

INF 491/791: APPLIED DATA SCIENCE

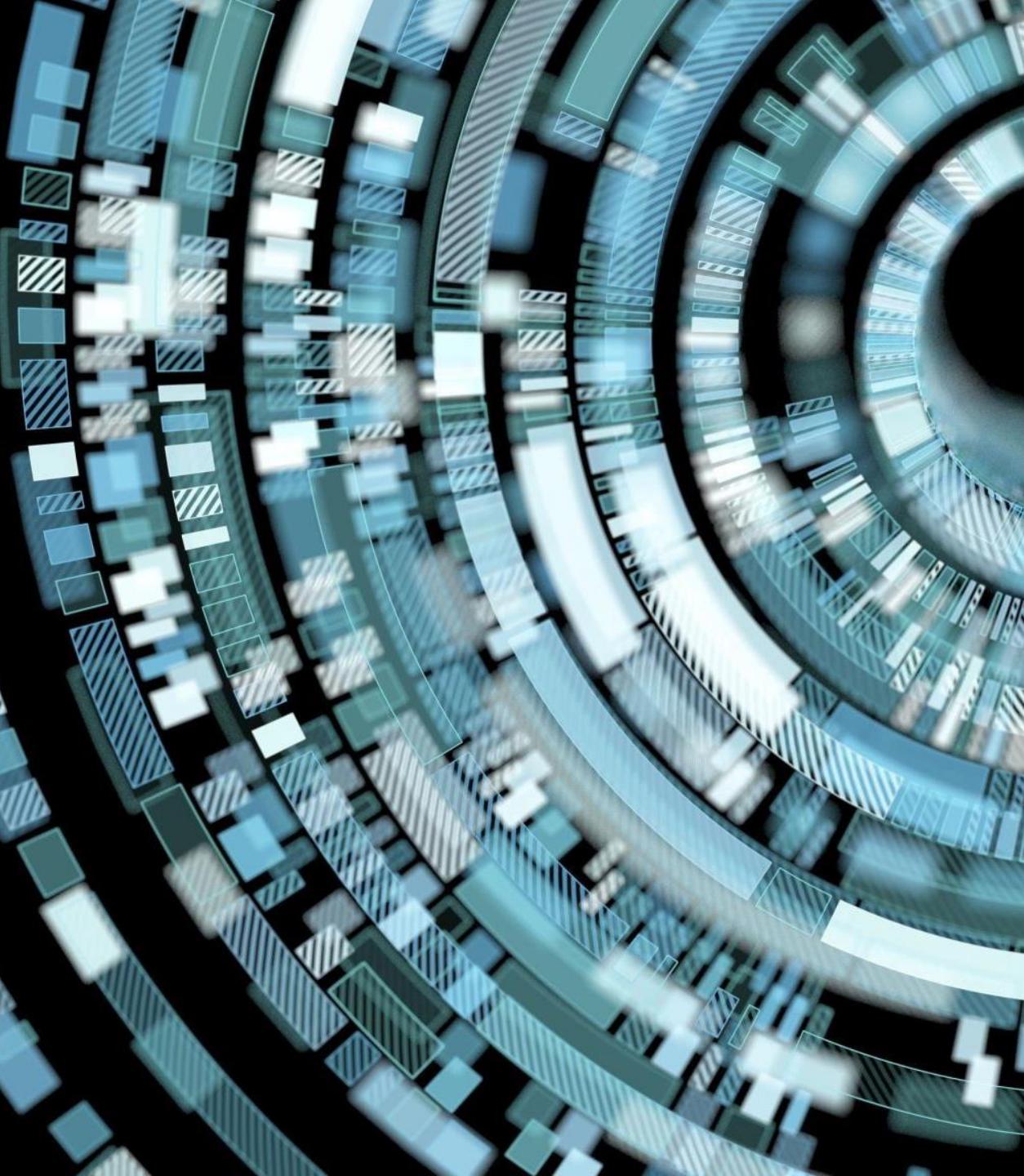
Data Mining and Big Data
Mr Mike Wa Nkongolo



Introduction to Data Mining and Big Data

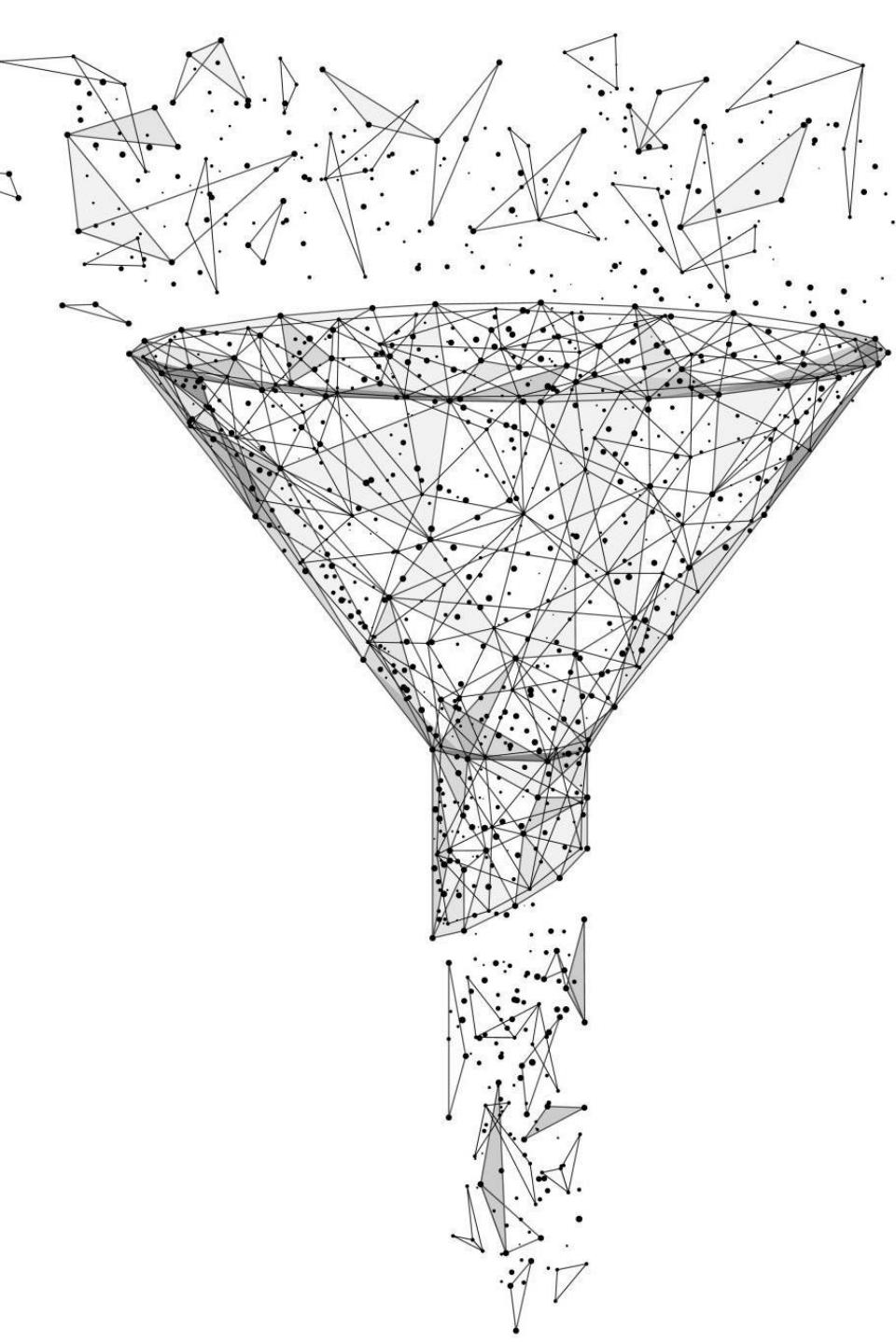
- Definition of Data Mining
- The Importance of Big Data in Data Mining
- Data Preprocessing
- Data Cleaning
- Data Integration
- Data Transformation
- Data Reduction
- Exploratory Data Analysis (EDA)





Introduction to Data Mining and Big Data

- Univariate Analysis
- Bivariate Analysis
- Multivariate Analysis
- Visualization Techniques
- Data Mining Techniques
- Classification and Prediction
- Clustering & Association Rule Mining
- Machine Learning



Introduction to Machine Learning

- Supervised vs. Unsupervised Learning
- Decision Trees and Random Forests
- Support Vector Machines
- Neural Networks and Deep Learning
- Big Data Technologies
- Challenges of Mining Big Data

