CAPSTONE 2 PROJECT PROPOSAL
Predicting Healthcare Provider Fraud

The goal of this project is to predict healthcare provider fraud. Fraudulent providers receive insurance payments for services they provided unnecessarily or didn't provide at all. Providers can misrepresent the service they provided by charging for a more expensive service. Duplicate claims are another potential indication of fraud. Fraud increases the cost of health insurance because fraudulent claims force public and private insurers to cover erroneous expenses.

By identifying fraudulent providers, public and private insurance providers can avoid paying for fraudulent claims. Without identifying healthcare provider fraud, ongoing fraud will inflate insurance costs and insurers will not be able to maintain competitive insurance rates.

I will use data from a kaggle competition to solve this problem:
https://www.kaggle.com/rohitrox/healthcare-provider-fraud-detection-analysis .

There are four data sources, describing beneficiaries, outpatient claims, inpatient claims, and provider fraud (detailed description on the next page). The challenge with working with this dataset is to preserve specific information about beneficiaries and claims while comparing healthcare providers. There are approximately 700,000 claims, 200,000 beneficiaries, and 5,410 providers. I will build various classification models (logistic, random forest, support vector, knn), to find the best model to predict healthcare provider fraud.

My deliverables will include code, a written report, and slide deck.

1. Beneficiary Data
    a. BeneID
    b. DOB
    c. DOD
    d. Gender
    e. Race
    f. RenalDiseaseIndicator
    g. State
    h. County
    i. NoOfMonths_PartACov
    j. NoOfMonths_PartBCov
    k. ChronicCond_Alzheimer
    l. ChronicCond_Heartfailure
    m. ChronicCond_KidneyDisease
    n. ChronicCond_Cancer
    o. ChronicCond_ObstrPulmonary
    p. ChronicCond_Depression
    q. ChronicCond_Diabetes
    r. ChronicCond_IschemicHeart
    s. ChronicCond_Osteoporasis
    t. ChronicCond_rheumatoidarthritis
    u. ChronicCond_stroke
    v. IPAnnualReimbursementAmt
    w. IPAnnualDeductibleAmt
    x. OPAnnualReimbursementAmt
    y. OPAnnualDeductibleAmt
2. Inpatient Data
    a. BeneID
    b. ClaimID
    c. ClaimStartDt
    d. ClaimEndDt
    e. Provider
    f. InscClaimAmtReimbursed
    g. AttendingPhysician
    h. OperatingPhysician
    i. OtherPhysician
    j. AdmissionDt
    k. ClmAdmitDiagnosisCode
    l. DeductibleAmtPaid
    m. DischargeDt
    n. DiagnosisGroupCode
    o. ClmDiagnosisCode_1
    p. ClmDiagnosisCode_2
    q. ClmDiagnosisCode_3
    r. ClmDiagnosisCode_4
    s. ClmDiagnosisCode_5
    t. ClmDiagnosisCode_6
    u. ClmDiagnosisCode_7
    v. ClmDiagnosisCode_8

w. ClmDiagnosisCode_9
x. ClmDiagnosisCode_10
y. ClmProcedureCode_1
z. ClmProcedureCode_2
aa. ClmProcedureCode_3
bb. ClmProcedureCode_4
cc. ClmProcedureCode_5
dd. ClmProcedureCode_6

3. Outpatient Data
    a. BeneID
    b. ClaimID
    c. ClaimStartDt
    d. ClaimEndDt
    e. Provider
    f. InscClaimAmtReimbursed
    g. AttendingPhysician
    h. OperatingPhysician
    i. OtherPhysician
    j. ClmDiagnosisCode_1
    k. ClmDiagnosisCode_2
    l. ClmDiagnosisCode_3
    m. ClmDiagnosisCode_4
    n. ClmDiagnosisCode_5
    o. ClmDiagnosisCode_6
    p. ClmDiagnosisCode_7
    q. ClmDiagnosisCode_8
    r. ClmDiagnosisCode_9
    s. ClmDiagnosisCode_10
    t. ClmProcedureCode_1
    u. ClmProcedureCode_2
    v. ClmProcedureCode_3
    w. ClmProcedureCode_4
    x. ClmProcedureCode_5
    y. ClmProcedureCode_6
    z. DeductibleAmtPaid
    aa. ClmAdmitDiagnosisCode
4. Provider Data
    a. Provider
    b. Potential Fraud