Portfolio of Learning

Personal Statement and Reflection

Harnessing the power of numbers as a Chartered Accountant (ACCA), my personal goals are aimed at unlocking the mysteries wrapped in data through the lens of data science and machine learning. With my core strengths in accounting, finance, and mathematics, coupled with an interest in coding, my adventure into the world of data science, has been both enlightening and transformative.

My affinity for integrating software, data, and analytics has evolved into a particular interest in data science and machine learning principles. Through the course of the Post Graduate Diploma in Data Science, exploring health analytics emerged as a new field, expanding my academic and professional horizons. Prior to my academic endeavors in data science, my insight into health analytics was notably sparse, confined to a minimal understanding of public healthcare in South Africa.
This limited perception was accumulated predominantly through family connections, with relatives serving in pivotal roles as Director Generals in the public healthcare sector. Consequently, my knowledge was rooted more in administrative and policy-oriented aspects, rather than the analytical and data-driven dimensions now being explored through my current studies.

By delving deeper into basic concepts of health analytics, I explored and understood the complexities of supervised and unsupervised machine learning. Supervised learning, in which algorithms are trained using labeled data, and unsupervised learning, in which algorithms explore unlabeled data to discover hidden patterns. These concepts have paved the way for valuable insights, including predicting and understanding patient outcomes, optimizing operational efficiency, and uncovering risk potential elements within health datasets.

Defining Health Analytics

Embarking on a deep dive into the realm of health analytics through my Post Graduate Diploma in Data Science, my understanding of this field has been holistically shaped by a series of assignments that intertwined the theoretical knowledge with hands-on application.

Initially, defining the foundational building blocks of a Health System through visual representation on a Miro Board, followed by crafting an essay to explore Contemporary Health Systems Challenges, set the initial stage of understanding the current healthcare setting and its inherent complexities. This was effectively contrasted with practical tasks, such as employing Python libraries like pandas and numpy to dissect datasets and leveraging Matplotlib for data visualization, which not only amplified my technical skillset but also enabled me to draw tangible correlations between data and healthcare quality improvement.

The courses also helped me understand the important role of data science in health systems science, through identifying and establishing the necessary skills of a health analyst, whilst concurrently illuminating my growth areas through practical application.

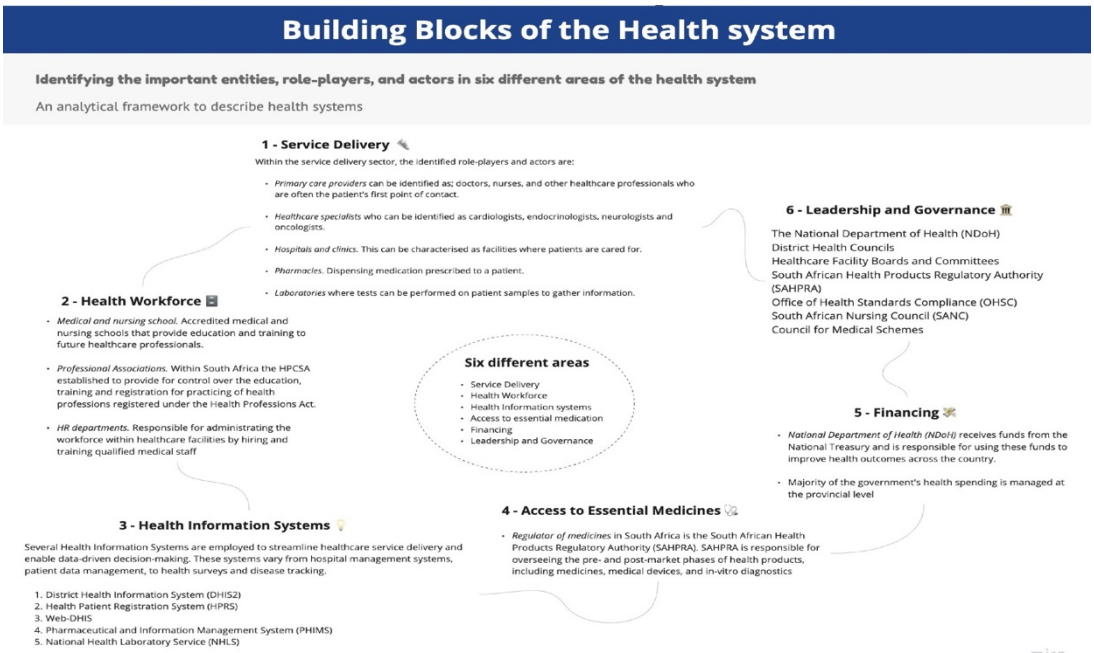
Leveraging different tools and technologies, such as cloud computing and markup, I deciphered the dominant technology groups in data analytics and contextualized the current state of data analytics in the medical field. As I navigate the various types of medical data sources and meticulously analyzed electronic health records, I learnt to place a significant emphasis on evaluating, preparing, and managing data quality, including developing strategic methods to eliminate null values and prepares data for multifaceted analyses.

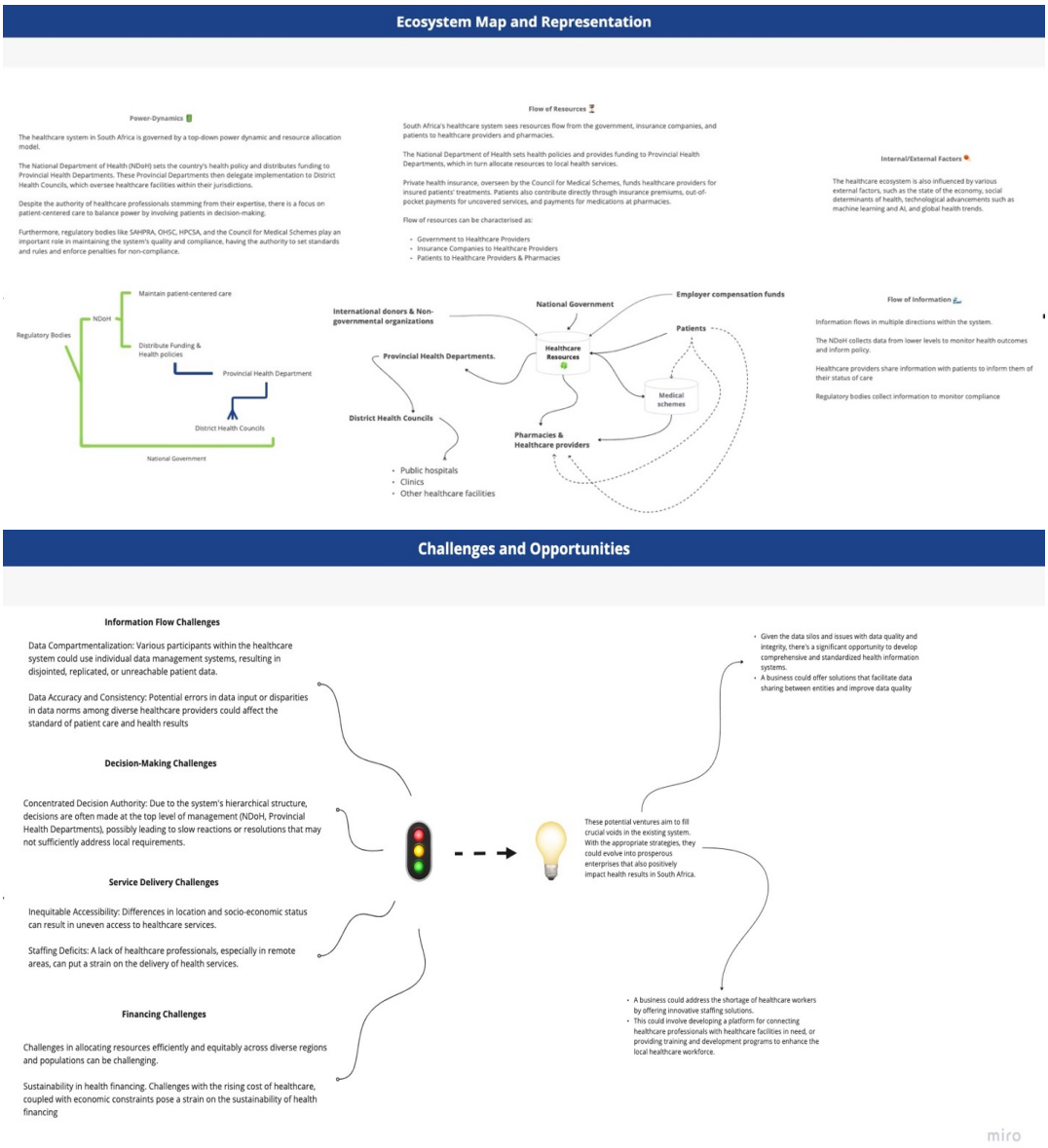
My journey also delved into the technicalities of managing dataset dimensions, where I applied Principal Component Analyses (PCA) and clustering algorithms such as K-means to principal components from datasets, consequently enhancing my practical understanding of both supervised and unsupervised learning methodologies.