Applications of Data Mining and Big Data



HEALTHCARE
ANALYTICS



BUSINESS
INTELLIGENCE



SOCIAL MEDIA
ANALYSIS



RECOMMENDER
SYSTEMS



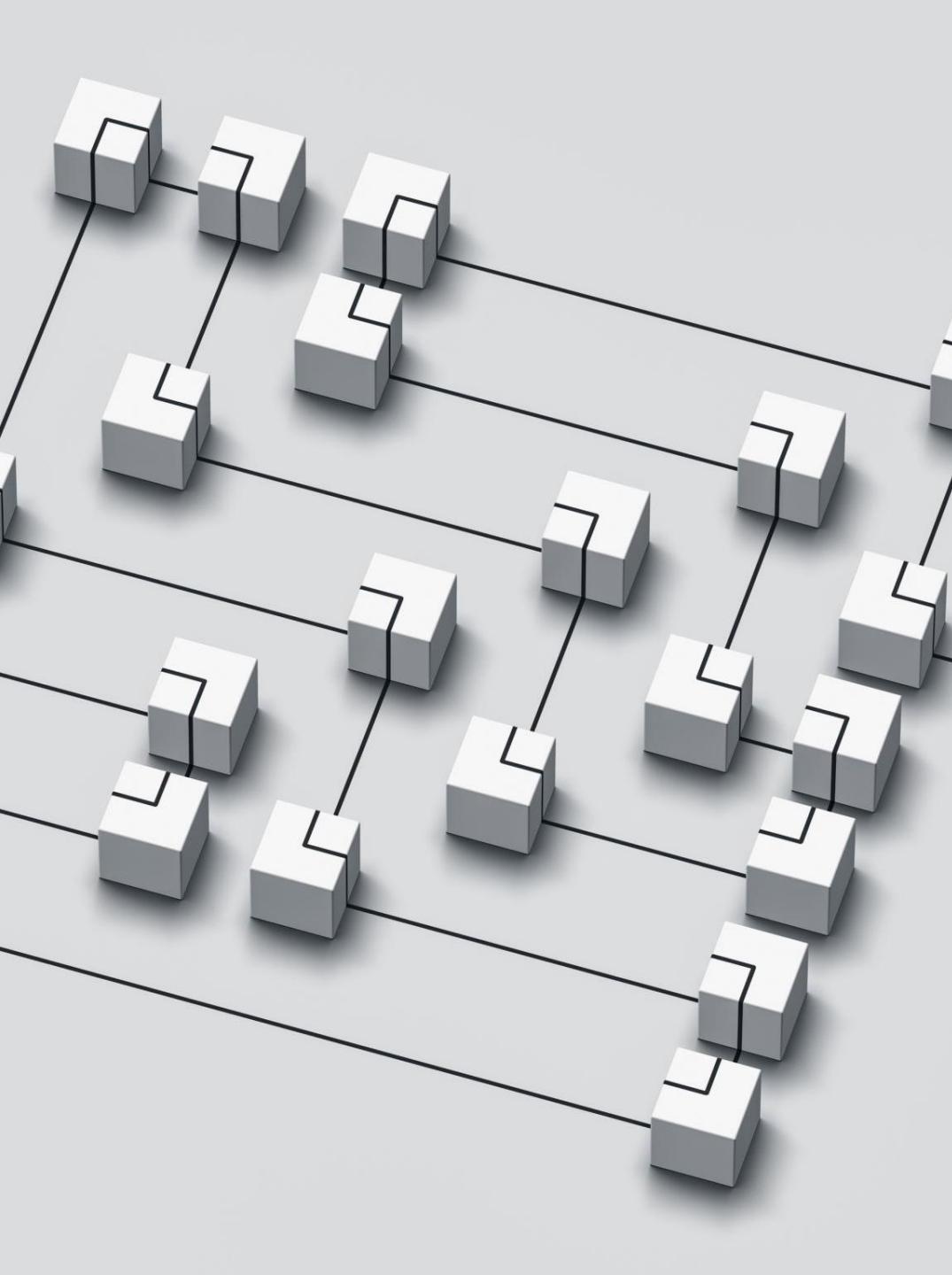
AI AND MACHINE
LEARNING
INTEGRATION



EDGE COMPUTING



BLOCKCHAIN AND
DATA SECURITY



Definition of Data Mining

- **Data Preparation:** Cleaning, transforming, and preprocessing the data to make it suitable for analysis. This includes handling missing values, dealing with outliers, and formatting data.
- **Pattern Discovery:** Employing algorithms to identify patterns, associations, correlations, or anomalies within the data.
- **Predictive Modeling:** Building predictive models that can forecast future trends or outcomes based on historical data.
- **Clustering and Classification:** Grouping similar data points into clusters or categorizing data into predefined classes or categories.
- **Data Visualization:** Presenting the results of data mining in visual formats such as charts, graphs, or reports for easier interpretation.

The Importance of Big Data in Data Mining



Real-Time Decision-Making: This is critical in applications like autonomous vehicles, where data must be analyzed instantly to ensure safety.



Scientific Discovery: In fields like genomics, particle physics, and climate science, big data plays a crucial role in scientific discovery. Researchers can process and analyze massive datasets to uncover new knowledge and make groundbreaking discoveries.



Identification of Rare Events: Big data facilitates the detection of rare events or anomalies. In fields like fraud detection, cybersecurity, and quality control, data mining techniques can identify unusual patterns or behaviors that might signal a problem or threat.



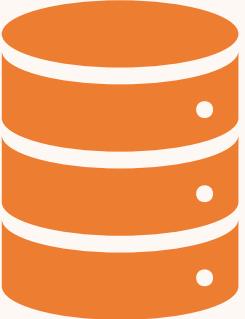
Improved Predictions: Larger datasets enable data mining models to make more accurate predictions. By analyzing vast amounts of historical and current data, these models can identify trends and patterns that lead to better forecasting, whether in financial markets, consumer behavior, or healthcare outcomes.



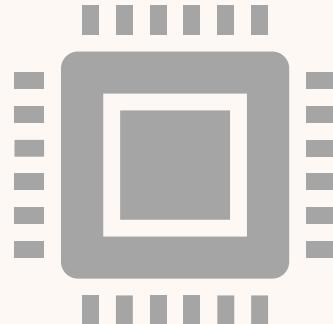
Data Preprocessing

- Involves identifying and correcting errors or inconsistencies in a dataset to improve its quality.
- **Example:** In a customer database, there may be entries with missing values, such as email addresses or phone numbers.
- Data cleaning would involve either filling in these missing values with reasonable estimates or removing records with too many missing values.

Data Integration

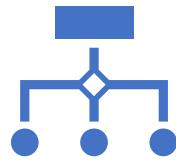


Data integration combines data from different sources into a unified view, providing a comprehensive and coherent dataset.



Example: An e-commerce company integrates data from various systems, including sales records, inventory databases, and customer information, to create a complete view of its business operations.

Data Transformation



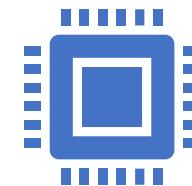
Data transformation involves converting data into a different format, structure, or scale to make it suitable for analysis or a specific application.



Example: Converting temperature readings from Fahrenheit to Celsius is a data transformation. Similarly, aggregating daily sales data into monthly totals is a transformation to a different scale.



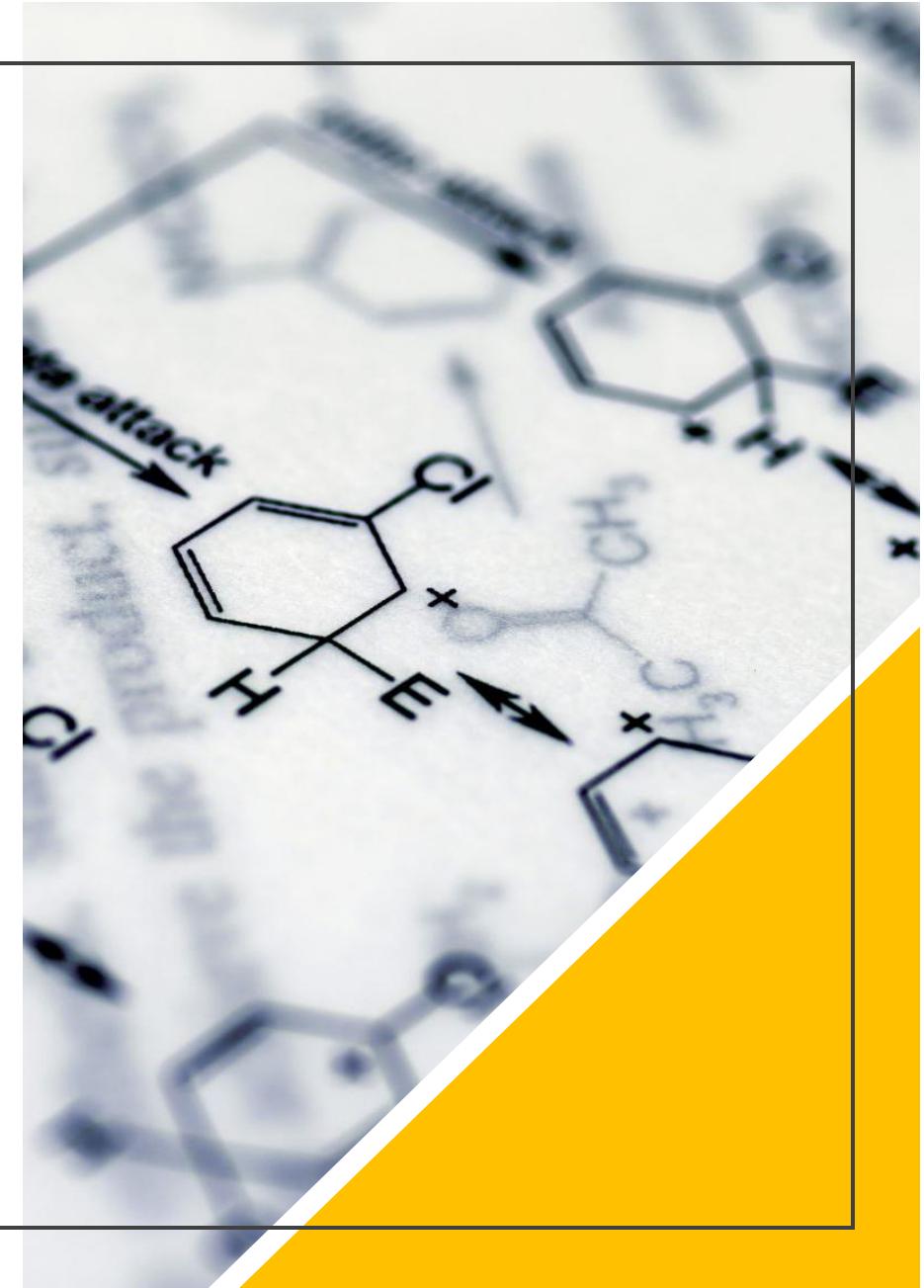
Data reduction aims to decrease the volume but produce the same or similar analytical results. It often involves selecting a representative subset of data.



Example: In machine learning, feature selection is a form of data reduction. Choosing the most relevant features (variables) from a dataset while discarding less important ones can improve model performance and reduce computational costs.

EDA

- EDA is the process of summarizing, visualizing, and understanding the main characteristics of a dataset before conducting formal analyses.
- **Example:** Before building a predictive model, a data scientist might use EDA techniques to create histograms, scatter plots, and summary statistics to uncover trends, anomalies, or relationships in the data.



