Data Science Portfolio of Projects

Dypsyl Dlamini   
B.Sc.(Eng), M.Sc.(Eng), M.Sc. (Big Data), MBA

1. BUSINESS MANAGEMENT
   1. Predicting Employee Attrition: Using HR data, this project aims to build a model to predict employee attrition.

The dataset is available on Kaggle:  
<https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset>

1.2 Bank Customer Churn: Project from Upwork – devise a predictive model to tell which customers are likely to leave. The dataset was presented with the project.

1. CRIME
   1. Analysing Crime Data: This project involves analysing crime data to identify patterns and trends.

The dataset is the UK Crime dataset available on data.police.uk:  
<https://data.police.uk/data/>

* 1. Detecting Credit Card Fraud: This project involves detecting fraudulent credit card transactions using a dataset that includes anonymized credit card transactions made over a two-day period in September 2013 by European cardholders.

The dataset is available on Kaggle:   
<https://www.kaggle.com/mlg-ulb/creditcardfraud>

* 1. Exploring Gun Violence in the United States: This project involves analysing and visualizing gun violence incidents in the US using a dataset that includes information on gun deaths and injuries, mass shootings, and gun laws from 2013 to 2018.

The dataset is available on Kaggle:   
<https://www.kaggle.com/jameslko/gun-violence-data>

* 1. Uniform Crime Reporting Program Data: The objective of this project is to analyse crime data and identify patterns in criminal activity using a dataset that contains information on various offenses known and cleared by arrest. The dataset includes information such as the type of offense, the number of incidents, the location of the incident, and the clearance status of the offense.

The dataset is available on the National Archive of Criminal Justice Data website:  
<https://www.icpsr.umich.edu/icpsrweb/NACJD/NIBRS>

1. *COMMERCE*
   1. Analysing Netflix Movie Ratings: This project involves analysing and visualizing the movie ratings and reviews on Netflix using a dataset that includes information on movies, TV shows, and documentaries available on Netflix as of 2019.

The dataset is available on Kaggle:   
<https://www.kaggle.com/shivamb/netflix-shows>

* 1. Analysing Stock Prices: The objective of this project is to analyse and visualize the historical stock prices of various companies and identify patterns or trends using a dataset available on Yahoo Finance. The dataset contains information such as the opening price, closing price, volume, and more for various companies.

The dataset is publicly available on the Yahoo Finance website:  
<https://finance.yahoo.com/>

* 1. Customer Segmentation: The goal of this project is to segment customers into different groups based on their purchasing behaviour using data from an online retailer.

The dataset is available on the UCI Machine Learning Repository:  
<https://archive.ics.uci.edu/ml/datasets/online+retail>

* 1. Predicting Customer Churn: The objective of this project is to predict customer churn in a telecommunications company using a dataset which includes features such as customer demographics, services used, contract information, billing information, and more.

The Telco Customer Churn dataset is available on Kaggle:  
<https://www.kaggle.com/blastchar/telco-customer-churn>

* 1. Predicting House Prices: The goal of this project is to build a predictive model for house prices using housing data and features such as the number of rooms, location, and age of the house.

The dataset is available on Kaggle:  
<https://www.kaggle.com/harlfoxem/housesalesprediction>

* 1. Sentiment Analysis of Product Reviews: This project involves analysing product reviews and classifying them as positive or negative using natural language processing techniques.

The dataset is available on the Yelp Dataset Challenge website:  
<https://www.yelp.com/dataset_challenge/dataset>

1. HEALTH SCIENCES
   1. Predicting Air Quality: The objective of this project is to predict air quality based on various features using a dataset which includes features such as carbon monoxide, nitrogen oxides, benzene, temperature, humidity, and more.

The dataset is available on the UCI Machine Learning Repository:  
<https://archive.ics.uci.edu/ml/datasets/Air+Quality>

* 1. Predicting Air Quality in Beijing:   
     This project involves predicting the PM2.5 levels in Beijing using a dataset that includes hourly air pollutants data and meteorological data from 2010 to 2015.

The dataset is available on the UCI Machine Learning Repository: <https://archive.ics.uci.edu/ml/datasets/Beijing+PM2.5+Data>

* 1. Predicting Diabetes Onset:   
     This project involves predicting the onset of diabetes using a dataset that includes diagnostic measures for diabetes, such as glucose and insulin levels, and other health-related variables for Pima Indian women.

The dataset is available on the UCI Machine Learning Repository: <https://archive.ics.uci.edu/ml/datasets/pima+indians+diabetes>

1. SOCIAL SCIENCE
   1. Analysing social media sentiment: In this project, the goal is to analyse the sentiment of social media posts using the Sentiment140 dataset, which contains over 1.6 million tweets with sentiment labels (positive or negative).

The dataset is available on Sentiment140:  
<http://help.sentiment140.com/for-students/>

* 1. Whithorn Trust: This was a commission in which the Whithorn Trust had 2 databases which they required us to convert into a single table, preparatory to uploading this to a 3rd Party application.