The program naturally evolved to address critical aspects like data security, ensuring that my insights into health analytics are not only rooted in optimizing data utility but also safeguarding sensitive information, thereby synthesizing technical know-how with ethical practice. Consequently, this multifaceted approach has scaffolded my transition from a nascent understanding of health analytics to developing a robust, hands-on expertise that cohesively intertwines technical proficiency with a comprehensive understanding of healthcare systems and challenges.

Health Ecosystem Mapping

Link to Miro board - https://miro.com/app/board/uXjVMySc-wM=/?share_link_id=657743105618





Addressing Contemporary Health Challenges: The Role of Data Science

This argumentative essay seeks to undertake a comprehensive review of a contemporary health challenges, within the South African context. The main objective is to explore the potential applications of data science to solve this problem and to identify the relevant data sources that can contribute to the design of an effective solution.

In South Africa's current health landscape, HIV/AIDS emerges as a predominant concern and challenge. South Africa ranks among the world's leaders in HIV prevalence. While commendable progress has been achieved in broadening the accessibility to antiretroviral therapies several barriers remain, including sustained treatment adherence, mother-to-child transmission of HIV and co-infection of TB and HIV is a significant challenge in South Africa. The projected HIV prevalence rate within the South African demographic stands at roughly 13.7%. As of 2021, the total count of individuals diagnosed with HIV in the country is approximated to be around

8.2 million. Notably, within the age bracket of 15–49 years, it is estimated that 19.5% of the population is afflicted with the virus.

The subsequent points will list various facets pertaining to the analysis of data challenges and prospects linked to HIV/AIDS in the South African context.
With the advent of big data, the vastness of available information poses challenges, not only overwhelming storage but also taxing processing capacities. A fundamental step towards comprehending the HIV/AIDS landscape within communities is the acquisition of trustworthy data. Nonetheless, the large volume of data can compromise its integrity and precision. Such inconsistencies or inaccuracies in the data have the potential to exponentially skew analyses, resulting in decisions that will not be optimal. Additionally, the surge in expansive datasets can culminate in the emergence of 'Data Silos'. In such scenarios, data becomes segregated within distinct databases or organizational departments, subsequently obstructing a holistic analysis.

Although real-time processing is imperative to harness the full potential of big data in healthcare infrastructures, its tangible deployment will face considerable technical challenges. A significant challenge, particularly in township-based clinics, is the widespread paper-based management system for HIV/AIDS patients. Each patient's record is maintained in individual folders, which are updated manually during visits by administrative personnel or nursing staff. Owing to the heavily manual nature of the established management system, it is intrinsically vulnerable to human error, including the potential misplacement or loss of patient records.

There exists a significant array of data-related challenges; nonetheless, accompanying opportunities suggest a promising trajectory for the integration of data within South Africa's healthcare system, specifically targeting the prevalence of HIV/AIDS.
One potential avenue of opportunity lies in data personalization, allowing healthcare services to be tailored to individuals or defined groups of patients. Furthermore, predictive analytics is emerging as a promising tool in addressing the HIV/AIDS health challenges. Through the analysis of prevailing trends and patterns, collaborative entities can forecast future efficacious treatments or anticipate potential HIV/AIDS outbreaks in communities at risk of virus resurgence. Simultaneously, opportunities arise to boost operational efficiency and advocate for open data initiatives within the public health sphere. Through a meticulous assessment of healthcare data, administrative agents of public and private healthcare can identify sectors of suboptimal performance, facilitating enhancements in patient engagement and streamlining of procedural protocols.

Utilizing data models, algorithms, and sophisticated analytical techniques emerges as a potent strategy to confront health adversities such as HIV/AIDS within the South African context. By using personalized strategies, the healthcare infrastructure can more adeptly address the distinct requirements and circumstances of both individuals and broader communities, potentially enhancing prevention, treatment, and care efficacy.

User profiles may embody comprehensive patient records, chronicling their medical history, socio-economic variables, and patterns of treatment adherence. Such profiles can facilitate the crafting of individualized treatment or preventive measures. Concurrently, within the realm of data science, these user profile datasets can serve to separate how specific demographics or communities react to treatments or interventions, enabling predictions regarding the potential benefits for equivalent sub-populations.

Algorithms and intricate analytical techniques can encompass methods of filtering. This involves identifying certain demographic sectors or communities that exhibit favourable outcomes from designated awareness campaign protocols. Consequently, equivalent strategies can be advised for groups with similar characteristics. Such filtering facilitates the tailoring of awareness campaigns or preventive measures to individuals, contingent on their risk profiles. The technique of Association Rule Mining can be employed to discern co-morbidities or social behaviours frequently linked with HIV infection, thereby enabling a more comprehensive approach to treatment and prevention. Time Series Analysis offers the capability to monitor the evolution of the HIV/AIDS epidemic chronologically, forecasting impending trends which are instrumental in resource distribution and strategic foresight. Concurrently, via time series evaluations, clustering techniques can be deployed to detect clusters of HIV/AIDS instances, thereby identifying high-vulnerability regions or communities, and optimizing prevention initiatives.

Through integrating these data-driven methods with on-the-ground healthcare initiatives can greatly enhance the effectiveness of strategies aimed at combating HIV/AIDS in South Africa. Given the major health challenge of HIV/AIDS in South Africa, the integration of contemporary data models, algorithms, and analytical methods offers a promising frontier. By linking technology and healthcare, South Africa stands to make significant strides in the ongoing battle against HIV/AIDS.

References:

Statistics South Africa. (2021). STATISTICAL RELEASE P0302. Mid-year population estimates. Available at: https://www.statssa.gov.za/publications/P0302/P03022021.pdf (Accessed: 9 August 2023)

Busgeeth, K., Rivett, U. *The use of a spatial information system in the management of HIV/AIDS in South Africa*. Int J Health Geogr 3, 13 (2004).