Data Analysis



Univariate analysis focuses on a single variable or feature in a dataset to understand its distribution, central tendencies, and variability.



Example: If you want to understand the distribution of ages in a population, you can create a histogram or a box plot showing the frequency or density of different age groups.



Bivariate analysis involves analyzing the relationship or association between two variables in a dataset.



Example: To study how the price of a product is affected by its advertising spending, you can create a scatter plot with advertising spending on the x-axis and product price on the y-axis.

Data Analysis



Multivariate analysis deals with the simultaneous analysis of more than two variables in a dataset to uncover complex relationships and patterns.



Example: Principal Component Analysis (PCA) is a multivariate technique that reduces the dimensionality of a dataset while preserving as much variance as possible. It's often used for feature reduction in machine learning.



Visualization Techniques: involves representing data graphically to provide insights, identify trends, or convey information effectively.

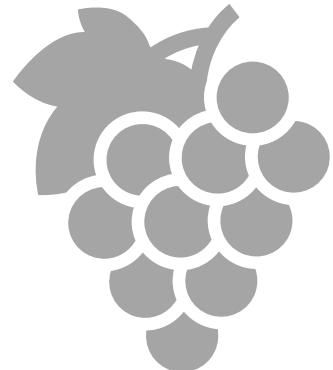


Example: Creating a heatmap to visualize the correlation matrix of variables in a dataset. High correlations are shown in one color, while low correlations are shown in another.

Data Mining



Data mining techniques are methods used to extract patterns, knowledge, or valuable information from large datasets.



Example: Using association rule mining to identify patterns in customer purchase data, such as discovering that customers who buy cereal and milk are also likely to buy bread.

Classification and Prediction



Classification involves assigning data instances to predefined categories or classes based on their characteristics.



For example, classifying emails as spam or not spam based on their content.



Prediction aims to estimate a numerical or categorical outcome for a data instance based on historical data.



For instance, predicting a student's future GPA based on their past academic performance.



Clustering groups similar data points together based on their inherent characteristics, without predefined categories.



Example: Clustering customer data to segment them into distinct groups, such as "high spenders," "budget shoppers," and "window shoppers," based on their shopping behavior.



Association Rule Mining

- Association rule mining identifies patterns or relationships between items in a dataset. It's commonly used in market basket analysis to discover which items are frequently purchased together.
- **Example:** Supermarkets analyzing customer purchase data to find associations like "Customers who buy bread also tend to buy butter."
- Anomaly Detection: Anomaly detection identifies data instances that deviate significantly from the norm or exhibit unusual behavior. It's used to find rare events or outliers.
- **Example:** Detecting fraudulent credit card transactions by identifying transactions that significantly differ from a user's typical spending behavior.

Machine Learning

- Machine learning algorithms are computational techniques that enable computers to learn and make predictions or decisions from data without being explicitly programmed.
- **Example:** Using a decision tree algorithm to predict whether a loan applicant is likely to default based on features like credit score, income, and loan amount.



Unsupervised Learning

- In supervised learning, the algorithm learns from a labeled dataset, where each input data point is associated with a corresponding target or output label.
- The algorithm's goal is to learn a mapping or relationship between the input data and the output labels. It aims to make predictions or classifications based on new, unseen data.
- **Example:** Consider a spam email classifier. You have a dataset of emails, and each email is labeled as either "spam" or "not spam" (ham). The features of these emails, such as the words used and email metadata, serve as input data.
- In supervised learning, the algorithm learns from this dataset to predict whether incoming emails are spam or not based on their features. It uses the labels (spam or not spam) to train and fine-tune its predictive model.



Supervised Learning

- In unsupervised learning, the algorithm works with unlabeled data, meaning it doesn't have access to explicit target labels or categories. Instead, the algorithm tries to discover patterns, structures, or relationships within the data without any prior guidance.
- **Unsupervised learning** is often used for tasks like clustering, dimensionality reduction, and anomaly detection.
- **Example:** Imagine you have a dataset of customer purchase history, including what items they bought and when. In unsupervised learning, you might use clustering to group similar customers together based on their purchase behavior. The algorithm would identify patterns within the data, such as customers who frequently buy electronics, customers who prefer clothing, etc., without any predefined categories. This can help businesses understand customer segments for targeted marketing.

A blackboard filled with mathematical calculations and diagrams. At the top right, there is a diagram of a rectangle divided into two triangles by a diagonal line, with the area labeled $\sqrt{2456.96} - 10$. Below it is a graph of a curve labeled $D(x) = a + b * x^c$, with points (x_1, y_1) and (x_2, y_2) marked. To the right, there is a shaded region with a boundary labeled x^2 . Further down, there is a circle with radius r and a shaded sector. Various equations are scattered across the board, including $\sqrt{a^2 + b^2} = x^2$, $c(x, y) \begin{cases} xy = c \\ cx - cy \\ 2\pi = c \end{cases}$, and $\text{men} = 584. + n^{0.9} (x^2 +$. There are also some diagrams of triangles and a grid at the bottom.

Learning

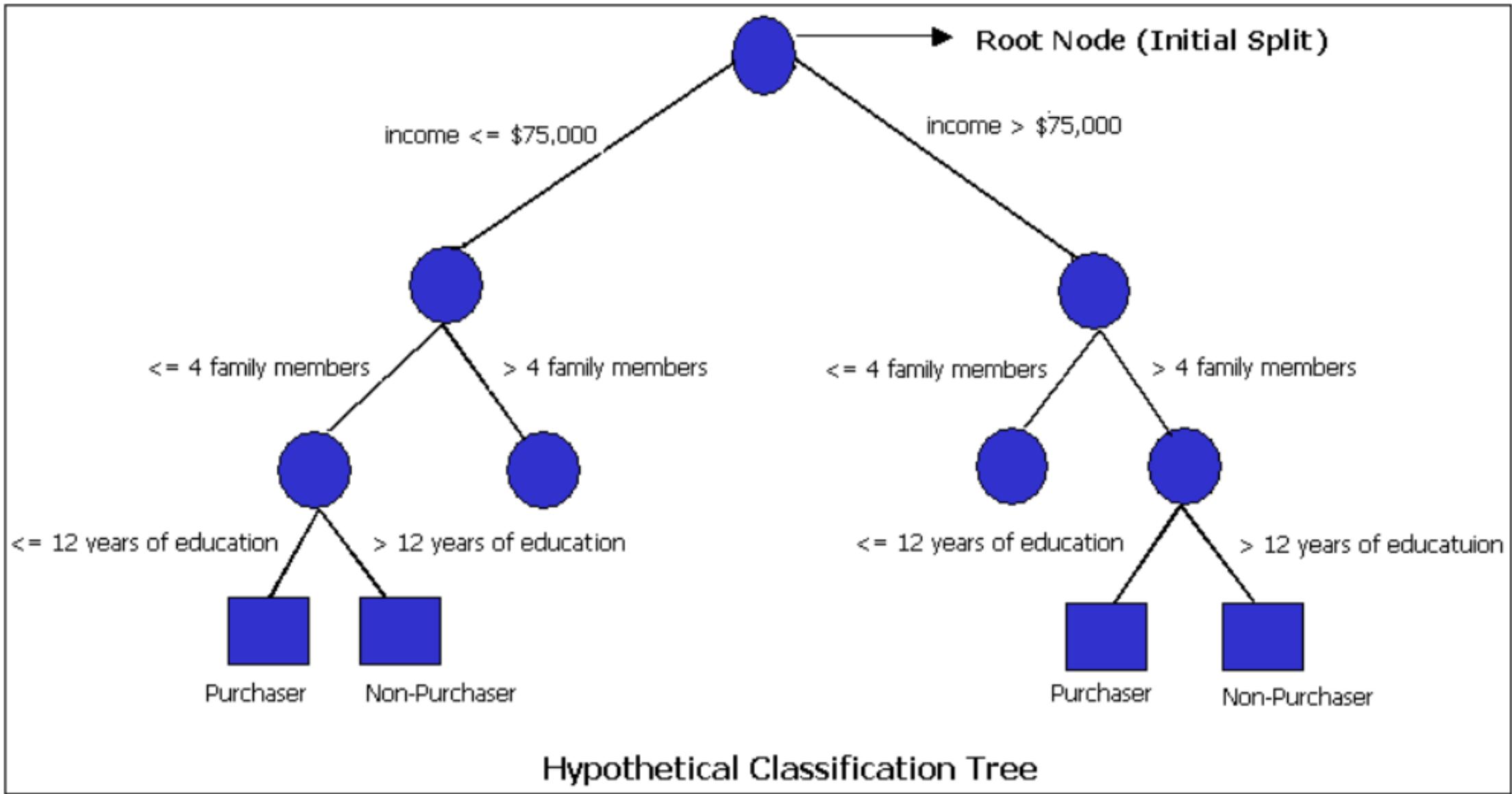
- In summary, supervised learning relies on labeled data to make predictions or classifications, while unsupervised learning seeks to discover patterns and structures in unlabeled data. Both approaches have distinct use cases and are essential in various machine learning applications.



