Group 14 - Andrew Vuong, Henry Wang, Aaron Johnson

1. Describe the dataset. (ex: The number of microaneurysms found in a patient's eye and whether or not they have diabetic retinopathy; or Lending Tree loan data, including who defaulted and who paid off their loan.)

From the dataset abstract found on Harvard Dataverse collected by Westre B, Giske A, Guttormsen H, Wergeland Sørbye S, Skjeldestad FE (2019) Quality control of cervical cytology using a 3-type HPV mRNA test increases screening program sensitivity of cervical intraepithelial neoplasia grade 2+ in young Norwegian women—A cohort study. doi: 10.1371/ journal.pone.0221546.

Essentially we are given deidentified patient data with marker tests for cervical cancer including a graded level of cancer biopsy result serving as the label. There are various other features we are also taking into consideration to determine its impact on the classification.

2. How many records does the dataset have?

4366 records

3. How many features does the dataset have? List or describe a few of them.

10 features with label of grade of cancer or no cancer:

```
['Age', 'study_gr', 'study_gr1', 'Diag_cyt', 'Diag_cyt_rev', 'SEE_01',
    'SEE_16', 'SEE_18', 'SEE_45', 'Ind_biop', 'biop_status']
```

Mainly different cervical HPV mRNA markers, types of study, age, method of cytology, meta study.

4. What can you try to predict in this dataset? (ex: We can use the number of microaneurysms measured in the patient's eye to predict whether or not they have diabetic retinopathy; or We can try using the features, including age, income, home ownership status, etc, to predict whether or not someone will default on their loan.)

The project will mainly attempt to correct markers and age with either binary classification of cancer / no cancer by binning the graded level of cancers or a multi-classification of graded level of cancer. We feel it would be more beneficial to do a binary classification because given the time of cancer detection affects the grade level of cancer, meaning the longer left untreated, all cancer progress towards the highest grade.

5. Is this a **labeled** dataset, appropriate for a supervised learning classification problem? (In other words, if you are trying to predict whether or not someone has a disease, does your dataset contain whether or not each record has the disease?)

This is a labeled dataset with labels of diseased or not for each record.

6. Provide a link to the dataset, if there is one. If you are getting your data from somewhere other than a link, where are you getting it from?

Dataset found from Harvard Dataverse Database:

https://dataverse.harvard.edu/dataverse/harvard?q=&fq0=subject_ss%3A%22Medicine%2C+Health+and+Life+Sciences%22&types=dataverses%3Adatasets&sort=dateSort&order=desc&page=2

Exact link to dataset - Quality control of cervical cytology using a 3-type HPV mRNA test increases screening program sensitivity of cervical intraepithelial neoplasia grade 2+ in young Norwegian women:

https://dataverse.no/dataset.xhtml?persistentId=doi:10.18710/TDJV8X