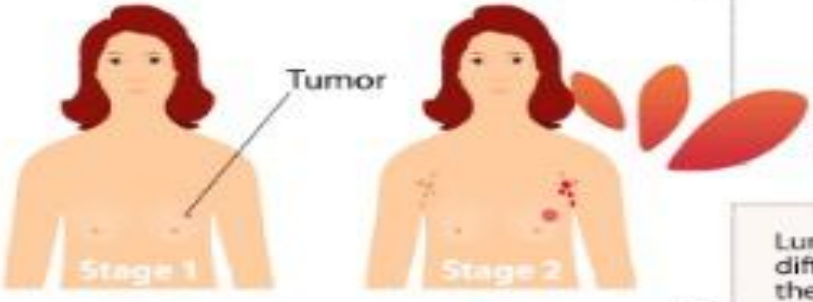
19%

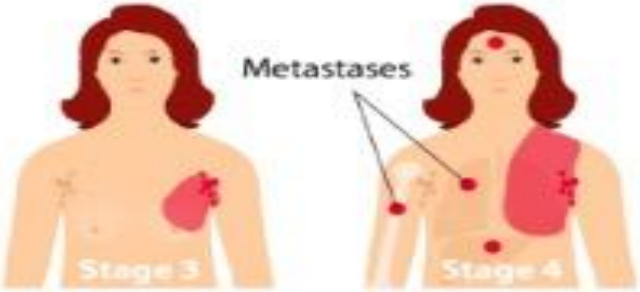
decline in lung cancer in women from 2002-2015