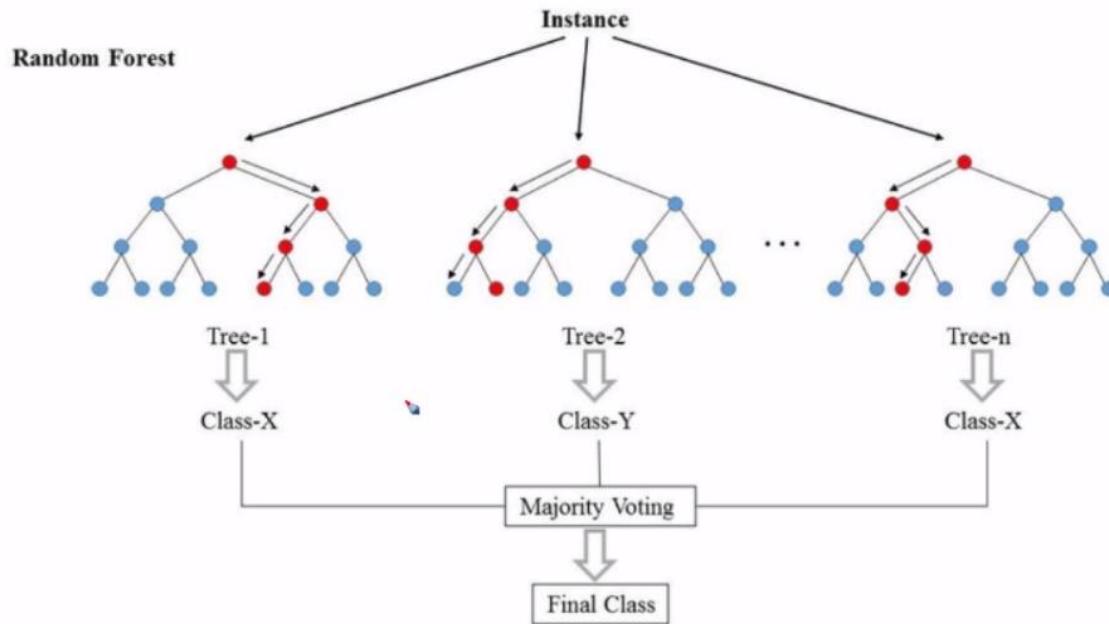
Machine Learning Algorithms

- Decision trees are a supervised learning algorithm used for classification and regression tasks. They partition the input data into subsets based on features and recursively make decisions at each node to arrive at a final prediction. Random Forests, on the other hand, are an ensemble method that consists of multiple decision trees. They work by aggregating the predictions of multiple decision trees to improve accuracy and reduce overfitting.





Random Forest



As above, **Random Forest** consists of many trees which have different shape. In order to make better performance, the shape of each tree should be different. Each tree predicts classification or regression and the **Random Forest** make result with **majority voting**.

- Decision trees are simple to understand and interpret but can be prone to overfitting when they become too deep.
- Random Forests, on the other hand, mitigate overfitting by averaging the predictions of many decision trees, making them more robust and accurate. While decision trees can be used for both classification and regression, Random Forests are primarily used for classification tasks.
- Supervised or Unsupervised: Both Decision Trees and Random Forests are supervised learning algorithms because they require labeled data during training. In the case of classification, they need labeled examples to learn the decision boundaries.

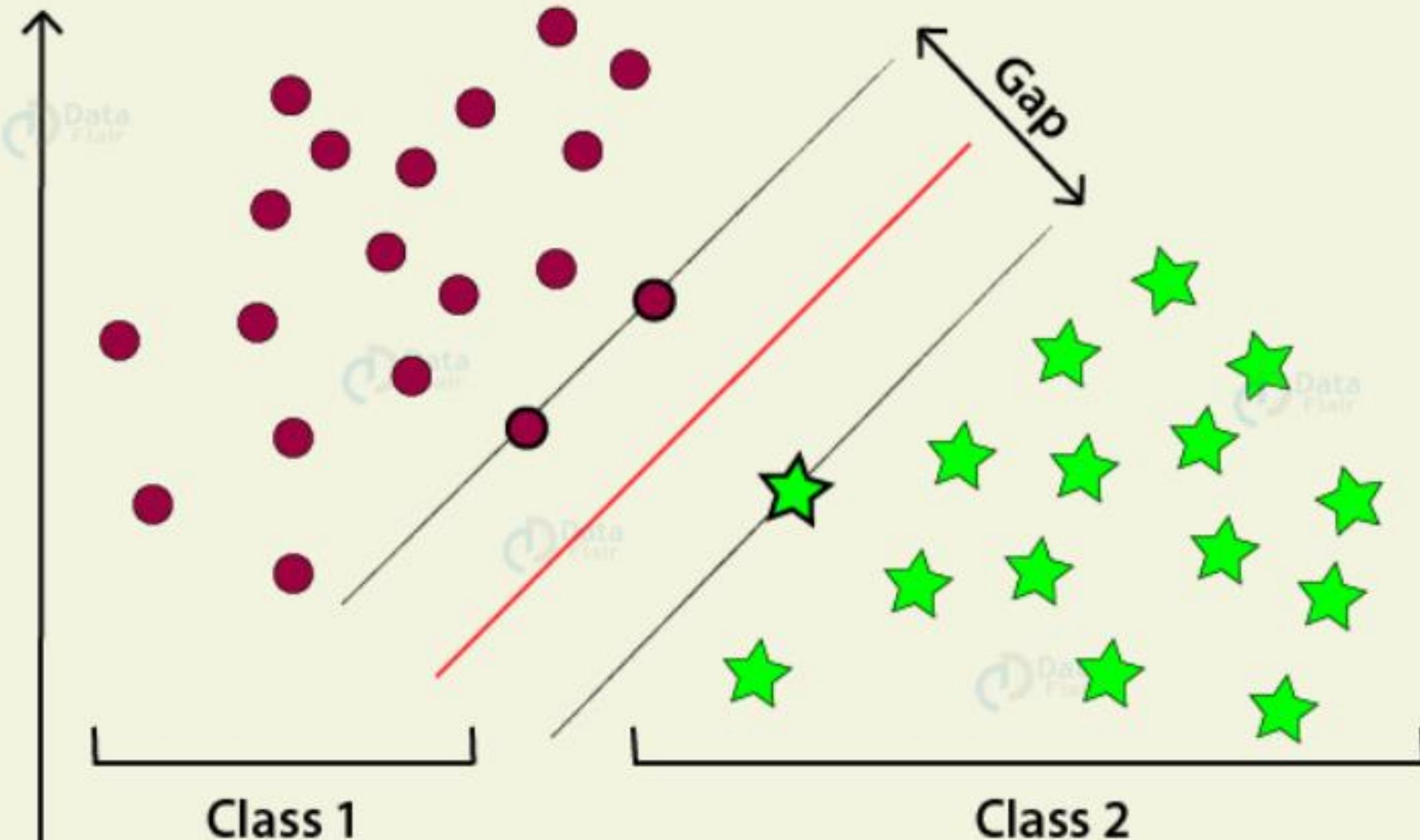


SVM

- SVM is a supervised learning algorithm used for classification and regression tasks. It finds a hyperplane or decision boundary that best separates different classes in the input data while maximizing the margin between them.
- SVM can also handle non-linear classification by using kernel functions.
- **Difference:** SVM aims to find the optimal decision boundary that maximizes the margin between data points of different classes. It is effective in high-dimensional spaces and can handle non-linear data using kernel tricks.
- SVM tries to find the "best" boundary by maximizing the margin, **while other algorithms like decision trees might create simpler boundaries based on recursive splits.**
- Supervised or Unsupervised: SVM is a supervised learning algorithm, as it relies on labeled data during training to learn the decision boundary.



Introduction to SVM



Neural Networks

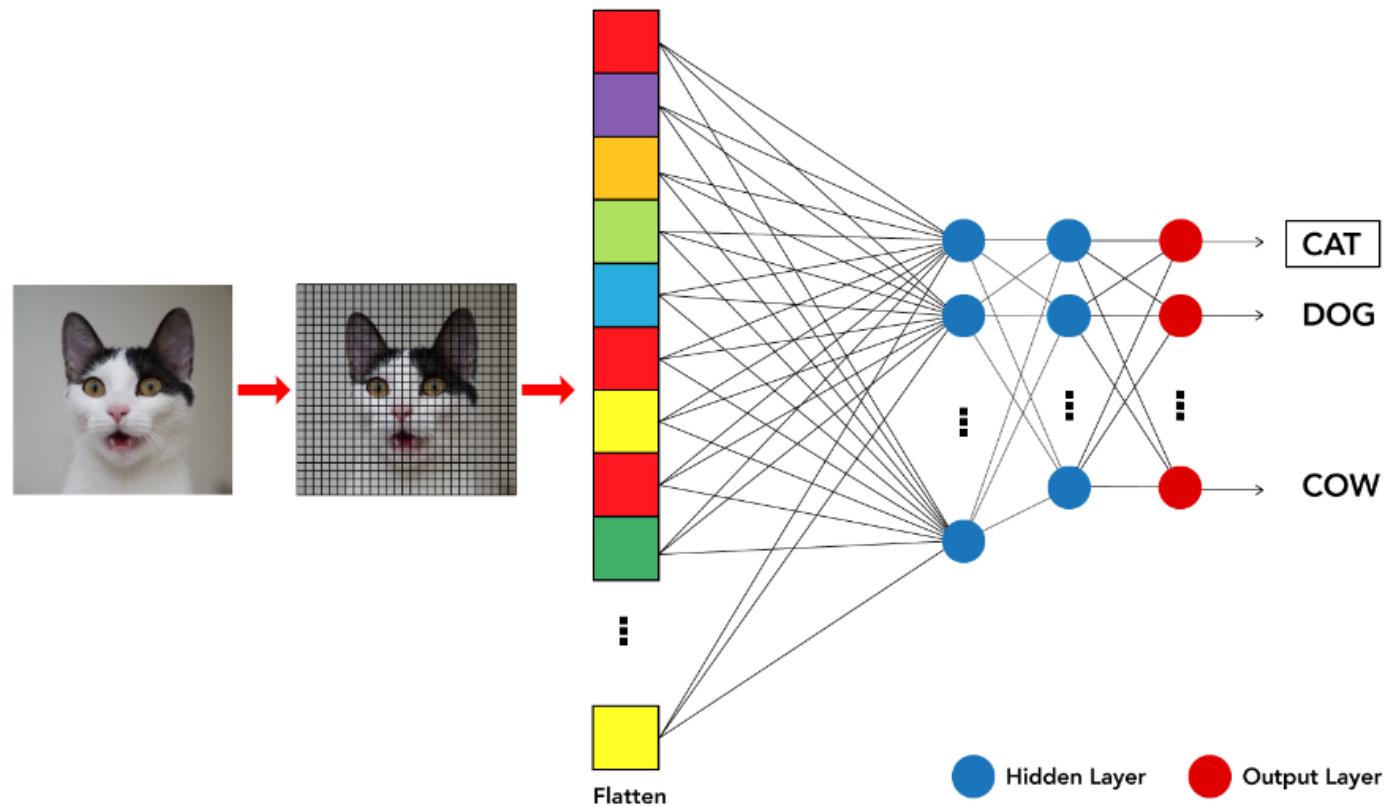


Neural networks are a family of machine learning algorithms inspired by the structure and function of the human brain.