Akintola, O. (2008). *Defying all odds: coping with the challenges of volunteer caregiving for patients with AIDS in South Africa*. Available at: https://onlinelibrary.wiley.com/doi/full/10.1111/j.1365-2648.2008.04704.x?casa_token=L8QStuqQpLsAAAAA%3AEMI2QvvnXK-8UbP5K79QEmUBTLtBksmAN0cDGA1cJitggnhrMNPp8L1_n4gL6vwAHUkGeD7Z6sv73WA9 (Accessed: 9 August 2023)

Olatosi B, Zhang J, Weissman S, et al. (July 19, 2019). *Using big data analytics to improve HIV medical care utilization in South Carolina: A study protocol* BMJ Open 2019. Available at: https://bmjopen.bmj.com/content/9/7/e027688 (Accessed: 9 August 2023)

W. I. Yudhistyra, E. M. Risal, I.- soon Raungratanaamporn, and V. Ratanavaraha, (Jun. 2020). "*Using Big Data Analytics for Decision Making: Analysing Customer Behaviour using Association Rule Mining in a Gold, Silver, and Precious Metal Trading Company in Indonesia*", pp. 57-71. Available at: https://ijods.org/index.php/ds/article/view/17 (Accessed: 9 August 2023)

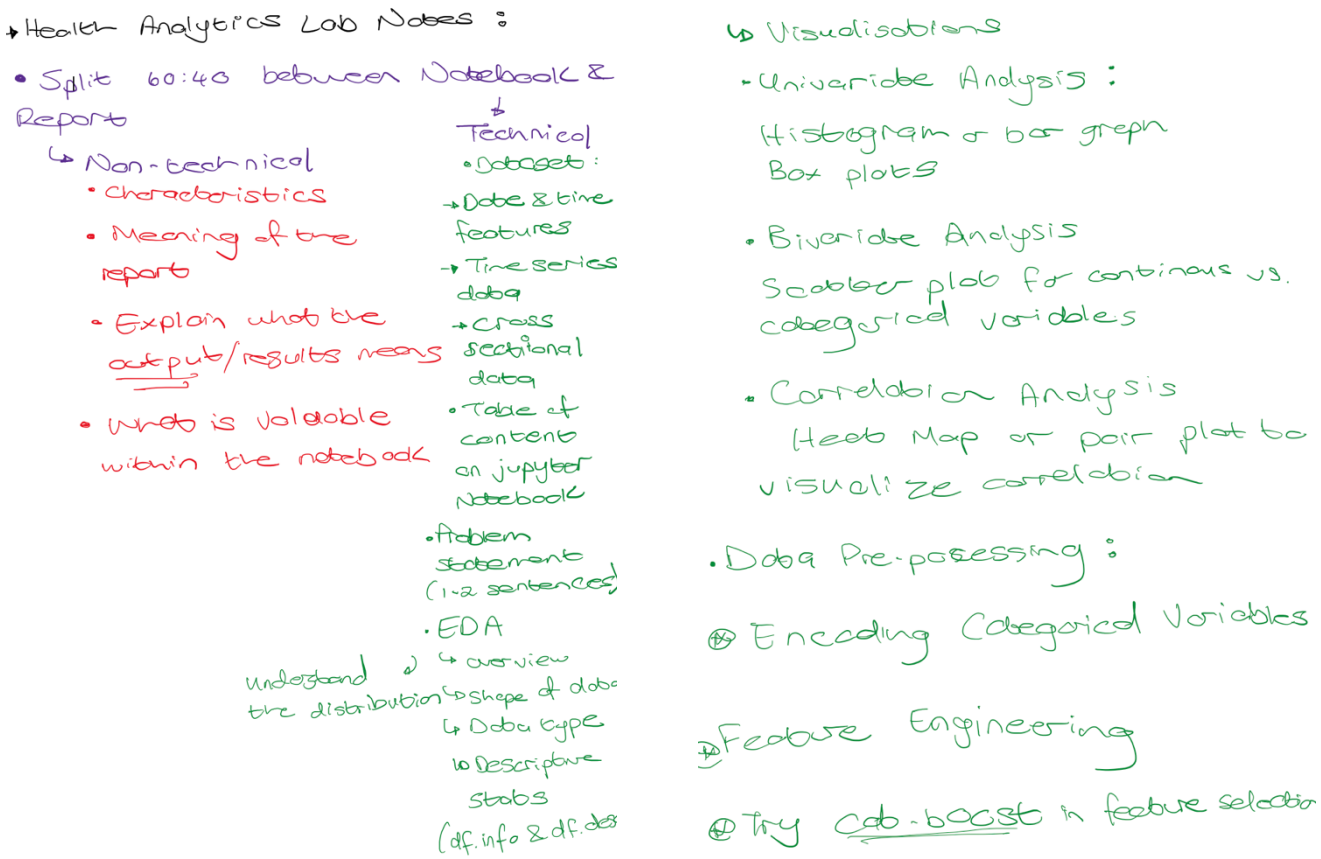
Lab 1 Reflection

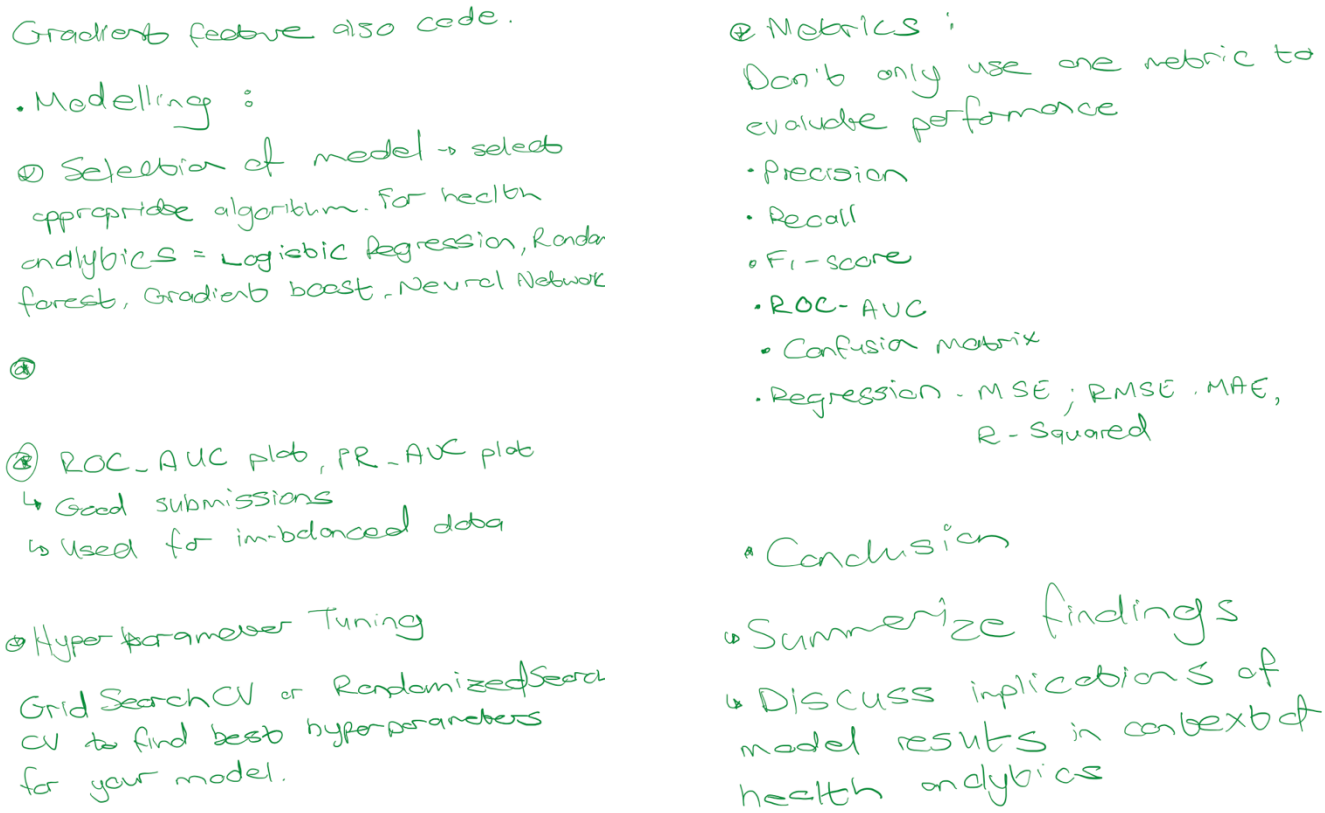
Embarking upon an assignment that melded theoretical knowledge with applied skills, for this group task we delved deeply into the nuances of linear regression modelling, utilizing a carefully selected dataset as the bedrock for our exploration.

