Capstone II – Possible Datasets

1. Cervical Cancer (Risk Factors)

https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29#

· Approximately 860 instances.
· Data available in .csv form.
· Predictive Problem: predict diagnosis of cervical cancer based on demographic information and sexual health history.
· 4 target variables: Hinselmann, Schiller, Cytology, and Biopsy.

2. Diabetes 130-US Hospitals for Years 1999-2008

https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008

· Approximately 100,000 instances.
· Data available in .csv form.
· Information available for patients regarding demographics, admission type, discharge disposition (data dictionary available), time in hospital, number of procedures (invasive or lab), and if readmission was necessary.
· Predictive Problem: predict admission type or discharge disposition based on all factors.
· Major problem: Race is only defined as “Caucasian”, “African American”, or “Other”.

3. Parkinson’s Telemonitoring

https://archive.ics.uci.edu/ml/datasets/Parkinsons+Telemonitoring

· Approximately 5,900 instances.
· Data previously used to predict patients’ Parkinson’s Disease scores on the UPDRS scale.
· Data is based on range of biomedical voice captures.