Deep Learning is a subfield of machine learning that focuses on neural networks with many layers (deep neural networks). Neural networks consist of interconnected nodes or neurons that process and transform data through multiple layers to make predictions.

Neural Networks and Layers



Machine Learning Evaluation

		Predicted Values	
		Negative	Positive
Actual Values	Negative	TN True Negative	FP False positive
	Positive	FN False Negative	TP True Positive

Confusion matrix with 2 class labels.

		True Class		
		Positive	Negative	Measures
Predicted Class	Positive	True Positive TP	False Positive FP	Positive Predictive Value (PPV) $\frac{TP}{TP+FP}$
	Negative	False Negative FN	True Negative TN	Negative Predictive Value (NPV) $\frac{TN}{FN+TN}$
	Measures	Sensitivity $\frac{TP}{TP+FN}$	Specificity $\frac{TN}{FP+TN}$	Accuracy $\frac{TP+TN}{TP+FP+FN+TN}$

Code

- Dataset used: UGRansome
- Code uploaded on ClickUp

```
In [32]: #The %%time command is typically used in Jupyter Notebook environments, such as Jupyter Notebook or JupyterLab.  
#It is called a "magic command" and is used to measure the execution time of a specific code cell.  
#when you include %%time at the beginning of a cell, it tells Jupyter to measure the time it takes to run the code within  
#that cell  
%%time  
  
# Import various libraries and tools for building and evaluating machine learning models in Python  
# Imported models: ensemble, random forest, SVM, Naive Bayes, genetic algorithm  
# Imported evaluation metrics: accuracy, precision, recall, f1 score  
  
from sklearn.ensemble import RandomForestClassifier  
from sklearn.svm import LinearSVC  
from sklearn.naive_bayes import GaussianNB  
from sklearn.model_selection import train_test_split  
  
from sklearn.ensemble import StackingClassifier #ensmbl method of stacking classify for ensmbling  
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score  
from sklearn.metrics import confusion_matrix  
from sklearn.metrics import classification_report  
  
from genetic_selection import GeneticSelectionCV #function to import gA -> cross validation  
from sklearn.tree import DecisionTreeClassifier #estimator in GA  
import numpy as np  
  
import warnings  
warnings.filterwarnings('ignore')
```

Wall time: 2.4 s

```
In [33]: rf = RandomForestClassifier(n_estimators=100, random_state=42) # It specifies the number of trees in the Random Forest.  
#In this case, there are 100 trees in the forest  
  
# random_state: This parameter is used to set the random seed for reproducibility.  
#By setting it to 42, the randomization process will be the same each time the code is run,  
#ensuring consistent results for the Random Forest model.
```

```
In [34]: rf.fit(X_train, y_train)  
  
#This code snippet trains the Random Forest classifier (rf) on the training data (X_train and y_train).  
#In other words, it uses the features in X_train to learn the patterns and relationships in the data that correspond to the  
#target labels in y_train. This is a crucial step in building a machine learning model, as it allows the model to learn from  
#the training data and make predictions on new, unseen data based on what it has learned during training.
```

```
Out[34]: RandomForestClassifier(random_state=42)
```

```
In [35]: rf_pred=rf.predict(X_test)  
  
#This code snippet uses the trained Random Forest classifier (rf) to make predictions on the test data (X_test).  
#The predict method takes the test features in X_test as input and produces predicted labels for these features.  
#The predictions are stored in the rf_pred variable, which can be used for further evaluation or analysis to assess how well  
#the model performs on unseen data.
```

```
In [36]: rf_accuracy = accuracy_score(rf_pred, y_test)  
rf_report = classification_report(rf_pred, y_test)  
rf_matrix = confusion_matrix(rf_pred, y_test)  
print('Accuracy of Random Forest : ', round(rf_accuracy, 3))  
print('Classification report of Random Forest : \n', rf_report)  
print('Confusion Matrix of Random Forest : \n', rf_matrix)
```

Activate Windows
Go to Settings to activate Windows.

Accuracy of Random Forest : 1.0

Classification report of Random Forest :

	precision	recall	f1-score	support
0	1.00	1.00	1.00	7394
1	1.00	1.00	1.00	13648
2	1.00	1.00	1.00	8767
accuracy			1.00	29809
macro avg	1.00	1.00	1.00	29809
weighted avg	1.00	1.00	1.00	29809

Confusion Matrix of Random Forest :

```
[[ 7394      0      0]
 [      0 13648      0]
 [      0      0 8767]]
```

Accuracy of SVM : 0.543

Classification report of SVM :

	precision	recall	f1-score	support
0	1.00	0.35	0.52	21026
1	0.00	1.00	0.00	16
2	1.00	1.00	1.00	8767
accuracy			0.54	29809
macro avg	0.67	0.78	0.51	29809
weighted avg	1.00	0.54	0.66	29809

Confusion Matrix of SVM :

```
[[ 7394 13632      0]
 [     0    16      0]
 [     0      0  8767]]
```

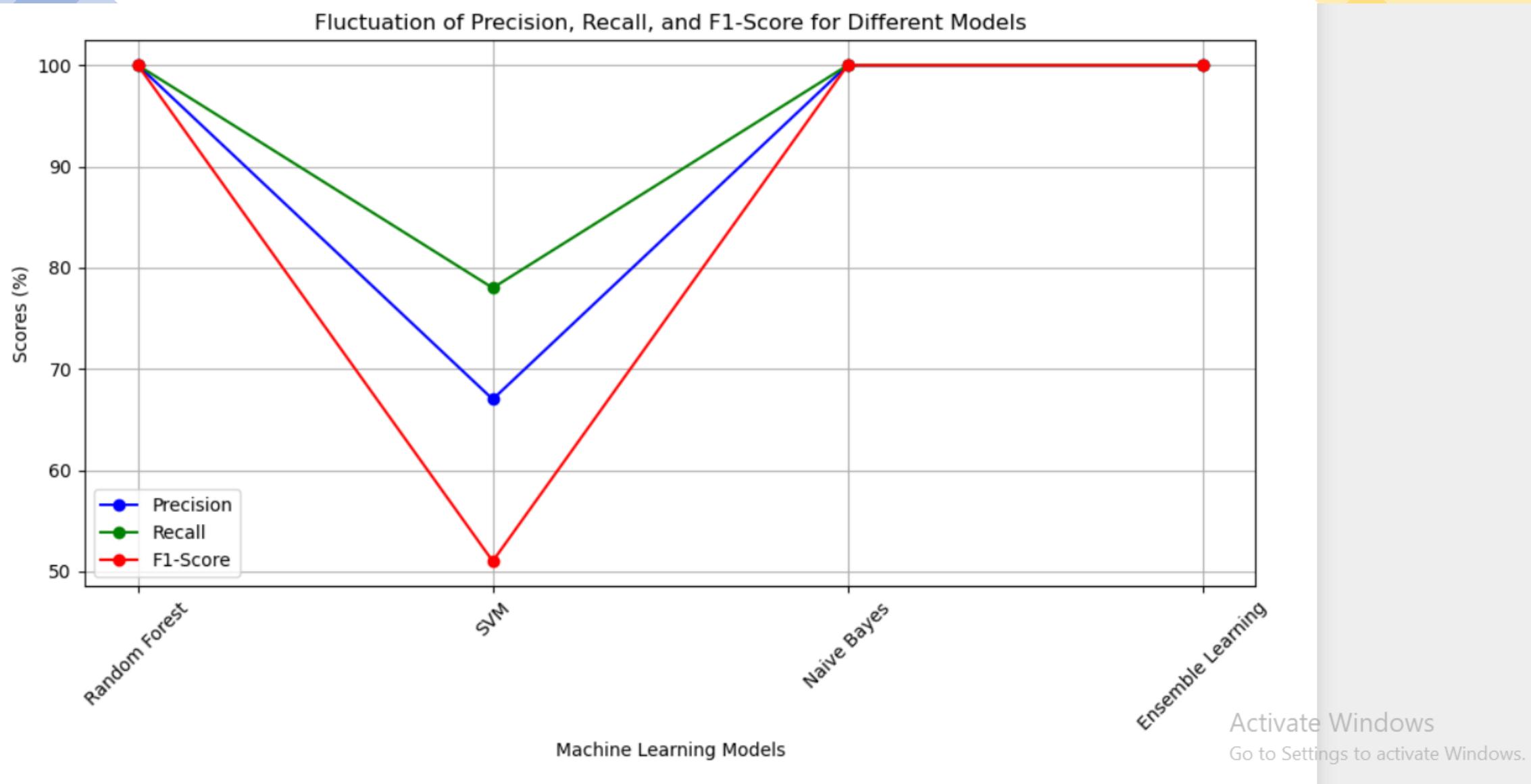
Accuracy of Naive Bayes : 1.0

Classification report of Naive Bayes :

	precision	recall	f1-score	support
0	1.00	1.00	1.00	7394
1	1.00	1.00	1.00	13648
2	1.00	1.00	1.00	8767
accuracy			1.00	29809
macro avg	1.00	1.00	1.00	29809
weighted avg	1.00	1.00	1.00	29809

Confusion Matrix of Naive Bayes :

```
[[ 7394    0    0]
 [    0 13648    0]
 [    0    0 8767]]
```



Activate Windows
Go to Settings to activate Windows.

Challenges of Mining Big Data

- **Volume:** Big data involves massive volumes of information that traditional data mining tools and techniques struggle to handle efficiently.
- **Velocity:** Data streams in at an unprecedented speed, such as social media updates, sensor data, and financial transactions. Real-time or near-real-time processing is essential to extract valuable insights promptly.