The initial phase demanded meticulous analysis of the dataset, wherein we dedicated substantial efforts towards understanding its features, distribution, and potential predictors. Upon selecting a relevant dataset, the journey ventured into practical implementation, utilizing Python libraries such as numpy, matplotlib, and sklearn to materialize a linear regression model. Here, we were able to put theory into practice, initializing the model, fitting it with data, and ultimately observing the resultant predictions and residuals.

This practical aspect not only solidified my understanding of the algorithm's workings but also provided invaluable hands-on experience with coding and model implementation. Furthermore, throughout the implementation, each decision-making juncture—from selecting independent and dependent variables, pre-processing data, to tuning the model—was anchored by thorough justification and clear delineation of the process followed. This required weaving theoretical knowledge of linear regression, gleaned from course materials, with the practical challenges and considerations encountered during implementation.

Online Revision Session for Lab 1





Link to online repository – Linear Regression:
https://github.com/christophertheys/LinearRegression

Types of Data Reflection – Exploring Data

Disease Classification - Algorithms can be trained to classify diseases like tumours as malignant or benign. This uses categorical data as the outcome is typically a distinct category

Prediction of Patient Readmission Rates - Hospitals might use patient data to predict the likelihood of a patient being readmitted within a certain time frame. This would be categorical data

Blood Sugar Level - Machine learning models can predict future blood sugar levels. This problem leverages continuous data as blood sugar levels are measured on a continuous scale.

Estimation of Drug Response - Analysing patient genetics and other biomarkers, machine learning models can be designed to predict how a patient will respond to a particular medication. This would be classified as categorical data as either a positive response, neutral response, or a negative response.

Progression Tracking of Neurological Diseases – Diseases such as Parkinson's or Alzheimer's, machine learning models can be utilised to analyse patient data and predict the rate of disease progression. This uses continuous data, because the disease progression will be measured on a continuous basis

Lab 2 Reflection

Participating in Lab 2, the objective of the task was undertaking a process of building a complex logistic regression model, starting with the important step of feature selection. Together, we used different feature selection techniques, such as recursive feature elimination and using feature importance scores, with the aim of distilling the most relevant predictors for our model.

Navigating the group's tasks proved to be a difficult task as we faced significant obstacles in our collective working dynamics, especially in the area of communication. The lack of consistent and clear communication between us often disrupts our workflow, creating not only procedural obstacles but also affecting the synergy and efficiency of our efforts. our collaborative efforts. This common problem has highlighted the urgent need to establish strong communication channels and standards in future collaborative projects.

Transitioning to model implementation, utilizing numpy, matplotlib, and sklearn libraries allowed us to not only apply theoretical knowledge practically but also deepen our collective understanding of logistic regression's functional form and optimization procedures.

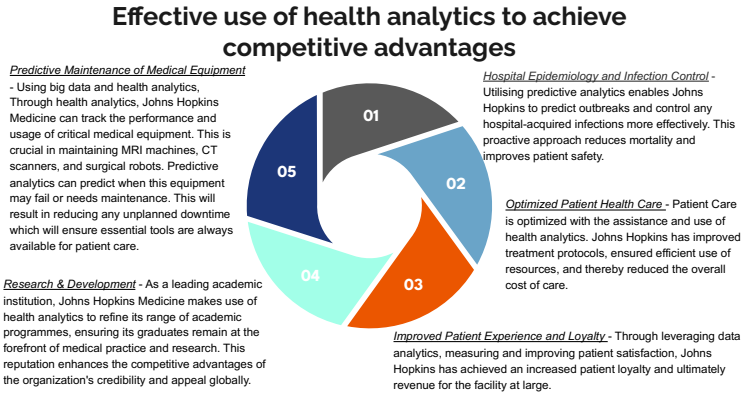
The subsequent model optimization posed a challenging yet enlightening experience, wherein we navigated through hyper-parameter tuning and regularisation techniques. While being mindful of the dataset's characteristics and the peril of overfitting, our group strategically employed techniques like GridSearchCV for hyperparameter tuning and implemented L1 and L2 regularization, ensuring our model was both accurate and generalizable to unseen data.

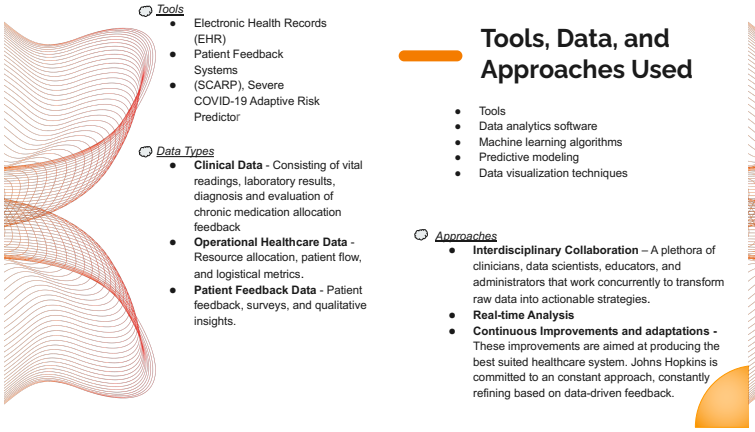
Link to online repository – Logistic Regression:

https://github.com/christophertheys/LogisticalRegression

Activity 2

Using Health Analytics in a Healthcare Setting





Reflection on Activity

I focused my lens on Johns Hopkins Hospital, which has manifested tangible competitive and life-saving advantages through its strategic deployment of health analytics. Sourcing information through a myriad of academic articles, industry reports, and case studies, my exploration uncovered how health analytics has been vital, not only in shaping superior health outcomes but also in driving operational excellence within such establishments.

Whilst crafting a presentation highlighting Johns Hopkins Hospital, I gained intricate insights into the multifaceted ways through which the organization adeptly utilizes health analytics to achieve varied outcomes. This includes amplification of patient care, optimization of operations, cost reduction, and the preservation of lives. A particularly remarkable discovery was the extensive and profound array of data sources exploited, ranging from Electronic Health Records (EHRs) to patient feedback surveys and numerous additional data points, all of which were thoroughly analysed to extract meaningful insights and perpetuate an ethos of continuous enhancement.