All About Breast Cancer

1 out of 10 women have a breast cancer

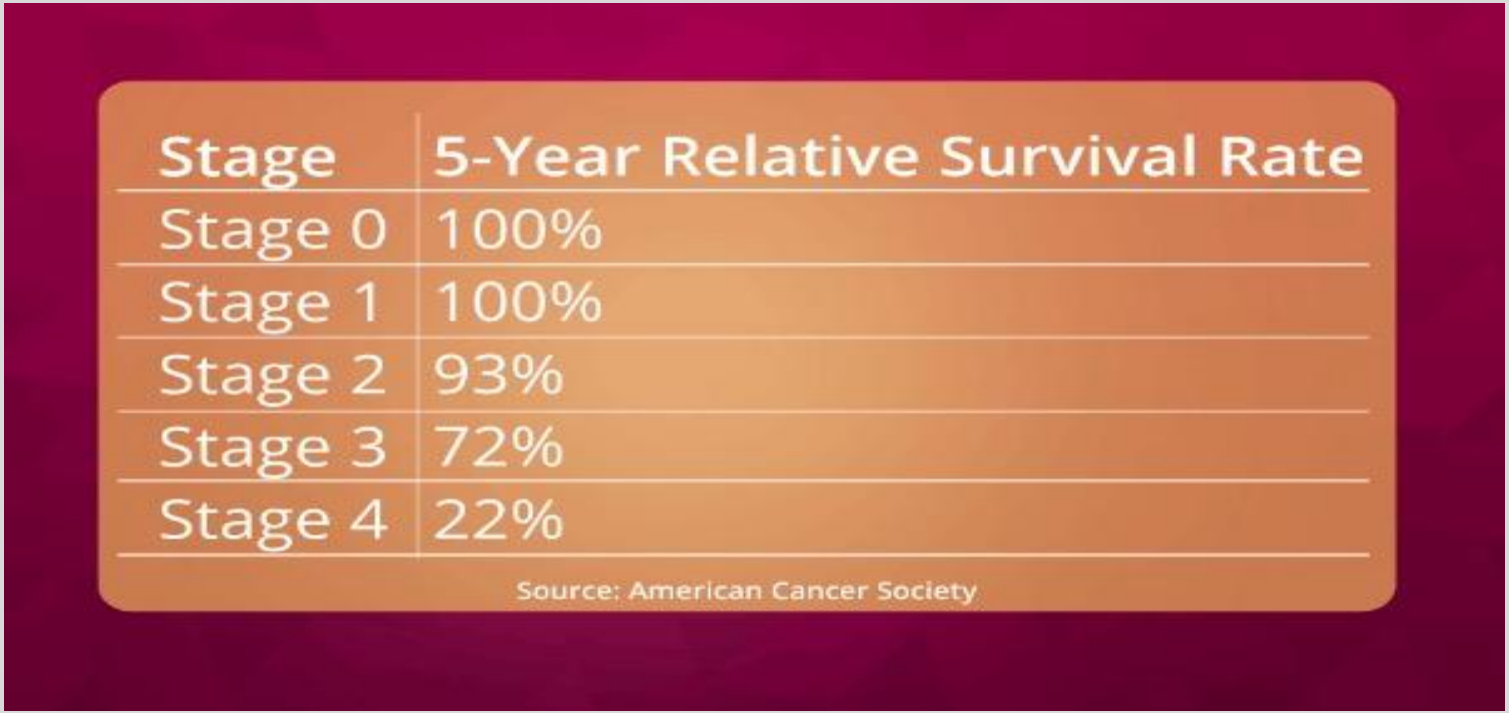
Risk





Symptoms

Why we need early detection?


Stage
5-Year Relative Survival Rate
Stage 0
100%
Stage 1
100%
Stage 2
93%
Stage 3
72%
Stage 4
22%
Source: American Cancer Society

Currently Research Outcome:

Machine Learning Method

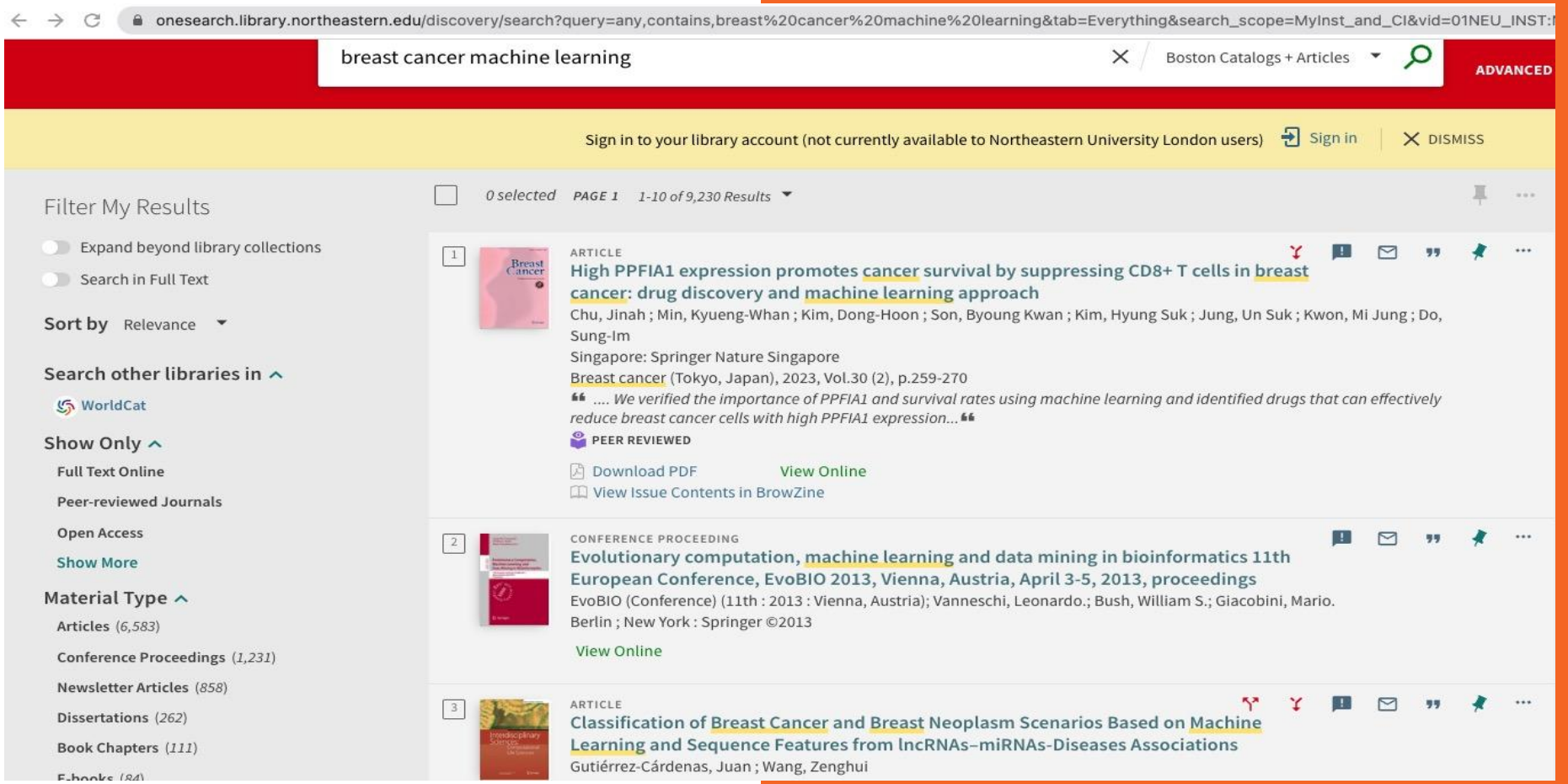
Medical area

Tip

Don't wait till the end of the presentation to give the bottom line.

Reveal your product or idea (in this case a translation app) up front.

Currently Research Outcome:



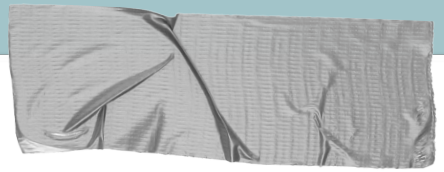
Currently Research Topic:

Images Analysis
Classification
Models for prediction or diagnosed the stages for breast cancer.
Decision Tree for the treatments or further detection
Computing
Data Visualization
…

Tip

Don't wait till the end of the presentation to give the bottom line.

Reveal your product or idea (in this case a translation app) up front.



. My Approach

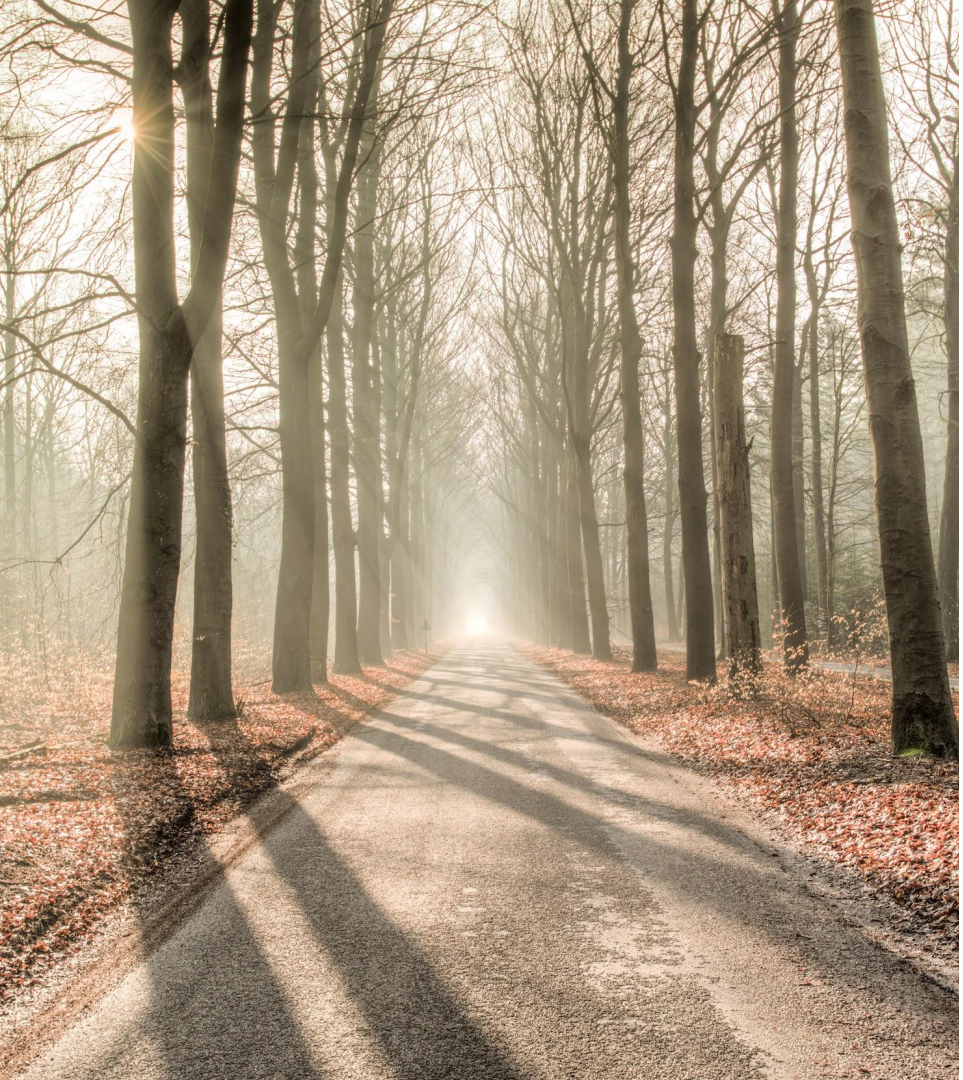
- ➔ **Data Cleaning and Transform**
- ➔ **Data Visualization**
- ➔ **Evaluation Metric in Machine Learning**

