Project Proposal

Problem Statement: Can we create an AI model that can predict the results of various cervical cancer tests with an accuracy of 80% or higher?

This problem uses a dataset similar to the cancer risk prediction dataset from the project topics before. I chose this one specifically because the other cancer set looked too clean to the point where there would not be much of any data wrangling to be done, and this one has blank values and will require more effort, so it will be a better project to learn from.

This project uses data collected at a hospital in Venezuela that contains details of 800 patients. The data includes their age, details about how long they have been sexually active, what STDs they have, how long they have smoked, how many times they have been pregnant, their diagnosis for a few different conditions, etc. The labels in this data set are results of different tests for cervical cancer including Hinsellman and Schiller tests, citologies, and biopsies.

This problem follows the SMART principles:

Specific - It gives a baseline percentage of accuracy we would like to see from the AI model trained

Measurable - The accuracy of the predictions can be measured by simply dividing the correct predictions by the total number of predictions to find the ratio of them which correctly predict how the tests perform.

Action-Oriented - We must actively clean the data and train an AI model on it and perform different optimizations in order to make the model more accurate and reach the 80% accuracy benchmark.

Relevant - This problem would help to find which test may be the best at detecting cervical cancer in different situations so we can better help protect the lives of people at risk.

Time-Bound - This problem must be solved before the end of this bootcamp, so it is bound by the time this bootcamp will take.

The dataset for this project can be found here:
https://data.world/uci/cervical-cancer-risk-factors