Navigating through the specific tools and approaches employed, I uncovered a panorama of analytics initiatives, ranging from predictive analytics for patient outcomes, operational analytics for resource optimization, to prescriptive analytics for enhanced decision-making across the healthcare value chain. Johns Hopkins Hospital's endeavours to intricately intertwine data and analytical tools, like machine learning algorithms and data visualization, into their operational and clinical frameworks. This offered a profound insight into the impact data can have on healthcare delivery and management.

Some challenges were also identified that the healthcare organizations encounter in their health analytics journeys. Data privacy, exchange of data, and management emerged as prevailing hurdles. Johns Hopkins Hospital, in this context, illuminated how these challenges can be adeptly managed through robust data governance, stringent security protocols, and fostering a culture that seamlessly blends clinical expertise with analytical acumen.

Reflecting upon this exploration, the lessons extend beyond the technical and strategic facets of health analytics. This task emphasized the vital role that data and analytics play in enhancing healthcare outcomes and operational efficacy. Furthermore, it has instilled a deeper appreciation for the complexity and multifaceted nature of implementing health analytics within organizations, thereby shaping a well-rounded perspective that connects technical, strategic, and ethical considerations in the realm of health analytics.

Lab 3

Breast Cancer Prediction Using Decision Tree

Background on breast cancer

Breast cancer is one of the most common cancers worldwide, predominantly affecting women, although men can also be diagnosed with this type of cancer. It arises from the cells of the breast, often from the inner lining of milk ducts or the lobules that supply these ducts with milk. The World Health Organization reports that breast cancer impacts over 2.1 million women each year, marking it as a major global health concern.

In this comprehensive report, we delineate the findings derived from employing a decision tree algorithm in the predictive analysis of breast cancer incidences. Through meticulous data preparation and model optimization, we strive to leverage the decision tree algorithm's capabilities to foster substantial advancements in predictive diagnostics, thereby demonstrating the potential of this model as a pivotal tool in mitigating the global health challenge posed by breast cancer.

The Data

We used a dataset found at which contained 569 rows and 32 columns containing information from breast mass biopsy samples. The dataset also has the target variable of diagnosis, which we are trying to predict whether the observation is benign or malignant based on the other attributes.

For this study, we utilized a dataset retrieved from, https://data.world/health/breastcancer- Wisconsin, which encompasses 569 instances and 32 attributes derived from breast mass biopsy samples. Each instance in this dataset represents critical data gathered from individual biopsy samples, with a range of attributes that facilitate a detailed analysis aimed at the predictive diagnosis of breast cancer.

The pivotal component of this dataset is the target variable denominated as 'diagnosis', establishing the binary classification that we aim to predict; benign or malignant.

The Process

Suite of Libraries

NumPy: Engaged for efficient handling of arrays and matrices, which is essential in managing our dataset efficiently.

Pandas: Utilized for data manipulation and analysis, aiding in cleaning, and organizing our dataset for optimal performance.

Matplotlib: Employed for crafting visualizations to better understand our data and to create graphical representations of our findings.

Seaborn: Leveraged in conjunction with Matplotlib to enhance the visualization of data through the creation of aesthetically pleasing and informative statistical graphics.

Scikit-learn (sklearn): This library played a pivotal role, offering a range of tools used throughout the process. Splitting our data into training and test sets, a critical step in training our decision tree model.
Within this library:

- Modules aided in feature selection, helping identify the most relevant attributes for our predictive model.
- Served as the foundation of our decision tree model, facilitating the training, and testing process.

- Facilitated the hyperparameter tuning process, helping optimize the model for better performance.

Preliminary Data Analysis

We ensured that the dataset doesn't have any duplicate observations and that all the required attributes are available.

A detailed error message that specifies which columns are missing is generated using a formatted string that joins the names of the missing columns with a comma. This approach ensures that the script will halt execution and alert the user to the specific issue, thereby preventing silent failures and facilitating debugging.

Exploring the Data

To get a better understanding of the dataset we examined the data visually and statistically to uncover patterns, relationships, and potential outliers.

Preparing the Data

The dataset underwent pre-processing which included steps such as duplication, the elimination of irrelevant columns, and the creation of a numerical depiction through correlation analysis. Additionally, the relations between different features were visualized, and verification was carried out to confirm the accurate representation of various classes in the 'diagnosis' column.

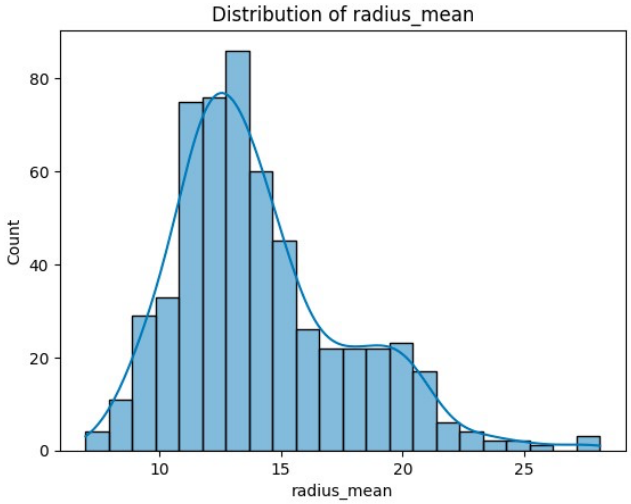
Feature Selection and Analysis

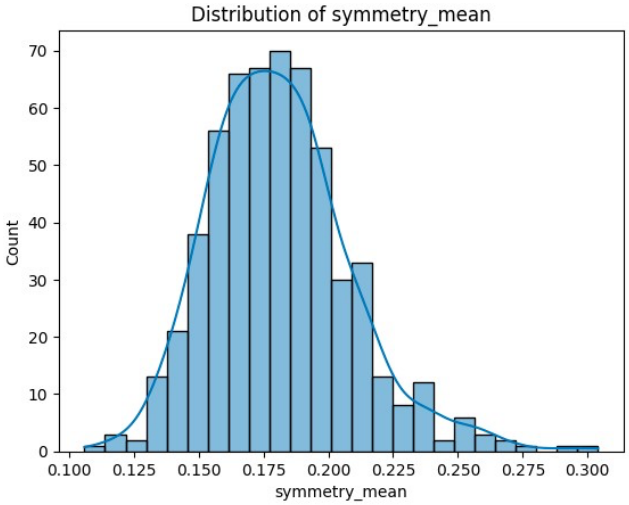
Based on the insights gained from exploring the data and further analysis we determined which features were most relevant for predicting breast cancer. In the initial stage of our feature selection process, we undertook a systematic exploration of each numerical attribute in our dataset to gain a foundational understanding of the underlying distributions and to identify potential outliers that could influence our model's performance.

Distribution Analysis:

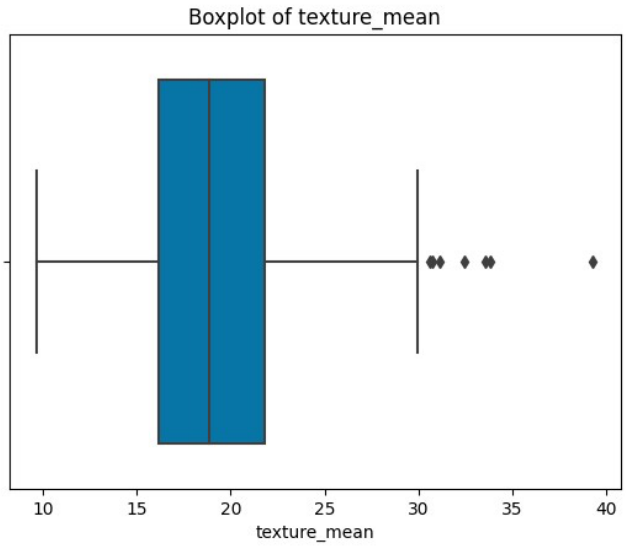
This step was crucial in understanding the data distribution, which could influence the feature selection based on the specific patterns or distributions observed.

