My research project was about AI (Artificial Intelligence) taking data from other patients about what types of symptoms they had during cervical cancer like what is their age, how many sexual partners they had etc. Then based on that data, can AI detect if a random patient is also diagnosed with cervical cancer based on the data they have provided?

Cervical cancer is cancer that starts in the cells of the cervix (a lower narrow end of the uterus). Before cancer appears in the cervix, the cells of the cervix go through changes known as dysplasia, in which abnormal cells begin to appear in the cervical tissue[1]. Over time, if these cells are not removed or destroyed these abnormal cells can lead to cancer cells and will start to grow and spread deeply into the cervix and to the surrounding areas[2]. I am doing this because cervical cancer was a sickness on which data could easily be found and many factors/diagnoses could lead to cervical cancer. There are many things that cause cervical cancer, but one of the main causes of cervical cancer is Human Papillomavirus(HPV)[3]. HPV affects your genitals, and the most common sign of the virus is warts in your genital area. Genital warts are rough, cauliflower-like lumps that grow on your skin. Genital warts are contagious but not harmful like all forms of HPV. High risk forms of HPV often don't cause symptoms until they have progressed to cancer. Cervical cancer is the most common type of HPV related cancer[4] . This is a virus that can be passed from one person to another during sex. The dataset I have chosen also includes this virus as a risk factor of Cervical Cancer. Women have to take a PAP test in order to figure out if they are diagnosed with cervical cancer.

At the start, I found my data on cervical cancer from a website called UCI Machine Learning Repository. On this website, there was a data set about cervical cancer and it included the different causes that could potentially lead to cervical cancer. After that, I uploaded that file to google collaboratory and cleaned the data. By cleaning the data I mean more than half of the data was impractical because it was missing. To clean the data I set up a threshold in my code that dropped any rows and columns where more than half of the data was missing. Then I imported a bunch of different libraries which were Numpy, Pandas, and learn. These libraries have their unique purposes. After importing the libraries I took the first 27 columns as my input data and the last column called biopsy as my output data. Here is a overview of what my inputs my data had:

Age	STDs	STDs:Hepatitis B
Number of sexual partners	STDs (number)	STDs:HPV
First sexual intercourse	STDs:condylomatosis	STDs: Number of diagnosis
First sexual intercourse	STDs:cervical condylomatosis	STDs: Time since first diagnosis
Smokes	STDs:vaginal condylomatosis	STDs: Time since last diagnosis
Smokes (years)	STDs:vulvo-perineal condylomatosis	
Smokes (packs/year)	STDs:syphilis	
Hormonal Contraceptives	STDs:pelvic inflammatory disease	
Hormonal Contraceptives (years)	STDs:molluscum contagiosum	

IUD (years)	STDs:HIV	
Schiller	Cytology	

From there I split my data into tests and training. The test size of my data was 30% and the train size of my data was 70%. After that, I used a StratifiedShuffleSplit, which was used to split my data randomly. Then I implemented a feature scale for helping my ROC AUC value get better. Then I choose Logistic Regression as the model. The metric that I chose for my data was called ROC AUC and then I implemented that in my code. Some things that I added to make my ROC AUC value get better was that I used max iter(1 * 10^5), classweight, standard fit transform and standard transform. After that, I did 10 trials to find out the average of my ROC AUC value. The average value of the ROC/AUC I got was 0.68. The standard deviation of this plot was 0.0932.

Trials that were done to figure out the average value:

Trials	Test/Train Split	Random State	ROC/AUC value
1	Test:20% Train:80%	1	0.688
2	Test:20% Train:80%	2	0.489
3	Test:20% Train:80%	3	0.781
4	Test:20% Train:80%	4	0.652
5	Test:20% Train:80%	5	0.630
6	Test:20% Train:80%	6	0.532
7	Test:20% Train:80%	7	0.658
8	Test:20% Train:80%	8	0.449
9	Test:20% Train:80%	9	0.586

10	Test:20% Train:80%	10	0.603

The code to figure out the Standard Deviation and ROC/AUC value:

```
[87]  1  # Figuring out the mean
      2  average_scores = [0.688, 0.489, 0.781, 0.652, 0.630, 0.532, 0.658, 0.449, 0.586, 0.603]
      3  np.mean(average_scores)
```

```
0.6068
```

```
1  # This is going to be the standered diviation value
2  np.std(average_scores)
```

```
0.09325320369831805
```

The top 4 Coefficients that were the most important and which showed as important factors with almost every random state that was changed were STDs:condylomatosis, STDs HIV, STDs:molluscum contagiosum, Smokes(years), and Number of sexual partners. These coefficients impacted the data in a lot of good ways by helping the ROC/AUC value get better.

In conclusion, the average ROC/AUC(a.k.a regression value) came to a 0.68 which is not that good in the field of medical studies however this value could be somewhat more promising if there was feature engineering feature selection, algorithm tuning and cross validation. All these types of methods can help enhance the accuracy of the model and the AI can detect better than it was before.

Work Cited

1) “Basic Information About Cervical Cancer | CDC.” *Centers for Disease Control and Prevention*, https://www.cdc.gov/cancer/cervical/basic_info/index.htm. Accessed 7 April 2023.

2) “Cervical cancer - Symptoms and causes.” *Mayo Clinic*, 14 December 2022, https://www.mayoclinic.org/diseases-conditions/cervical-cancer/symptoms-causes/syc-20352501. Accessed 7 April 2023.

3) “Cervical cancer - Symptoms and causes.” *Mayo Clinic*, 14 December 2022, https://www.mayoclinic.org/diseases-conditions/cervical-cancer/symptoms-causes/syc-20352501. Accessed 7 April 2023.

4) “HPV Human Papillomavirus: Causes, Symptoms & Treatment.” *Cleveland Clinic*, 4 August 2022, https://my.clevelandclinic.org/health/diseases/11901-hpv-human-papilloma-virus. Accessed 7 April 2023.