**Capstone Project 2 Proposal**

**(Predicting Recidivism Using LR and RF)**

**General information:** Statistics show that recidivism is a major issue in the US: The Bureau of Justice Statistics (BJS) reported 67 percent and 76.6 percent of state prisoners released in 2005 were re-arrested within three and five years of release respectively. The justice system in the US predicts an individual’s recidivism potential when making key decisions such as parole, probation, and sentencing. The court systems rely on a software package called Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) to predict the likelihood of recidivism. The recidivism prediction made by an individual could be biased depending on one’s life experiences and perspectives. Software packages could also yield biased results if the weight of features is inexplicably fixed and provided by biased individuals. Machine Learning algorithms can be useful in limiting biases and improving recidivism predictions. The capstone project will attempt to employ data science tools to investigate the possibility of predicting recidivism from learning historical recidivism data.

**Data set**: **Data set**: 3-year recidivism for offenders released from prison for the state of Iowa. (<https://catalog.data.gov/dataset/3-year-recidivism-for-offenders-released-from-prison>). The data includes information that can be subdivided in to general information (age, sex, race), original offense information (offense classification, offense type, offense subtype, type of release), and recidivism information (days to recidivism, new offense classification, new offense type, new offense subtype).

**Who would care about the problem?** Social workers and government (justice system): An excellent recidivism prediction results would help the Justice System to rely on machine learning tools when making key decisions. This would lead to sound use of resources and “fair and unbiased” ruling. An accurate prediction can help prevent crimes and avoid wasting resources on individuals who are least to commit repeat offenses.

**Procedures**

1. Obtain publicly available recidivism data.
2. Data wrangling/cleaning: Investigate type of variables, valid values, and missing values. Clean data including missing data (N/A, empty strings, invalid values)
3. The project will be approached as a classification problem. The dataset has labels (recidivism flag) that show whether a prisoner committed a similar offense after release. I will experiment with Logistic Regression and Random Forest Classifiers (and possibly more algorithms).

**Final Package**

The final submitted package will include:

1. Python notebook with all the data wrangling, exploration, and analysis codes
2. Report
3. Power point slides summarizing the main findings