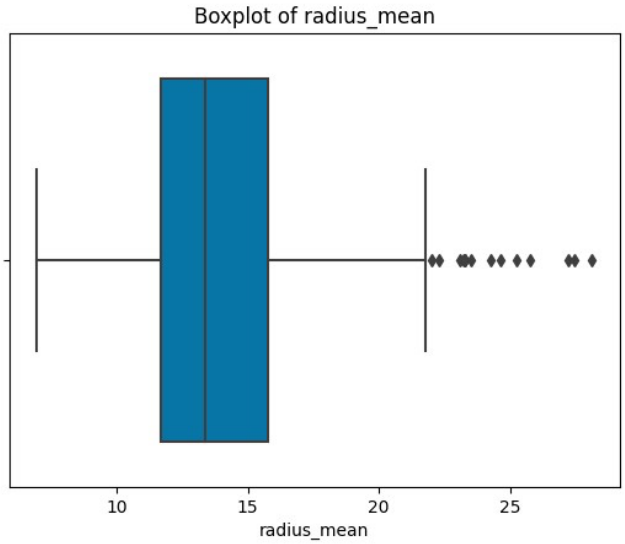
Seaborn's histplot function -





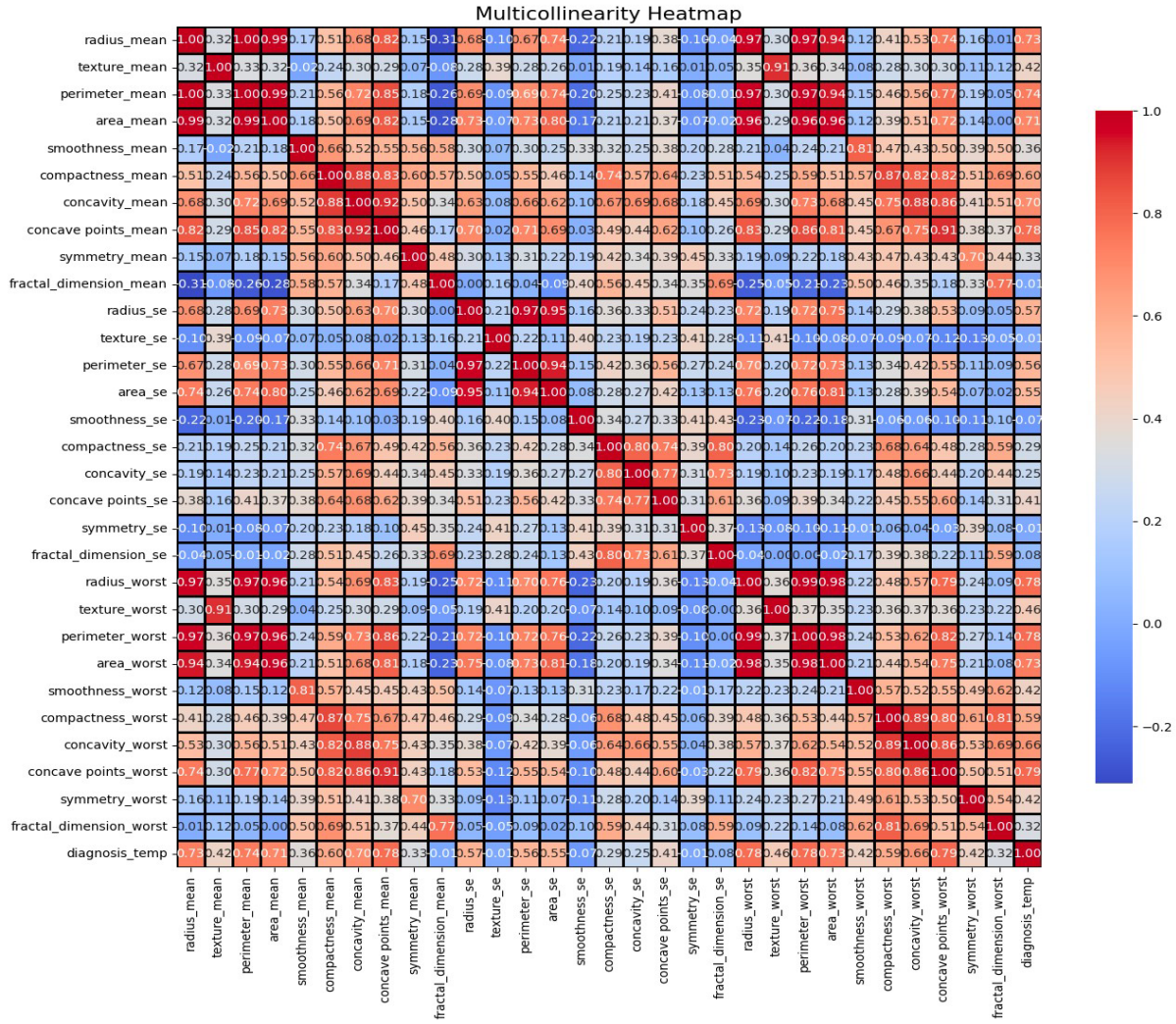
Subsequently, we utilized Seaborn's boxplot function to create boxplots for each numerical column, providing a graphic representation of the central tendency and variability of the data, alongside showcasing potential outliers.





Correlation Analysis:

The correlation matrix helps you identify relationships between features. High correlations between features can indicate multicollinearity, which may affect the model's performance. This analysis aids in selecting features that are less correlated with each other.



Implementing the Decision Tree Algorithm

Per the assignment requirements, we were tasked to implement a decision tree model using numpy, matplotlib, or Sklearn libraries. Scikit-learn's DecisionTreeClassifier typically employs an optimized version of the CART (Classification and Regression Trees) algorithm. The CART algorithm is efficient and scalable, making it suitable for datasets of various sizes.

Implementing the decision tree algorithm involved training the decision tree algorithm on the data. This allows the algorithm to learn from the dataset how to distinguish between benign and malignant cases based on the features we provided.

The classifier is trained utilizing the training dataset, thereby learning the underlying patterns present within the data. This knowledge equips the classifier with the ability to accurately forecast outcomes on previously unseen data.

Model Optimization and Hyper-parameter Tuning

The objective is to enhance the decision tree model by identifying and incorporating the best hyperparameters, ultimately improving the model's ability to make accurate predictions.

Including a varied range of values for these hyperparameters ensures that the grid search can explore a broad space of possible models, which is essential to finding a well-tuned model

Results

Our model was assessed using various metrics to evaluate its performance:

The model achieved an accuracy score of approximately 90.06%. Accuracy, however, isn't a good metric especially for imbalanced datasets and applications like medical diagnosis. It was then important to consider other metrics like precision, recall, and the confusion matrix to properly evaluate model performance.

Confusion Matrix

The confusion matrix provides insights into the model's ability to make accurate predictions.

In our case: [100 7] [10 54]

- → The top left cell (100) represents true negatives (benign cases correctly predicted)
- → The top right cell (7) represents false positives (benign cases predicted as malignant)
- → The bottom left cell (10) represents false negatives (malignant cases predicted as benign).
- → The bottom right cell (54) represents true positives (malignant cases correctly predicted).