Choose Dataset :

My skills and knowledge,

Currently resource,

Reasonable techniques



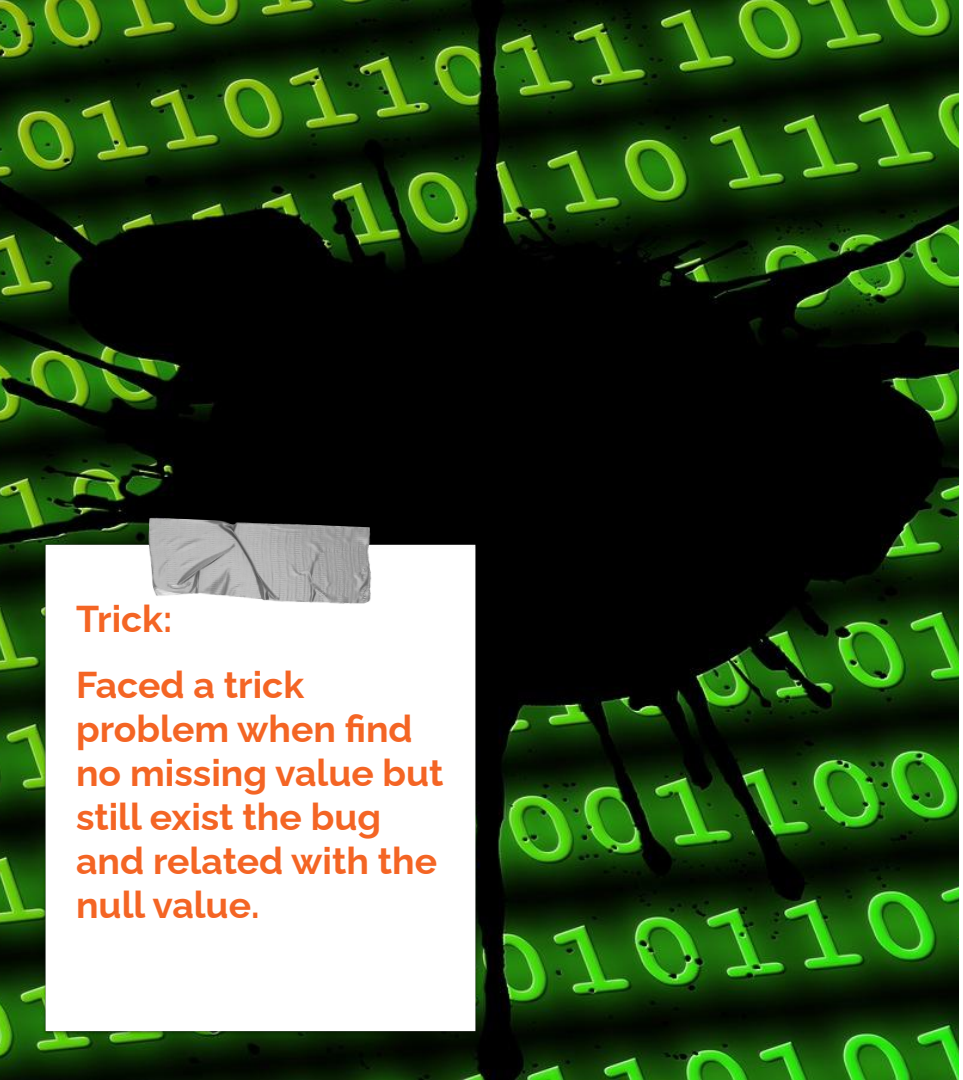
Where is my path?