- **Veracity:** Big data often contains noisy, incomplete, or inaccurate information. Ensuring data quality and dealing with uncertainties can be a substantial challenge in mining big data effectively.
- **Value:** Identifying valuable insights from big data can be challenging. With vast amounts of data, it's easy to get lost in irrelevant information.
- **Privacy and Security:** The sheer amount of data increases the risk of privacy breaches and security threats. Safeguarding sensitive information while allowing data mining is a significant challenge.
- **Scalability:** Traditional data mining algorithms may not scale efficiently to handle big data. Developing scalable algorithms and distributed computing solutions is essential.
- **Interoperability:** Integrating big data tools and platforms into existing IT infrastructures can be complex. Ensuring that new technologies work seamlessly with legacy systems is a challenge.
- **Regulatory Compliance:** Big data mining must adhere to various regulations and legal constraints, such as GDPR in Europe or HIPAA in healthcare. Ensuring compliance while mining data can be challenging.
- **Resource Constraints:** Processing and storing big data require significant computational and storage resources. Ensuring access to these resources can be a challenge, especially for smaller organizations.
- **Ethical Concerns:** As data mining becomes more pervasive, ethical concerns regarding data privacy, surveillance, and potential biases in algorithms need to be addressed.
- **Complexity:** Big data projects often involve complex data preparation, feature engineering, and modeling processes. Managing this complexity and ensuring the interpretability of results can be challenging.

Practical Demonstration

- Application of Decision Tree, Random Forest, Support Vector Machine, Naïve Bayes, and Ensemble Learning algorithms using the UGRansome dataset.
- Application of Genetic Algorithm using the UGRansome dataset.
- See code sample uploaded on Clickup.

Bibliography

- Fortino, A., 2023. **Data Mining and Predictive Analytics for Business Decisions: A Case Study Approach.** Stylus Publishing, LLC.
- Olson, D.L. and Araz, Ö.M., 2023. **Data mining and analytics in healthcare management: Applications and tools (Vol. 341).** Springer Nature.
- Kahil, M.S., Bouramoul, A. and Derdour, M., 2023. **Big data visual exploration as a recommendation problem.** International Journal of Data Mining, Modelling and Management, 15(2), pp.133-153.
- Nkongolo, M.W., **Downloading the UGRansome Dataset.** 10.13140/RG.2.2.23570.07363/1

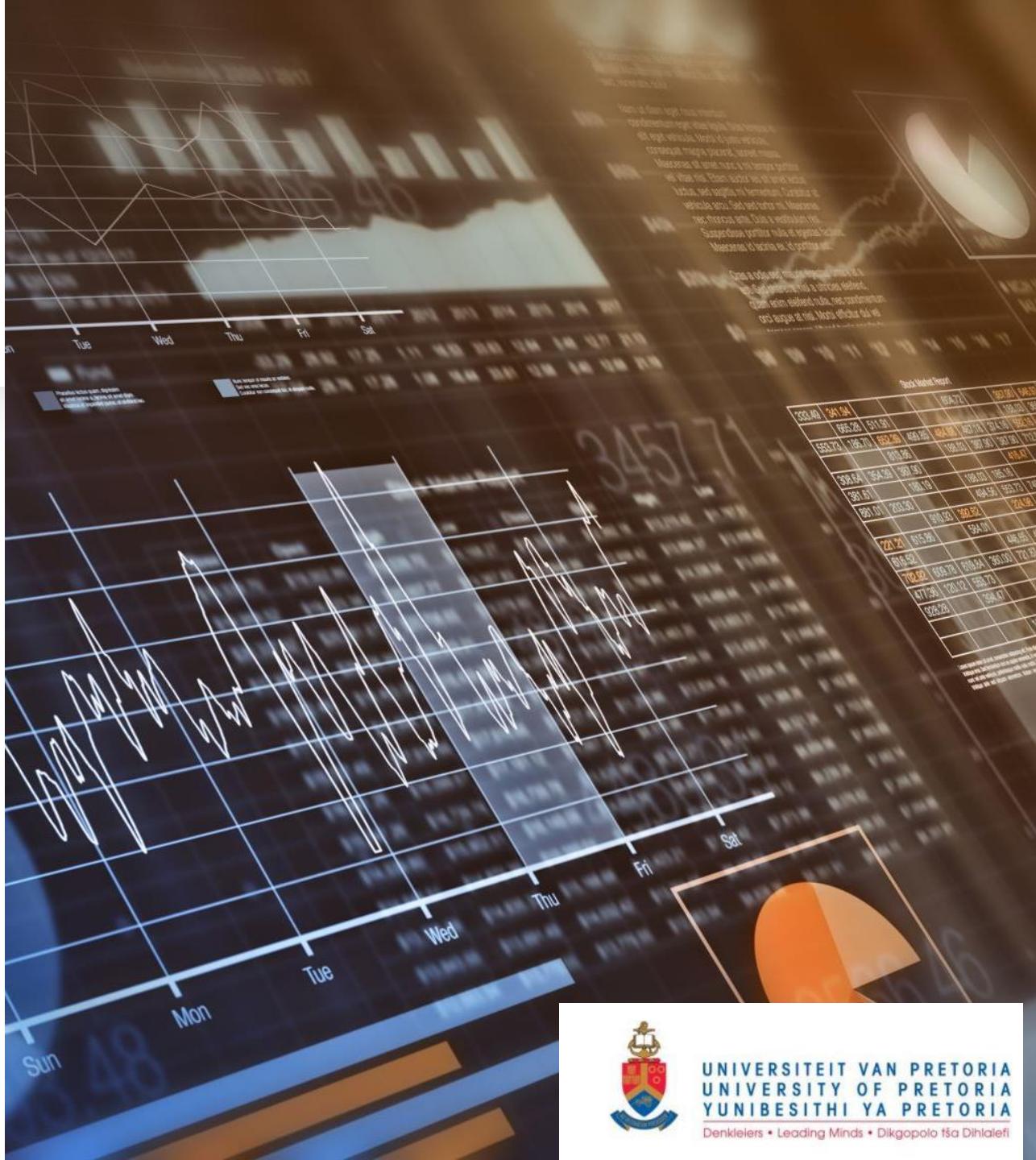
INF 491/791: APPLIED DATA SCIENCE



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA
Denkleiers • Leading Minds • Dikgopolo tsa Dihlalefi

Data Exploration and Visualization

- Importance of Data Exploration and Visualization in Data Science.
- Overview of Jupyter Notebook as a powerful tool for interactive data analysis and visualization.
- Understanding the Dataset:
 - Loading and inspecting data using Pandas.
 - Overview of dataset structure, columns, and data types.
 - Basic summary statistics and information.
- Data Cleaning and Preprocessing:
 - Handling missing values, duplicates, and outliers.
 - Data transformation and normalization techniques.
- Exploratory Data Analysis (EDA):
 - Generating descriptive statistics and visualizations.
 - Distribution plots, histograms, and box plots.
 - Correlation analysis and scatter plots.



- **Introduction to Data Visualization:**
 - Role of visualization in understanding data patterns.
 - Benefits of using Matplotlib and Seaborn libraries.
- **Creating Basic Plots:**
 - Line plots, bar plots, and scatter plots using Matplotlib.
 - Customizing plot appearance, labels, and legends.



- **Case Study and Hands-on Practice:**
- Real-world Data Exploration:
 - Applying data exploration and visualization techniques to a real dataset.
 - Exploring patterns, trends, and insights.
- Hands-on Jupyter Notebook Exercises:
 - Guided exercises using Jupyter Notebook, Pandas, Matplotlib, Seaborn, and Plotly.
 - Data cleaning, visualization, and interpretation.



Importance of Data Exploration and Visualization in Data Science

- Data Exploration and Visualization play a fundamental and pivotal role in the field of Data Science, serving as the cornerstone of informed decision-making and insightful analysis. Their importance lies in their ability to unearth hidden patterns, trends, and relationships within complex datasets, translating raw information into actionable insights.



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA
Denkleiers • Leading Minds • Dikgopololo tša Dihlatlefi

Importance of Data Exploration and Visualization in Data Science

- **Data Quality Assessment:** Visualization aids in identifying data quality issues such as outliers, missing values, or inconsistencies. These visual cues guide data cleaning and preprocessing, enhancing the accuracy and reliability of subsequent analyses.
- **Feature Selection and Engineering:** Effective data exploration helps in selecting relevant features for modeling while also inspiring the creation of new features that might enhance predictive performance. This contributes to more robust and accurate machine learning models.
- **Enhanced Creativity:** Visualization encourages creative thinking and innovative problem-solving. Exploring data visually can lead to the discovery of new angles or perspectives that may not have been evident through traditional numerical analysis alone.



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA
Denkleiers • Leading Minds • Dikgopololo tša Dihlatlefi

Overview of Jupyter Notebook as a powerful tool for interactive data analysis and visualization

- You can save hours of manual and tedious EDA work.

While EDA is vital, it is often a time-consuming task.

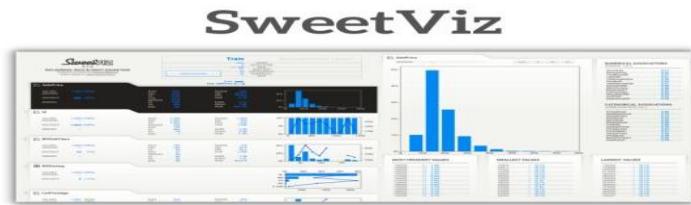
The below visual summarizes 8 powerful EDA tools. These tools automate many redundant steps of EDA and help you profile your data in quick time.



