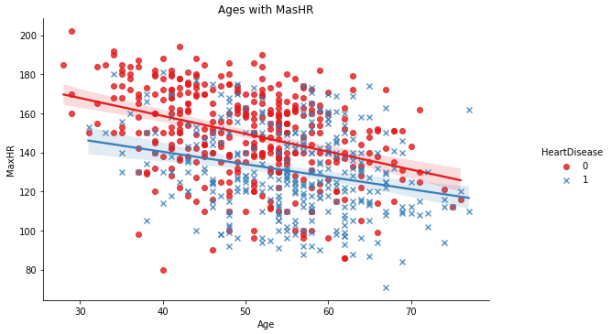
Mireille Feudjio Data Science Portfolio

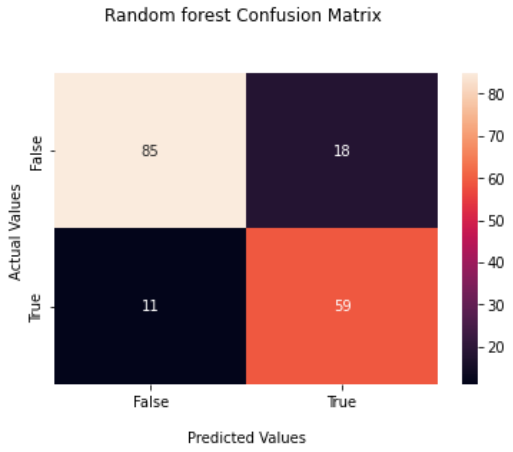
Data Scientist | Machine Learning | Managing Data for Better Life

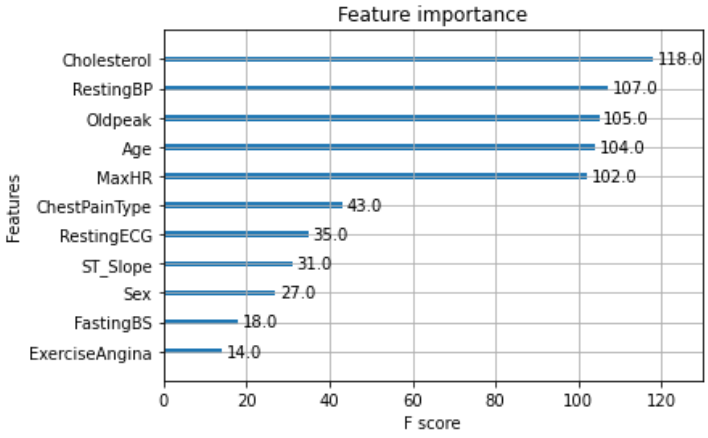
Classification

Heart Disease: What is my status?

Performed multiple classification algorithms using grid search and random search for hyperparameter tuning. 3 models were developed (KNN, Random Forest, XGBOOST).







Keys Output

3 models were developed (KNN, Random Forest, XGBOOST). Form lab exam of the patient, all the models predicted that, the patient has no risk of heart disease. - We cared about correct prediction, and we have an unbalanced dataset, so f1_score helped to select the best model - Random Forest.

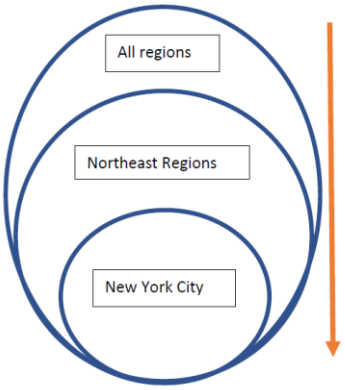
Person with heart disease is more likely to be:

- Person with normal resting electrocardiogram.
- Person who exercise-induced angina;
- Person with Asymptomatic chest pain type;
- Person with fasting blood sugar > 120 mg/dl (1) ;
- Female with heart disease tend to have Hight cholesterol than male.
- mean value of aged person with heart disease is 55.58.
- mean value of resting blood pressure of person with heart disease is 134.18.
- mean value of Max heart rate of person with heart disease is 127.65.