Google drive?

Local device?

Cloud?

…



Trick:

Faced a trick problem when find no missing value but still exist the bug and related with the null value.

Missing data or null data:

Play around and check.

Focus.

Meet different people.

Research online.

Learn from the practice.

Take a break.

Story for illustration purposes only


Some techniques and knowledge will come after these.
FOCUS.
otherwise you will find life becomes a blur.
purehappylife.com

Outcome:

Research interest.

[Heatmap]

[K-Fold Cross Validation]

HeatMap

```python
# just learn the orginall idea for heatmap implementation, can see it is a type of data plot function.
import numpy as np
import matplotlib.pyplot as plt

# Create some random data
data = np.random.rand(5, 5)

# Create a heatmap of the data
fig, ax = plt.subplots()
im = ax.imshow(data)

# Add a color bar
cbar = ax.figure.colorbar(im, ax=ax)

# Set the x and y axis tick labels
ax.set_xticks(np.arange(data.shape[1]))
ax.set_yticks(np.arange(data.shape[0]))
ax.set_xticklabels(['A', 'B', 'C', 'D', 'E'])
ax.set_yticklabels(['1', '2', '3', '4', '5'])

# Rotate the tick labels and set their alignment
plt.setp(ax.get_xticklabels(), rotation=45, ha="right",
         rotation_mode="anchor")

# Loop over data dimensions and create text annotations.
for i in range(data.shape[0]):
    for j in range(data.shape[1]):
        text = ax.text(j, i, f'{data[i, j]:.2f}',
                       ha="center", va="center", color="w")
```

Display the relationship between different values in a matrix or table of data,

Color encoding

Correlation vs coefficients between two variables,

Displaying the correlation between different features in a multidimensional dataset.

Tip

Stories become more credible when they use concrete details such as the specific complex moves Alberto learned through Translate and his 30 goals in 21 games performance stats.

```
# using  Pandas library and Seaborn library in Python to create a heatmap of a correlation matrix.

# 1)Selects a subset of columns from a Pandas DataFrame called 'df' using the variable 'features_mean'.
# 2)Computes the correlation between these columns using the Pandas corr() method and stores the resulting correlation matrix in the variable 'corr'.
# 3)creates a square figure with a size of 14 by 14 using the Seaborn plt.figure() function.
# uses the Seaborn heatmap() function to create a heatmap of the correlation matrix with the following arguments:

#cbar=True: show a colorbar next to the heatmap to indicate the range of values.
#square=True: make the cells of the heatmap square.
#annot=True: show the correlation coefficients as annotations inside the cells.
#fmt='.2f': format the annotation values as floating-point numbers with two decimal places.
#annot_kws={'size': 15}: set the font size of the annotations to 15.
#xticklabels=features_mean: label the x-axis ticks with the names of the selected columns.
#yticklabels=features_mean: label the y-axis ticks with the names of the selected columns.
#cmap='coolwarm': use the 'coolwarm' colormap to represent the corr

corr = df[features_mean].corr()
plt.figure(figsize=(14,14))
sns.heatmap(corr, cbar = True,  square = True, annot=True, fmt= '.2
           xticklabels= features_mean, yticklabels= features_mean,
           cmap= 'coolwarm')
```

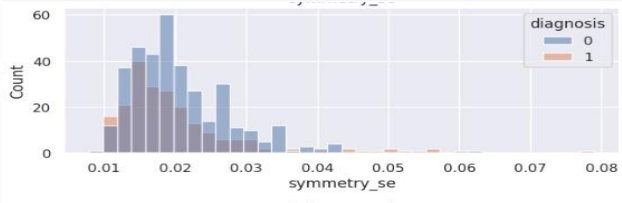
Histogram:

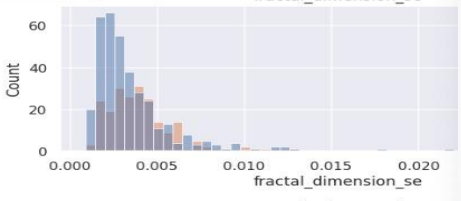
A powerful tool to see what going on in the data set.

```
    'radius_se', 'texture_se', 'perimeter_se', 'area_se', 'smoothness_se',
    'compactness_se', 'concavity_se', 'concave_points_se', 'symmetry_se',
    'fractal_dimension_se', 'radius_worst', 'texture_worst',
    'perimeter_worst', 'area_worst', 'smoothness_worst',
    'compactness_worst', 'concavity_worst', 'concave_points_worst',
    'symmetry_worst', 'fractal_dimension_worst']
ig = plt.figure(figsize=(20,30))

or i in range(len(num_list)):
    plt.subplot(10,3,i+1)
    plt.title(num_list[i])
    sns.histplot(data=df,x=df[num_list[i]],hue='diagnosis')

lt.tight_layout()
```





texture_worst

perimeter_worst

K-Fold Cross Validation

Machine Learning Model Evaluation Method

K-Fold Cross Validation

Detect whether the model is overfitting

Improve the accuracy of evaluation

```
[ ] model = svm.SVC()
    classification_model(model,data,prediction_var,outcome_var)

[ ] model = KNeighborsClassifier()
    classification_model(model,data,prediction_var,outcome_var)

[ ] model = RandomForestClassifier(n_estimators=100)
    classification_model(model,data,prediction_var,outcome_var)

[ ] model=LogisticRegression()
    classification_model(model,data,prediction_var,outcome_var)

[ ] data_X= data[prediction_var]
    data_y= data["diagnosis"]

[ ] def Classification_model_gridsearchCV(model,param_grid,data_X,data_y):
        clf = GridSearchCV(model,param_grid,cv=10,scoring="accuracy")
        clf.fit(train_X,train_y)
        print("The best parameter found on development set is :")
        print(clf.best_params_)
        print("the best estimator is ")
        print(clf.best_estimator_)
        print("The best score is ")
        print(clf.best_score_)

[ ] param_grid = {'max_features': ['auto', 'sqrt', 'log2'],
                  'min_samples_split': [2,3,4,5,6,7,8,9,10],
                  'min_samples_leaf':[2,3,4,5,6,7,8,9,10] }
    model= DecisionTreeClassifier()
    Classification_model_gridsearchCV(model,param_grid,data_X,data_y)
```