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA
Denkleiers • Leading Minds • Dikgopololo tša Dihlatlefi

Overview of Jupyter Notebook as a powerful tool for interactive data analysis and visualization

8 Automated EDA Tools



Pandas-Profiling



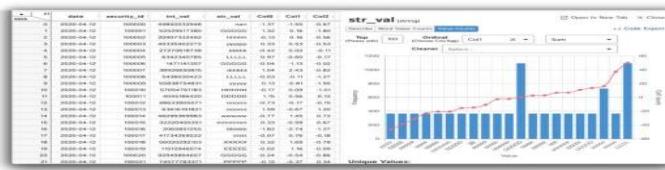
DataPrep



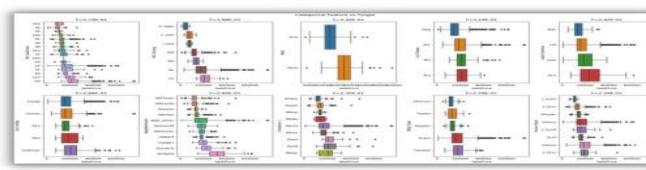
AutoViz



D-Tale



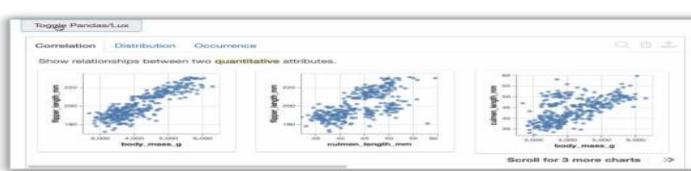
dabl



QuickDA



Lux



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Denkleiers • Leading Minds • Dikgopolo tša Dihlatleli

Understanding the Dataset

Contextualizing Analysis: A deep understanding of the dataset provides context to the data scientist. It helps them comprehend the origin, source, and purpose of the data, enabling them to make informed decisions about how to preprocess, analyze, and interpret the information.

Data Quality Assessment: Understanding the dataset allows data scientists to assess the quality and reliability of the data. They can identify issues such as missing values, outliers, duplicates, and inconsistencies, which must be addressed before meaningful analyses can take place.

Feature Selection and Engineering: A thorough understanding of the dataset aids in selecting the most relevant features (variables) for analysis. It also inspires the creation of new features through feature engineering, potentially enhancing the performance of predictive models.

Bias and Limitations: By understanding the dataset's characteristics, data scientists can identify potential biases, limitations, and sources of noise. This awareness is crucial for interpreting results accurately and avoiding erroneous conclusions.

Optimal Analytical Approaches: Different datasets require different analytical techniques. Understanding the dataset helps data scientists choose appropriate statistical methods, machine learning algorithms, and visualization strategies that are best suited to reveal insights from the data.

Domain-Specific Insights: Understanding the dataset in the context of the domain it represents allows data scientists to uncover meaningful insights that might not be apparent through analysis alone. This domain expertise helps interpret the results in a way that aligns with real-world scenarios.



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Denkleiers • Leading Minds • Dikgopololo tša Dihlatlefi

Loading and inspecting data using Pandas

```
import pandas as pd

# Load the dataset
url = "https://raw.githubusercontent.com/OpenAI/data/master/uris/uris.csv"
uris_df = pd.read_csv(url)

# Display the first few rows of the dataset
print("First 5 rows of the dataset:")
display(uris_df.head())

# Display summary information about the dataset
print("\nSummary information:")
display(uris_df.info())

# Display basic statistics of the dataset
print("\nBasic statistics:")
display(uris_df.describe())
```



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA
Denkleiers • Leading Minds • Dikgopololo tša Dihlatlefi

Inspecting data using Pandas

	user_id	uri
0	0	https://openai.com/research/
1	0	https://github.com/openai/data/blob/master/URI...
2	0	https://arxiv.org/abs/1708.04729
3	0	https://arxiv.org/abs/1911.07820v2
4	0	https://arxiv.org/abs/1911.07820

First 5 rows of the dataset:

	user_id	uri
0	0	https://openai.com/research/
1	0	https://github.com/openai/data/blob/master/URI...
2	0	https://arxiv.org/abs/1708.04729
3	0	https://arxiv.org/abs/1911.07820v2
4	0	https://arxiv.org/abs/1911.07820



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Denkleiers • Leading Minds • Dikgopololo tša Dihlatlefi

Overview of dataset structure, columns, and data types

Summary information:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 732838 entries, 0 to 732837
Data columns (total 2 columns):
 #   Column   Non-Null Count   Dtype  
 --- 
  0   user_id    732838 non-null    int64  
  1   uri        732838 non-null    object 
dtypes: int64(1), object(1)
memory usage: 11.2+ MB
```

Basic statistics:

	user_id
count	732838.000000
mean	19814.113824
std	11497.468144
min	0.000000
25%	9464.000000
50%	19375.000000
75%	30288.000000
max	39646.000000



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Denkleiers • Leading Minds • Dikgopololo tša Dihlatlefi

Loading and inspecting data using Pandas

```
In [128]: df.describe() #explain numerical column frm dataset
```

Out[128]:

	50	1	1.1	500	5	5061
count	207533.000000	207533.000000	207533.000000	207533.000000	207533.000000	207533.000000
mean	21.520009	2.377930	35.497988	14179.514265	2283.817509	5064.014696
std	15.863390	2.883349	116.785406	26435.795386	2667.948833	2.722092
min	-10.000000	1.000000	1.000000	1.000000	1.000000	5061.000000
25%	8.000000	1.000000	9.000000	454.000000	365.000000	5062.000000
50%	19.000000	1.000000	13.000000	2360.000000	1115.000000	5062.000000
75%	32.000000	2.000000	30.000000	18454.000000	3484.000000	5066.000000
max	96.000000	12.000000	1980.000000	126379.000000	12360.000000	5068.000000

```
In [129]: df.info() #datatype formation
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 207533 entries, 0 to 207532
Data columns (total 14 columns):
 #   Column      Non-Null Count  Dtype  
--- 
 0   50          207533 non-null   int64  
 1   TCP          207533 non-null   object  
 2   A            207533 non-null   object  
 3   WannaCry    207533 non-null   object  
 4   1            207533 non-null   int64  
 5   1DA11mPS    207533 non-null   object  
 6   1BonuSr7    207533 non-null   object  
 7   1.1          207533 non-null   int64  
 8   500          207533 non-null   int64  
 9   5            207533 non-null   int64  
 10  A.1          207533 non-null   object  
 11  Bonet        207533 non-null   object  
 12  5061         207533 non-null   int64
```



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Denkleiers • Leading Minds • Dikgopololo tša Dihlatlefi

Data Cleaning and Preprocessing

- **Handling Missing Values, Duplicates, and Outliers:**
- **Missing Values:** When data points are missing, it can lead to skewed analysis and biased results. Handling missing values involves imputing or removing them. For example, in a dataset of survey responses, some participants might not have provided their age. You can impute missing ages with the median age of the respondents.
- **Duplicates:** Duplicate entries can distort analysis and lead to overrepresentation. Detecting and removing duplicates ensures data integrity. For instance, in an e-commerce dataset, there might be multiple identical orders due to system errors. Removing duplicates ensures accurate order counts.
- **Outliers:** Outliers are data points that significantly deviate from the norm. Outliers can affect statistical measures and model performance. Addressing outliers may involve removing or transforming them. In a dataset of exam scores, a single exceptionally high score might be an outlier. You can replace it with a more reasonable value based on the distribution of scores.



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA
Denkleiers • Leading Minds • Dikgopololo tša Dihlatlefi

Data Transformation and Normalization Techniques

- Data Transformation: Data transformation involves converting data into a more suitable format or scale. For instance, transforming skewed data using logarithms can help achieve a more normal distribution. In a dataset of income levels, applying a logarithmic transformation can reduce the impact of extreme incomes.
- Normalization: Normalization scales data to a common range, making comparisons meaningful. In machine learning algorithms, normalized data prevents certain features from dominating others. For example, consider a dataset with attributes like age (0-100) and income (0-100000). Normalizing these features to a 0-1 range ensures balanced contributions to the analysis.
- Standardization: Standardization scales data to have zero mean and unit variance. This is particularly useful for algorithms that assume normally distributed data. For instance, in a dataset containing height and weight, standardization ensures that both attributes contribute equally to clustering algorithms.
- Encoding Categorical Variables: Machine learning models often require numerical inputs. Categorical variables (e.g., "red," "blue," "green") need to be encoded into numerical values.



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA
Denkleiers • Leading Minds • Dikgopolo tša Dihlatleli

Data Transformation and Normalization Techniques

	A	B	C	D	E	F	G	H	I
1	50,"TCP","A","WannaCry","1","1DA11mPS","1BonuSr7","1","500","5","A","Bonet","5061","SS"								
2	40,"TCP","A","WannaCry","1","1DA11mPS","1BonuSr7","1","504","8","A","Bonet","5061","SS"								
3	30,"TCP","A","WannaCry","1","1DA11mPS","1BonuSr7","1","508","7","A","Bonet","5061","SS"								
4	20,"TCP","A","WannaCry","1","1DA11mPS","1BonuSr7","1","512","15","A","Bonet","5061","SS"								
5	57,"TCP","A","WannaCry","1","1DA11mPS","1BonuSr7","1","516","9","A","Bonet","5061","SS"								
6	41,"TCP","A","WannaCry","1","1DA11mPS","1BonuSr7","1","520","17","A","Bonet","5061","SS"								
7	22,"TCP","A","WannaCry","1","1DA11mPS","1BonuSr7","1","524","11","A","Bonet","5061","SS"								
8	18,"TCP","A","WannaCry","1","1DA11mPS","1BonuSr7","1","528","19","A","Bonet","5061","SS"								
9	3,"TCP","A","WannaCry","1","1DA11mPS","1BonuSr7","1","532","13","A","Bonet","5061","SS"								
10	26,"TCP","A","WannaCry","1","1DA11mPS","1BonuSr7","1","536","21","A","Bonet","5061","SS"								
11	22,"TCP","A","WannaCry","1","1DA11mPS","1BonuSr7","1","540","15","A","Bonet","5061","SS"								
12	7,"TCP","A","WannaCry","1","1DA11mPS","1BonuSr7","1","544","23","A","Bonet","5061","SS"								
13	30,"TCP","A","WannaCry","1","1DA11mPS","1BonuSr7","1","548","17","A","Bonet","5061","SS"								
14	26,"TCP","A","WannaCry","1","1DA11mPS","1BonuSr7","1","552","25","A","Bonet","5061","SS"								
15	11,"TCP","A","WannaCry","1","1DA11mPS","1BonuSr7","1","556","19","A","Bonet","5068","SS"								
16	34,"TCP","A","WannaCry","1","1DA11mPS","1BonuSr7","1","560","27","A","Bonet","5068","SS"								
17	30,"TCP","A","WannaCry","1","1DA11mPS","1BonuSr7","1","564","21","A","Bonet","5068","SS"								
18	15,"TCP","A","WannaCry","1","1DA11mPS","1BonuSr7","18","568","29","A","Bonet","5068","SS"								
19	38,"TCP","A","WannaCry","1","1DA11mPS","1BonuSr7","18","572","23","A","Bonet","5068","SS"								
20	34,"TCP","A","WannaCry","1","1DA11mPS","1BonuSr7","18","576","31","A","Bonet","5068","SS"								
21	19,"TCP","A","WannaCry","1","1DA11mPS","1BonuSr7","18","580","25","A","Bonet","5068","SS"								
22	42,"TCP","A","WannaCry","1","1DA11mPS","1BonuSr7","18","584","33","A","Bonet","5068","SS"								
23	38,"TCP","A","WannaCry","1","1DA11mPS","1BonuSr7","18","588","27","A","Bonet","5068","SS"								
24	23,"TCP","A","WannaCry","1","1DA11mPS","1BonuSr7","18","592","35","A","Bonet","5068","SS"								
25	46,"TCP","A","WannaCry","1","1DA11mPS","1BonuSr7","18","596","29","A","Bonet","5068","SS"								
26	42,"TCP","A","WannaCry","1","1DA11mPS","1BonuSr7","18","600","37","A","Bonet","5068","SS"								
27	27,"TCP","A","WannaCry","1","1DA11mPS","1BonuSr7","18","604","31","A","Bonet","5066","SS"								
28	50,"TCP","A","WannaCry","1","1DA11mPS","1BonuSr7","18","608","39","A","Bonet","5066","SS"								
29	46,"TCP","A","WannaCry","1","1DA11mPS","1BonuSr7","18","612","33","A","Bonet","5066","SS"								



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA
Denkleiers • Leading Minds • Dikgopolo tsa Dihlaletsi

Data Transformation and Normalization Techniques