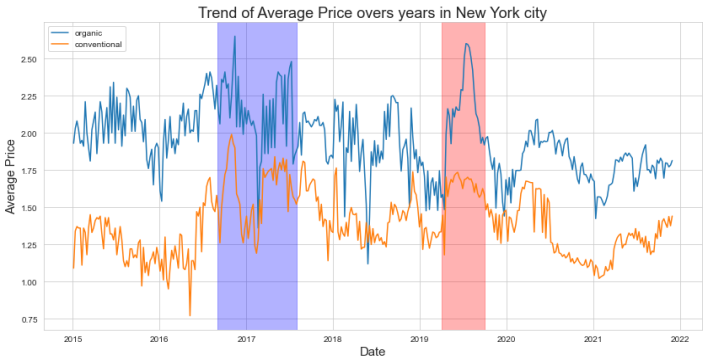
Time Serie analysis

Avocado: what is the expected price and consumption of avocado in 2022?

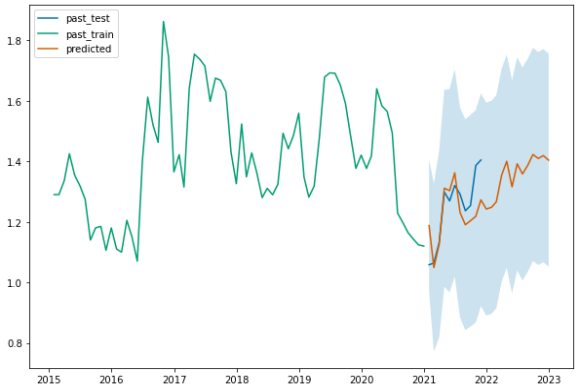
Avocado price and total volume were forecasted with SARIMAX, ARIMA algorithms. The mean absolute error (MAE) metric was used to evaluate and select models. MAE was 0.063 for the best model.



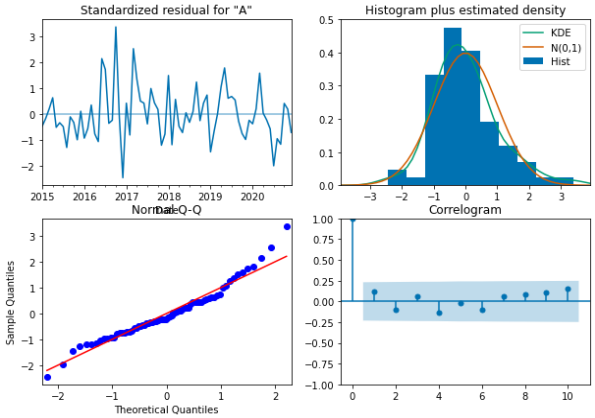
Conceptual framework for EDA



Weekly trend of avocado prices



Plot of the past, the test and prediction price of Avocado



Model Evaluation

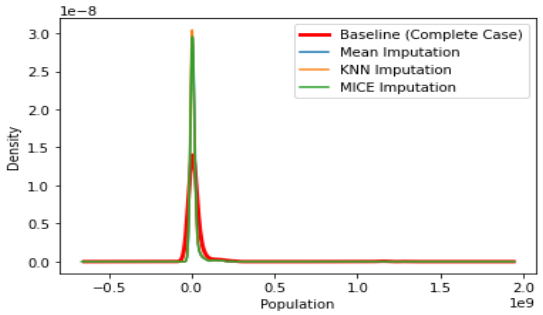
Keys Output

- Clear distinction between organic and conventional avocados.
- The price of the organic avocado is higher than that of the conventional avocado.
- The total volume of organic avocados sold is less than that of conventional avocados.
- New York is among the top 5 regions with high price (and high volume) in organic and conventional avocado (occupied 4th or the 5th rang).
- When total volume increases, price decreases and vice versa.
- 8 models have been developed for each type of avocado (conventional price, conventional volume, organic price, organic volume).
- The predicted prices of avocados (conventional and organic) are close to real values.
- The average of the difference between this value (pred – actual) MAE is 0.063.