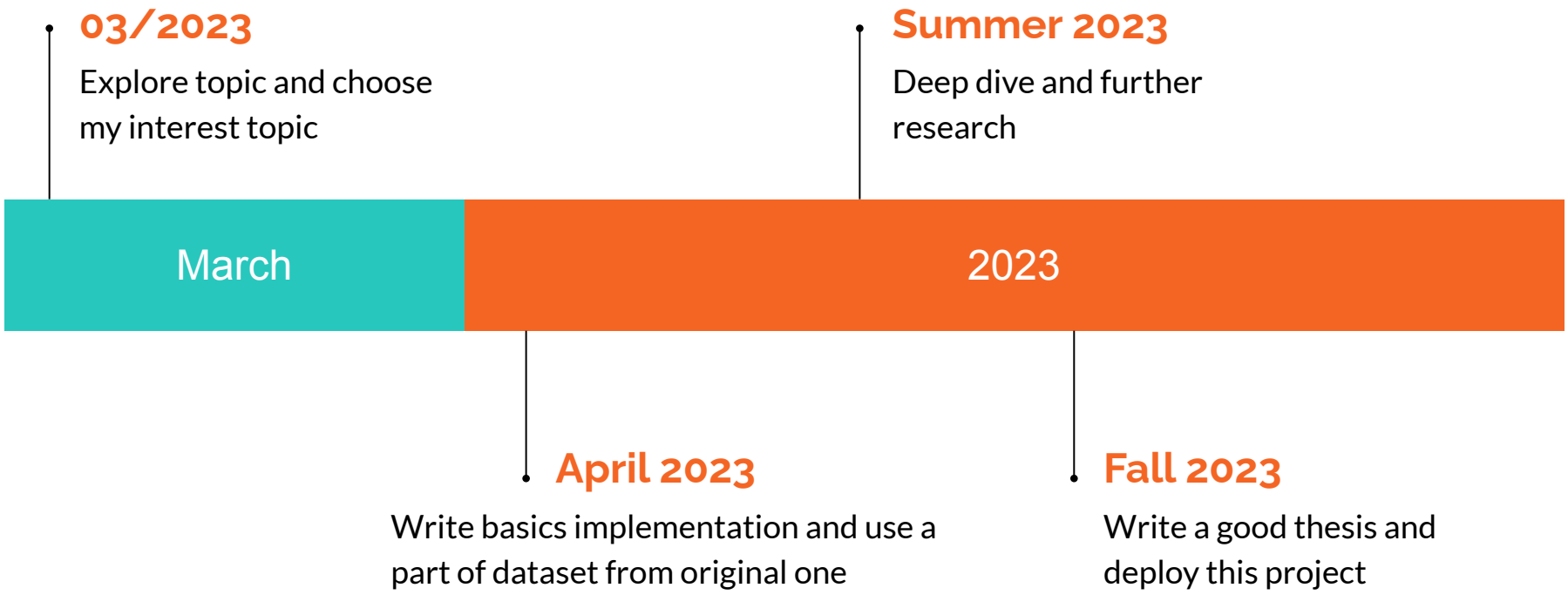
Milestones

(no ending in this course)

Test and further outcome

(data distributed, data visolization, data transform, model adjust and evaluate)

Milestones


03/2023
Explore topic and choose my interest topic
Summer 2023
Deep dive and further research
March
2023
April 2023
Write basics implementation and use a part of dataset from original one
Fall 2023
Write a good thesis and deploy this project

Reference for this project:

1.https://onesearch.library.northeastern.edu/discovery/fulldisplay?docid=cdi_crossref_primary_10_3991_ijoe_v18i05_29197&context=PC&vid=01NEU_INST:NU&lang=en&search_scope=MyInst_and_CI&adaptor=Primo%20Central&tab=Everything&query=any,contains,breast%20cancer%20machine%20learning&offset=10

2.https://www.kaggle.com/code/malik12345/breast-cancer-detection-using-ml/notebook

3.https://www.kaggle.com/datasets/nancyalaswad90/breast-cancer-dataset

4.https://www.kaggle.com/code/usakshaya/breast-cancer-prediction

5.https://machinelearningmastery.com/k-fold-cross-validation/

6.https://en.wikipedia.org/wiki/Histogram

7.https://towardsdatascience.com/different-ways-to-connect-google-drive-to-a-google-colab-notebook-pt-1-de03433d2f7a

8.https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9175124/

9.https://www.freecodecamp.org/news/how-to-handle-missing-data-in-a-dataset/

Reference for this project:

10..https://www.v7labs.com/blog/overfitting

11.https://www.youtube.com/watch?v=qxpKCBV60U4

12.https://scikit-learn.org/stable/modules/cross_validation.html

13.https://stackoverflow.com/questions/73443407/svm-problem-name-model-svc-is-not-defined

14.https://www.youtube.com/watch?v=tLhunN5Jhqs

15.https://www.youtube.com/watch?v=Hlk5psu5yFw

Healthy guideline:

Healthy Habits

Leading a healthy lifestyle is recommended to protect your overall health and may help reduce your risk for certain cancers.

Here are a few tips to follow:

- Eat five or more servings of fruits and vegetables each day.
- Get regular physical activity.
- Maintain a healthy weight.
- Limit alcohol intake to no more than one drink per day.
- Do not smoke. Or, quit smoking.

Check:

Scheduling Exams

While living a healthy life can help reduce your risk for cancer, women can be diagnosed with breast cancer at any age. Detecting breast cancer at an early stage, when treatment is more likely to be successful, still provides the best hope for survival. This is why it is so important for you to schedule regular exams. Below you will find some general guidelines for breast cancer early detection methods. *You should always consult with your doctor to create a screening schedule that is most appropriate for you.*

EXAM	AGE	FREQUENCY
Breast Self-Awareness	18+	Regularly/Monthly
Well-Woman Exam	21+	Yearly
Mammogram	40+	Yearly

Thanks!