The model achieved a precision score of 89%. Precision measures the accuracy of positive predictions (Malignant). The model achieved a recall score of 84%, it indicates that 84% of the actual malignant cases were correctly identified.

After hyperparameter tuning for our decision tree model, the model correctly predicted the outcomes for approximately 93% of the test cases. This means that hyperparameter tuning helped improve the model’s performance. The metrics indicate that our model is highly effective at identifying whether a patient has breast cancer or not.

Link to online repository – Decision Tree Model:

https://github.com/christophertheys/BreastCancerPrediction

Activity 5 - Principal Component Analysis (PCA) to a biomedical dataset

Link to online repository:

https://github.com/christophertheys/PCA

Concept and applications of Principal Component Analysis (PCA) in the context of biomedical data analysis:

Principle Component Analysis (PCA)

Principle Component Analysis (PCA) is a tool which performs a procedure that alters the original variables of a dataset into a new format. This does not alter the dataset but rather changes how it's represented. These new variables become statistically uncorrelated, termed as the 'Principal Components'.

This means that the changes in one variable are not associated with changes in another. This tool is important for reducing redundancy within the dataset and ultimately makes the data easier to analyse and visualize.

These principal components are created using a linear combination of the original variables. This further facilitates PCA to reduce the number of variables within the dataset, while still attempting to preserve as much relevant information as possible.

Key features in applications of PCA are:

- Process only works with numeric features.
- PCA is sensitive to scale.
- Remove/Constraint outliers due to influence on result.

Application and Process:

Import Libraries

```
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.decomposition import PCA
from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.impute import SimpleImputer
from sklearn.compose import ColumnTransformer
import numpy as np
from sklearn.pipeline import Pipeline
from scipy.stats import zscore
```

[32] ✓ 0.0s Python

Loading Dataset

```
# Enter full directory to load the .csv in submission file

# Dataset loaded into a pandas DataFrame

try:
    file_path = "/Users/christophertheys/Desktop/heart_disease.csv"
    data = pd.read_csv(file_path)
except FileNotFoundError:
    print(f"The file at {file_path} does not exist. Please check the file path and try again.")
    data = None
except pd.errors.EmptyDataError:
    print(f"The file at {file_path} is empty. Please provide a valid data file.")
    data = None
except Exception as e:
    print(f"An unexpected error occurred while loading the file: {e}")
    data = None

# Ensuring valid dataset

if data is not None:
# Displaying the first five rows of the dataset
    print(data.head())
    print("\n")
```

[2] ✓ 0.0s Python

```
...   Gender  age     education  currentSmoker  cigsPerDay  BPMeds
0     Male   39   postgraduate              0         0.0     0.0  \
1   Female   46  primaryschool              0         0.0     0.0
2     Male   48     uneducated              1        20.0     0.0
3   Female   61       graduate              1        30.0     0.0
4   Female   46       graduate              1        23.0     0.0

  prevalentStroke  prevalentHyp  diabetes  totChol  sysBP  diaBP    BMI  \
0              no             0         0    195.0  106.0   70.0  26.97
1              no             0         0    250.0  121.0   81.0  28.73
2              no             0         0    245.0  127.5   80.0  25.34
3              no             1         0    225.0  150.0   95.0  28.58
4              no             0         0    285.0  130.0   84.0  23.10

   heartRate  glucose Heart_ stroke
0       80.0     77.0           No
1       95.0     76.0           No
2       75.0     70.0           No
3       65.0    103.0          yes
4       85.0     85.0           No
```

Pre-processing insights

Through analysing the dataset, these feature types were identified:

- Numerical - num_features
- Categorical - cat_features Numerical feature pre-processing –

  Snippet indicates the method used in handling missing values for numerical features. The choice of ‘mean’ was the best suitable strategy for this assessment. This will indicate missing values and then be replaced by the mean value of each respective feature/column to aim in preserving data distribution.

```
numeric_transformer = Pipeline(steps=[ ('Imputer', SimpleImputer(strategy='mean'))
```

Preprocessing of Dataset

```
# Processing steps applied ensuring integrity of dataset

numerical_cols = ['age', 'cigsPerDay', 'totChol', 'sysBP', 'diaBP', 'BMI', 'heartRate', 'glucose']
categorical_cols = ['Gender', 'education', 'currentSmoker', 'BPMeds', 'prevalentStroke', 'prevalentHyp', 'diabetes', 'Heart_ stroke']


numeric_transformer = Pipeline(steps=[
    ('imputer', SimpleImputer(strategy='mean')),
    ('scaler', StandardScaler())
])

categorical_transformer = Pipeline(steps=[
    ('imputer', SimpleImputer(strategy='most_frequent')),
    ('onehot', OneHotEncoder(handle_unknown='ignore'))
])

preprocessor = ColumnTransformer(
    transformers=[
        ('num', numeric_transformer, numerical_cols),
        ('cat', categorical_transformer, categorical_cols)
    ])

Final_preprocessed = preprocessor.fit_transform(data)
```

[8] ✓ 0.0s Python

PCA Application

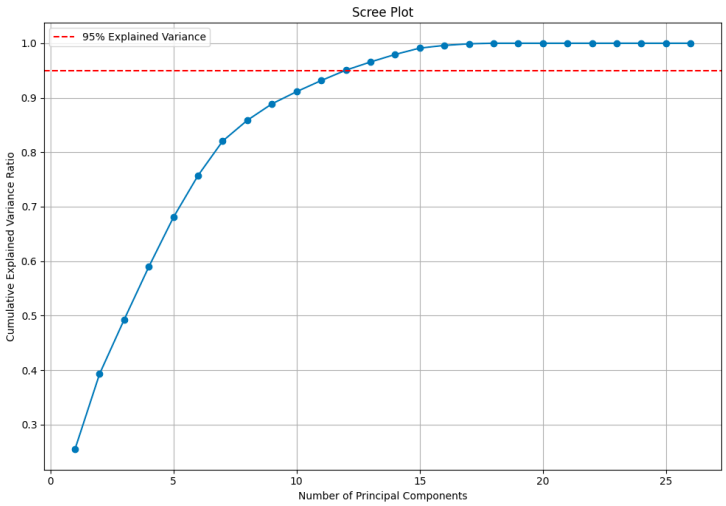
When choosing the right number of principal components, after some research I discovered that it often involves a combination of examining the explained variance and considering the specific use-case and objectives of the analysis.

One approach, which is most common to take, is looking at the cumulative explained variance ratio as a function of the number of components. A visualization is created to assist in its determination. Plotting visuals utilized to aim in deciding the number of components.

This visual aids in visually assessing how many components are necessary to explain enough variance.