Regression

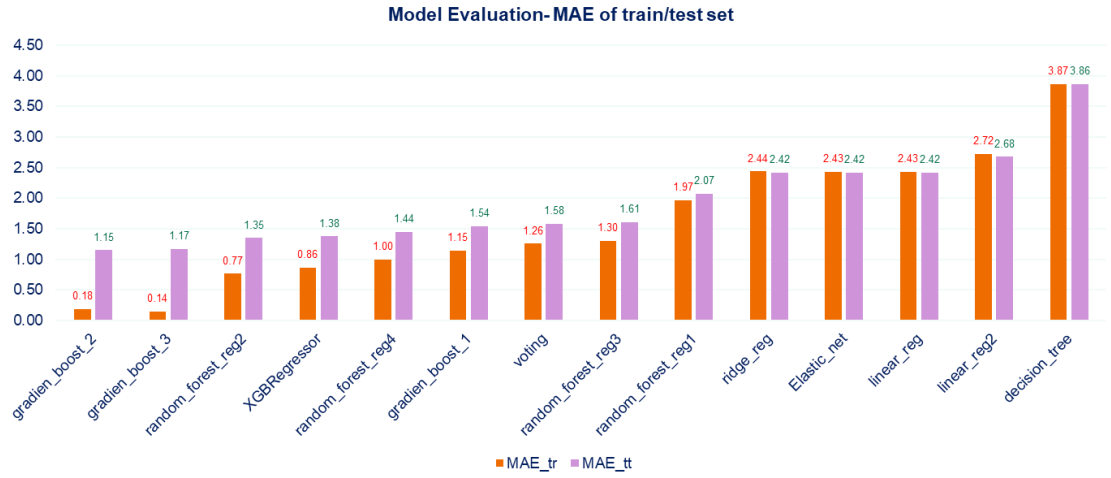
Life expectancy. Does immunization matter?

14 regression models have been developed to predict life expectancy, the chosen one is Gradient boost with MAE of 0.202 on train set and 1.431 on the test set, R square is 0.94 on the test set.



Assessment and treatment of missing value (if not, 69% of data will be drop).

Check the normal distribution of the data



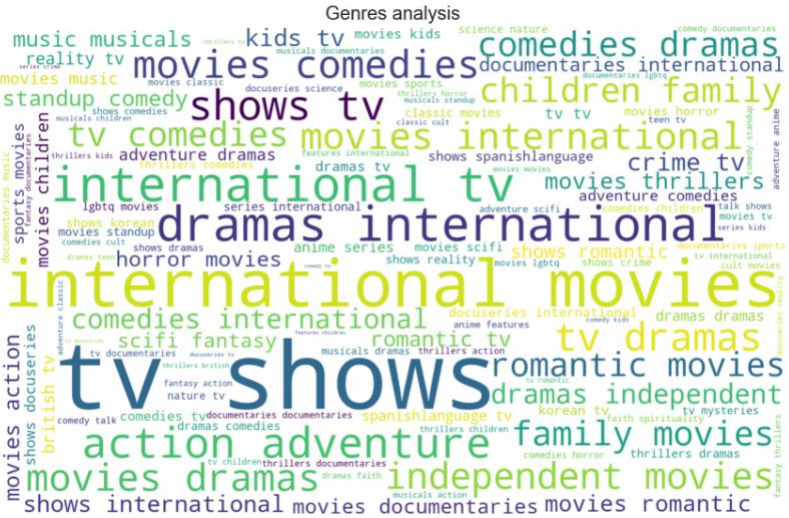
MAE of tree based model and linear model

