Our project was to do rule association mining on a dataset of traffic collisions in San Diego. Rule association mining involves translating the data into transactions, which are sets of unique occurrences for each event, mining which subsets of these sets are frequent itemsets, and then mining which frequent itemsets show high conditional probability for the occurrence of another set.
This type of machine learning is useful for exploring datasets for which patterns are not obvious, but can be discovered by the frequency of the items in the set.

____

This project uses a convolutional neural network in TensorFlow to perform image classification. Given an image of a cell, the model is able to determine, with roughly 96% accuracy, whether or not the cell is infected with malaria. The dataset consists of 27,558 images of cells split evenly between images of infected cells and images of uninfected cells. The original source of the dataset is from the NIH website at the following link: https://ceb.nlm.nih.gov/repositories/malaria-datasets/

___

This is an implementation of a spam filter utilizing scikit-learn. It's intended to identify spam messages in reviews and comments at my current job.

___

Primary Question: What is the relative risk of fatal crashes for different days of the year?
Secondary question: What is the relative risk of a fatal crashes for various risk factors, such day of the week and weather conditions.

I used raw data from the National Highway Traffic Safety Administration: ftp://ftp.nhtsa.dot.gov/fars/ Specifically, I gathered the Fatality Analysis Reporting System (FARS) for the 25 years from 1992 to 2016. https://www.nhtsa.gov/research-data/fatality-analysis-reporting-system-fars
I then used Spark for the first half and Pandas for the last part, to create visual representations of the relative risks.
I decided to use the average of a week before and a week after as my control. This approach has some limitations, since I am only considering two other points versus taking into account all the possible control points throughtout the year, which may give a more accurate result.

This project is working on movie rating prediction base on 9 features: genre(Genre), number of winning awards(win), number of nomination awards(nomination), is it a long move or not(time_indice), how long is this movie(time_long), county(Country), language(Language), released month(Month), how many people vote(imdbVotes). Variable names are in parentheses. Accordinhg to these 9 features run a linear regression model to predict ibdm user's movie rating. The r-square of this model is about 0.36 which shows that 36% of rating can be explained by selected data.

___

Objective: Classify skin lesions using a single convolutional neural networks (CNNs), trained end-to-end from images directly, using only pixels and disease labels as inputs.

___

This project was an initial foray into time series analysis and forecasting of time series data. A visual comparison of approximately four years of sales data containing multiple categories was carried out. Further analysis of furniture sales was then performed. Specifically, an Autoregressive Integrated Moving Average (ARIMA) model was developed and shown to not account for the seasonal nature of furniture sales. A Seasonal Autoregressive Integrated Moving Average (SARIMA) model was then developed and shown to better account for the seasonality of furniture sales.

___

Predicting Customer churn rate on Telco Customer churn dataset, taken from kaggale. We'll take a look at what types of customer data we have, do some preliminary analysis, and develop churn prediction models - all with Python/PySpark and different machine learning frameworks, like, ML Package and Scikit-learn.

The broad idea of this project is to develop machine learning models that could predict churn rate from given data. We will use different tools and methods like, numpy, pandas, seaborn, correlation, etc. to explore, extract and trasnform data.

___

This project predicts if a person in the dataset has diabetes or not based on their bmi , age, no of pregnancies they had, past medical history, insulin, glucose. Various machine learning models like logistic regression, decision trees, random forest, knn, svc, gaussian nb, gradient boosting, neural networks in python. Different models gave varying results on the test data. However, logistic regression , knn, decision trees in python seemed to perform better than other models to predict diabetes. Logistic regression in pyspark was attempted. It gave an accuracy of 84.2% on test data. The same spark code was tried on amazon aws. Results and cli export have been attached.

___

I have taken dataset from LendingClub (https://www.lendingclub.com/info/download-data.action). LendingClub is the world's largest peer-to-peer lending platform and it is headquartered in San Francisco, California. LendingClub is a marketplace for personal loans that matches borrowers looking for a loan to the lenders seeking to lend money and make a return.

Borrowers usually fills out an application with the reason for loan along with his past financial history. LendingClub evaluates each borrower's credibility based on his/her payment habits and assigns an interest rate to the borrower. Sometimes, the borrower may default the loans and as a result, there is a higher risk for lenders.

I have worked on the dataset for the year 2012-13. Based on borrower's financial history, I predited whether a borrower will payoff or default his loans. The dataset contains 188185 observations and 145 predictors. Loan status can be fully paid, charged off, current, in grace period or late (31-120 days). There is no enough data (0.0005%) on current, in grace period or late, so, I have ignored them and I did a binary classification on whether a borrower will be able to pay off his debt or not.

I have used both pandas and pyspark for this project. I used pandas for Data preprocessing, Exploratory data analysis, Correlation, Model Building. After doing all the cleanup, I saved the dataframe as csv file to be read from pyspark for model building.

The metrics I used to evaluate the model is classification accuracy. I downsampled the majority class to match minority class.
Accuracy of the model using pandas:

1. Random Forest - 81.4%
2. Logistic Regression - 78.7%
3. KNN classifier - 77.7%

4. Decision Tree Classifier - 82.7%

I used BinaryClassificationEvaluator to evaluate the models of pyspark, which uses areaUnderROC as the default metric.
For Test set, the AUC scores are as follows.
1. Logistic Regression model - 0.86
2. Random Forest Classifier - 0.86
3. Gradient Boosting - 0.91

This project predicts the success rate of a movie based on different attributes such as the actors, directors, year in which the movie was released, genre, total runtime, rating, number of votes, total revenue generated, metascore, age of the users watching and recording the votes or rating, the geographical areas where movie was released, any other influences eg: ongoing trends, etc.
The dataset is taken from the Kaggle dataset : https://www.kaggle.com/PromptCloudHQ/imdb-data

---

The project tries to build a regression model to predict purchase amount for each customer.The training dataset has the purchase amount for a retail store for selected high volume products.
The store wants to predict how much a customer might spend on a particular high volume product given their spending history and data about the customers like age, gender, occupation etc.,
To achieve this ,the project uses Decision Tree and Random Forest regressors.
The sample data is in the data folder.The files are outputted to the data folder.

___

Analyze the effect Elon Musk's tweets on Tesla stock price.

- Using data set available from kaggle
  - Elon Musk tweets from 2012 to 2017 https://www.kaggle.com/kulgen/elon-musks-tweets
    - This contains all tweets made by @elonmusk, his official Twitter handle, between 11/16/2012 and 09/29/2017.
  - Tesla stock price data https://www.kaggle.com/rpaguirre/tesla-stock-price
    - 06/29/2010 to 03/17/2017.