Part 1: Correlation, regression
We will use acoustic features to distinguish a male voice from female. Load the dataset from "voice.csv", identify the target variable and do a one-hot encoding for the same. Split the dataset in train-test with 20% of the data kept aside for testing.
[Hint: Refer to LabelEncoder documentation in scikit-learn]

2. Fit a logistic regression model and measure the accuracy on the test set.
[Hint: Refer to Linear Models section in scikit-learn]

3. Compute the correlation matrix that describes the dependence between all predictors and identify the predictors that are highly correlated. Plot the correlation matrix using seaborn heatmap.
[Hint: Explore dataframe methods to identify appropriate method]

4. Based on correlation remove those predictors that are correlated and fit a logistic regression model again and compare the accuracy with that of previous model.
[Hint: Identify correlated variable pairs and remove one among them]

Part 2:
1. Let's attempt to predict the survival of a horse based on various observed medical conditions. Load the data from 'horses.csv' and observe whether it contains missing values.
[Hint: Pandas dataframe has a method isnull]

2. This dataset contains many categorical features, replace them with label encoding.
[Hint: Refer to get_dummies methods in pandas dataframe or Label encoder in scikit-learn]

3. Replace the missing values by the most frequent value in each column.
[Hint: Refer to Imputer class in Scikit learn preprocessing module]

4. Fit a decision tree classifier and observe the accuracy.
5. Fit a random forest classifier and observe the accuracy.

Part 3: Project Banking domain
Challenge/requirement
PeerLoanKart is an NBFC (Non-Banking Financial Company) which facilitates peer to peer loan. It connects people who need money (borrowers) with people who have money (investors). As an investor, you would want to invest in people who showed a profile of having a high probability of paying you back.

You create a model that will help predict whether a borrower will pay the loan or not.

Key issues

Ensure NPAs are lower – meaning PeerLoanKart wants to be very diligent in giving loans to borrower

Fields in Data Description:

• credit.policy: 1 if the customer meets the credit underwriting criteria of PeerLoanKart, and 0 otherwise
• purpose: The purpose of the loan (takes values "credit_card", "debt_consolidation", "educational", "major_purchase", "small_business", and "all_other")
• int.rate: The interest rate of the loan, as a proportion (a rate of 11% would be stored as 0.11). Borrowers judged by PeerLoanKart to be more risky are assigned higher interest rates
• installment: The monthly installments owed by the borrower if the loan is funded
• log.annual.inc: The natural log of the self-reported annual income of the borrower
• dti: The debt-to-income ratio of the borrower (amount of debt divided by annual income)
• fico: The FICO credit score of the borrower
• days.with.cr.line: The number of days the borrower has had a credit line
• revol.bal: The borrower's revolving balance (amount unpaid at the end of the credit card billing cycle)
• revol.util: The borrower's revolving line utilization rate (the amount of the credit line used relative to total credit available)
• inq.last.6mths: The borrower's number of inquiries by creditors in the last 6 months
• delinq.2yrs: The number of times the borrower had been 30+ days past due on a payment in the past 2 years
• pub.rec: The borrower's number of derogatory public records (bankruptcy filings, tax liens, or judgments)
• not.fully.paid: This is the output field. Please note that 1 means borrower is not going to pay the loan completely

Business benefits

Increase in profits up to 20% as NPA will be reduced due to loan disbursal for only good borrowers

Part 4(Extra optional)

Implement Linear Regression or Logistic Regression (your choice) models. Estimators required

- Model.predict
- Model.fit
- Model.score
- Model.coef_
- Model.intercept_