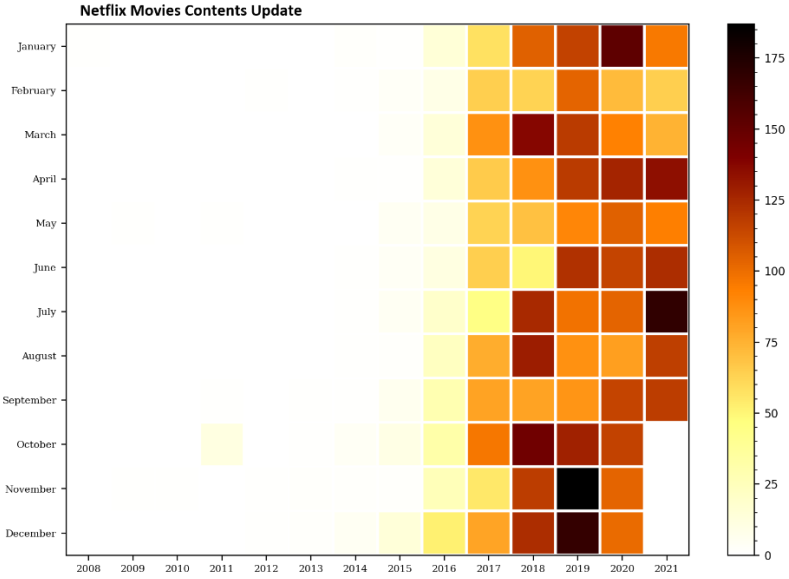
Keys Output

- Life expectancy has increased over years.
- The mean average of the life expectancy of developed countries is generally higher compared to that of developing countries
- However, the ratio of LE over the decade of 2005 to 2015 showed that life expectancy in developing countries has greatly increased
- Immunization has impacted the improvement of life expectancy and the reduction of infant deaths.
- Feature of importance reveals that immunization features has a very low contribution in the model .
- Economic factors and mortality factors play an important role in the system.
- (14) regression models have been developed to predict life expectancy.
- the chosen one is Gradient boost with MAE of 0.18 on train set and 1.15 on the test set.

Natural Language processing Playground

This project aims at exploring different concepts of NLP with a movie_ tv_show dataset. We explored issues like the most important genres, the spread of added movies and tv_shows across years and months, recommendation, and apply classification algorithms on the text description dataset. [Notebook]



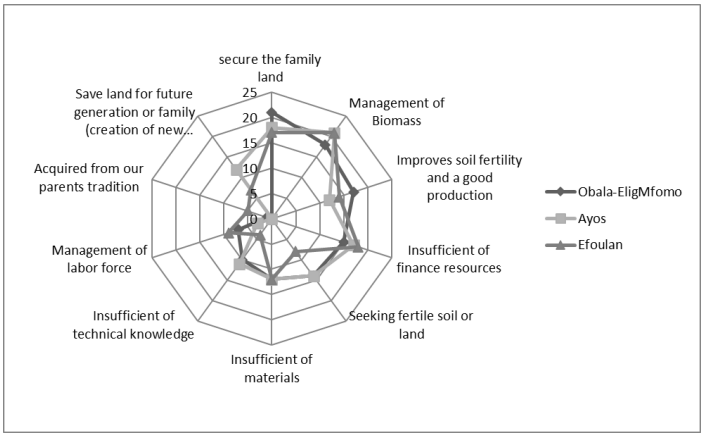


The notebook investigated concepts like: removing punctuation, tokenize, lemmatize, stemming, stop words, Create dictionary of the most frequent words and Word Cloud, Gensim and spacy libraries, apply Recommender system, use of Bag of words (unigram, bigram, …), exploring different models (random forest, gradient boost, Naïve Bayes) with vectorized data made with count vectorizer, TFIDF, analyzing similarity between reviews description and Classification with kmean clustering with the use of Elbow method.

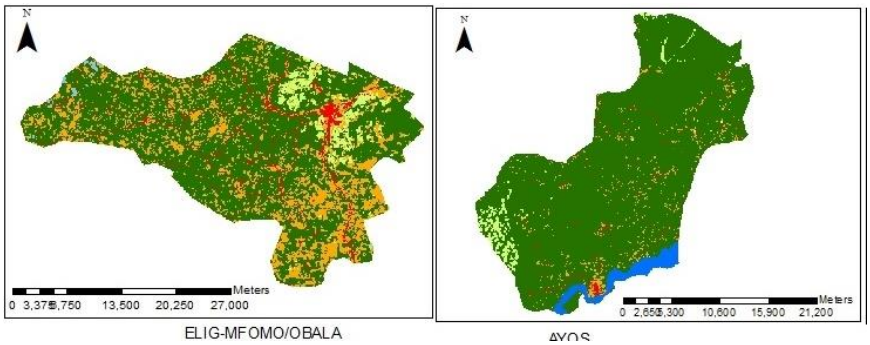
Research (statistic, GIS)

Drivers and levers options of shifting cultivation project

Define and carry on the project to identify drivers motivating farmers to practice shifting cultivation. Developed data collection tools, gathered data with CSPro, analyzed data with R, SPSS, and Excel. Analyzed and mapped the different trajectories of shifting cultivation practice across three landscapes with QGIS, and ArcGIS.



Weight of technical drivers across 3 cities



ELIG-MFOMO/OBALA AYOS

Land cover change and land use change

Other projects on Notebook

SQL, Sqlite3, Unsupervised_learning, A/B testing, Classification_project,