Analysis of results obtained from PCA:

Implications of the extracted principal components



Application of using 2 for the number of principle components:

Implementing PCA

```
pca = PCA()
pca.fit(Final_preprocessed)

explained_variance_ratio = pca.explained_variance_ratio_

# The cumulative variance explained
cumulative_explained_variance = np.cumsum(explained_variance_ratio)

# Visual used to assess
plt.figure(figsize=(10,7))
plt.plot(range(1, len(cumulative_explained_variance)+1), cumulative_explained_variance, marker='o')
plt.title('Scree Plot')
plt.xlabel('Number of Principal Components')
plt.ylabel('Cumulative Explained Variance Ratio')
plt.axhline(y=0.95, color='r', linestyle='--', label='95% Explained Variance')
plt.legend()
plt.grid(True)
plt.tight_layout()
plt.show()
```

[12] ✓ 0.4s Python

```
pca = PCA(n_components=2)
X_pca = pca.fit_transform(Final_preprocessed)

explained_variance_ratio = pca.explained_variance_ratio_
print(f'Explained variance ratio: {explained_variance_ratio}')

plt.figure(figsize=(8,6))
plt.scatter(X_pca[:,0], X_pca[:,1])
plt.xlabel('First principal component')
plt.ylabel('Second principal component')
plt.title('2D PCA Result')
plt.show()
```

[24] ✓ 0.3s Python

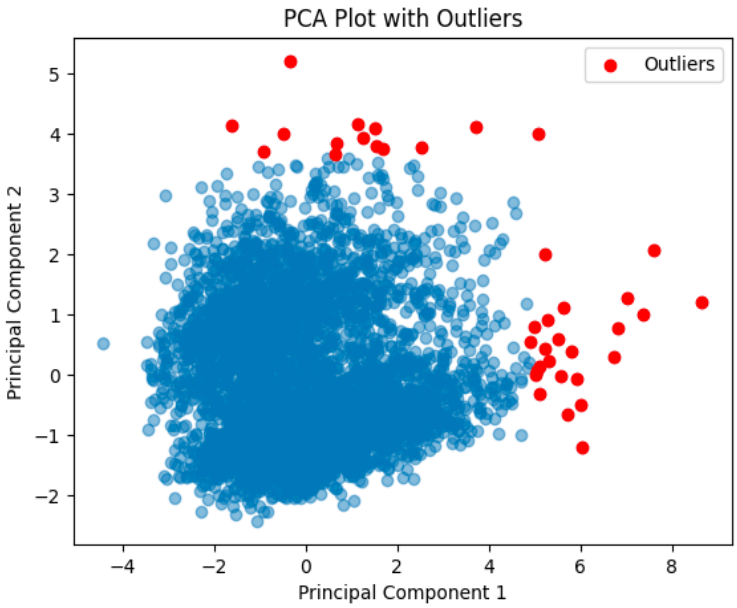
```
Explained variance ratio: [0.25479835 0.13834493]
```

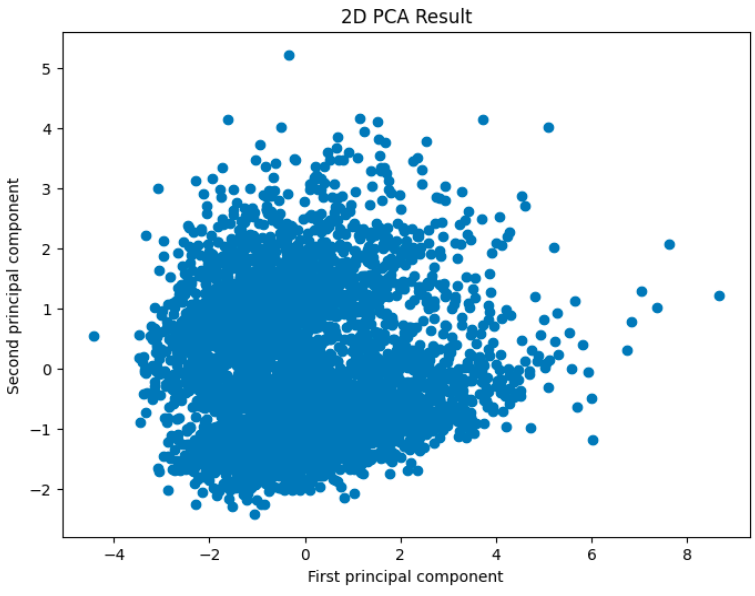
- 25.48% - The first principal component accounts for approximately 25.48% of the total variance in the dataset. This is substantial (a fourth of the total dataset) considering that this is a reduction from the original feature space from the dataset.
- 13.83% - The second principal component accounts for approximately 13.83% of the total variance in the dataset. The additional 13.83% brings the cumulative retained variance to 39.31%

Additional code of visualizing results

```
pca = PCA(n_components=2)
X_pca = pca.fit_transform(Final_preprocessed)

explained_variance_ratio = pca.explained_variance_ratio_
print(f'Explained variance ratio: {explained_variance_ratio}')

plt.figure(figsize=(8,6))
plt.scatter(X_pca[:,0], X_pca[:,1])
plt.xlabel('First principal component')
plt.ylabel('Second principal component')
plt.title('2D PCA Result')
plt.show()


z_scores = np.abs(zscore(X_pca))
# A commonly used threshold allocated for visual
threshold = 3

outliers = np.where(np.any(z_scores > threshold, axis=1))
plt.scatter(X_pca[:, 0], X_pca[:, 1], alpha=0.5)
plt.scatter(X_pca[outliers, 0], X_pca[outliers, 1], color='r', label='Outliers')

plt.xlabel('Principal Component 1')
plt.ylabel('Principal Component 2')
plt.title('PCA Plot with Outliers')
plt.legend()

plt.show()

[31]  ✓ 0.7s                                                    Python
...  Explained variance ratio: [0.25479835 0.13834493]
```

Visualization of results:





Insights in process:

Explanations

Dataset – Heart Disease Dataset
Source - https://www.kaggle.com/code/ekrembayar/heart-disease-uci-eda-pca-kmeans-hc-rf-with-r

Few issues initially faced was identifying a suitable dataset for meeting the objective of Principle Component Analysis (PCA).
General elements which are vital in achieving the objective of PCA are:

- Dimensionality Reduction
- Visualization
- Noise Reduction
- Feature Engineering
- Interpretation of Variability

Understanding the purpose of PCA was essential in selecting a dataset which meets key requirements for suitable application of the concept.
Similarly, the dataset selected should encompass these vital elements:

- Numerical Data
- High Dimensionality
- There are correlated variables.
- Has sufficient sample size.
- Able to handle pre-processing.
- Evidence of linearity

Reference List:

- Jonathon Shlens, (2014), Version 3.02, '*A Tutorial on Principal Component Analysis'.* https://arxiv.org/pdf/1404.1100&sa=U&ved=2ahUKEwi57Mfr0ZDpAhWtF6YKH fSxAck4ChAWMAZ6BAgEEAE&usg=AOvVaw2ccduDFnmcXvF-iGE-VXIM
- Md. Touhidul Islam, Sanjida Reza Rafa, Md. Golam Kibria. (2020). *Early Prediction of Heart Disease Using PCA and Hybrid Genetic Algorithm with k-Means.* *https://ieeexplore.ieee.org/abstract/document/9392655?casa_token=fYksPCsZtxsAAAAA:vWAI3wrPS2gVgDCCdGl0wy9QguxBP7zmkodPC16MsbAdvrgAZU8R4M8HZyp4lIRFJJYlpDDTMM4*

Activity 6 – Data Security

Reflection on Data Security

The task offered an insightful dive into the crucial realm of data security within project lifecycles, uncovering a spectrum of potential issues and consequential safeguard strategies. Reflecting on past projects allowed me to pinpoint potential data security vulnerabilities at various stages within each lab and activity, such as data collection, storage, transmission, and concerning aspects like access control and confidentiality.

Engaging in reflection of assignments brought forth a range of perspectives, experiences, and insights on practical data security implementations, enhancing my understanding of the practicalities and challenges of enforcing robust data security practices in real-world project contexts.

I found that this task enabled an approach of blending theoretical knowledge with practical application, fostering a well-rounded understanding of data security in the realm of project management and execution.