```
In [146]: from sklearn import preprocessing #preprocessing describe, nominal to numeric
```

```
In [147]: lab_encoder = preprocessing.LabelEncoder() # transformation of textual to numeric
df['Protcol'] = lab_encoder.fit_transform(df['Protcol'])
df['Flag'] = lab_encoder.fit_transform(df['Flag'])
df['Family'] = lab_encoder.fit_transform(df['Family'])

df['SeddAddress'] = lab_encoder.fit_transform(df['SeddAddress'])
df['ExpAddress'] = lab_encoder.fit_transform(df['ExpAddress'])
df['IPaddress'] = lab_encoder.fit_transform(df['IPaddress'])
df['Threats'] = lab_encoder.fit_transform(df['Threats'])
df['Prediction'] = lab_encoder.fit_transform(df['Prediction'])
```

```
In [148]: df #df is variabl stnd for data variable
```

```
Out[148]:
```

	Time	Protcol	Flag	Family	Clusters	SeddAddress	ExpAddress	BTC	USD	Netflow_Bytes	IPaddress	Threats	Port	Prediction
0	40	1	0	16	1	2	2	1	504	8	0	1	5061	2
1	30	1	0	16	1	2	2	1	508	7	0	1	5061	2
2	20	1	0	16	1	2	2	1	512	15	0	1	5061	2
3	57	1	0	16	1	2	2	1	516	9	0	1	5061	2
4	41	1	0	16	1	2	2	1	520	17	0	1	5061	2
...
207528	12	1	5	7	8	1	6	1964	2986	6081	0	8	5062	0
207529	8	1	5	7	8	1	6	1968	2992	6092	0	8	5062	0
207530	8	1	5	7	8	1	6	1972	2998	6103	0	8	5062	0
207531	8	1	5	7	8	1	6	1976	3004	6114	0	8	5062	0
207532	8	1	5	7	8	1	6	1980	3010	6125	0	8	5062	0

207533 rows × 14 columns



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA
Denkleiers • Leading Minds • Dikgopolo tsa Dihlaletsi

Exploratory Data Analysis (EDA)

- Generating descriptive statistics and visualizations.
- Distribution plots, histograms, and box plots.
- Correlation analysis and scatter plots.

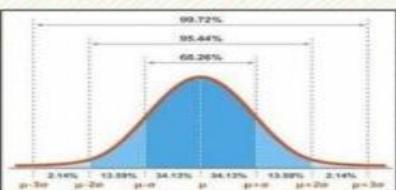


UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA
Denkleiers • Leading Minds • Dikgopolo tsa Dihlaletsi

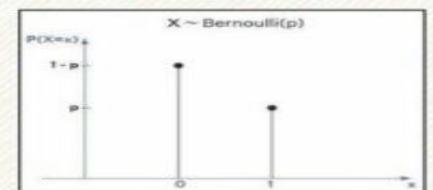
Exploratory Data Analysis (EDA)

Most Important Distributions in Data Science

Normal Distribution



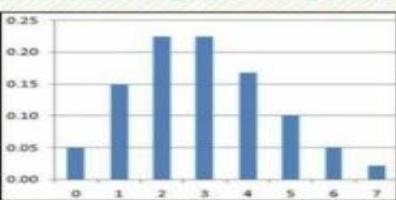
Bernoulli Distribution



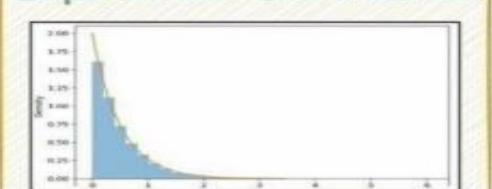
Binomial Distribution



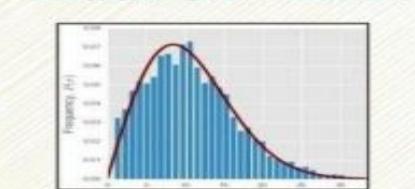
Poisson Distribution



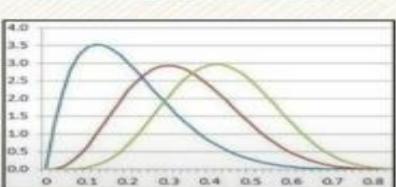
Exponential Distribution



Gamma Distribution



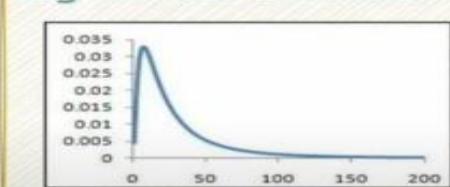
Beta Distribution



Uniform Distribution



Log Normal Distribution



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Denkleiers • Leading Minds • Dikgopolo tsa Dihalefe

Data Processing Framework

Import Libraries

- Numpy
- Pandas
- Matplotlib
- Import dataset

Read/Load dataset

- Generate EDA

Provide distribution insights of data

- df.value_counts()
 - df.shape()
- df.describe()
 - df.info()
- df.isnull().sum()

Specify Name of Columns

- df.columns=["x", "y", "z"]
- df["x"].value_counts()

Visualization

- Seaborn
- Scatterplot
- Piechart

Data Processing Framework: Data Transformation for Machine Learning

Data Transformation

- Sklearn.preprocessing

Cross validation

- Training Vs Testing

Machine Learning results

- Accuracy
- Precision
- Recall
- F1 score

Visualization of Machine Learning results

Visualization

- Seaborn
- Scatterplot
- Piechart

Role of visualization in understanding data patterns

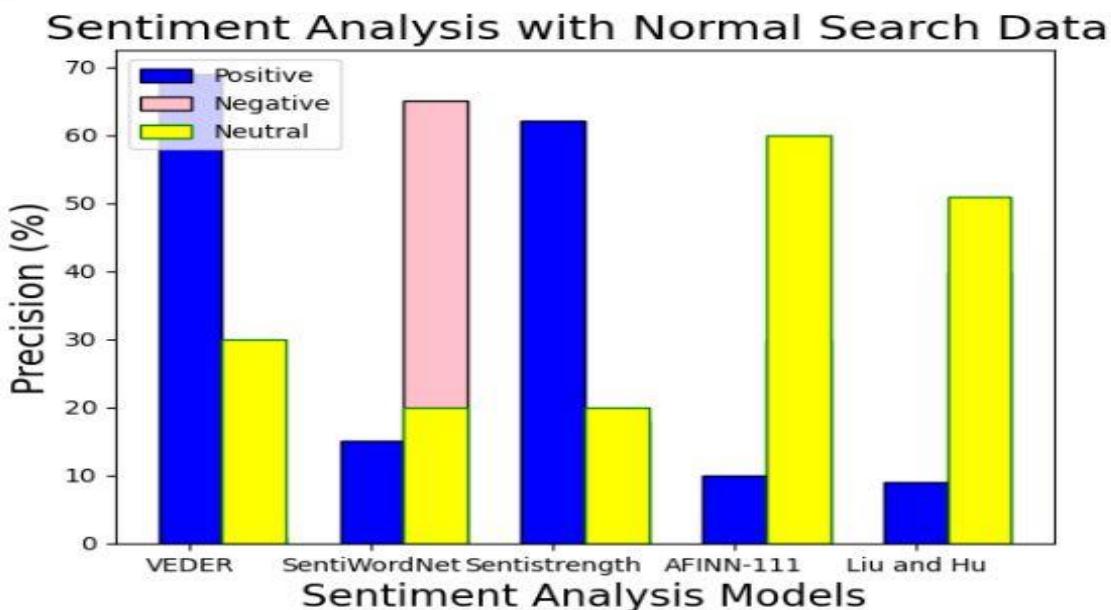
```
In [43]: import matplotlib.pyplot as plt
import numpy as np

city=['VEDER','SentiWordNet','Sentistrength','AFINN-111','Liu and Hu']
Gender=['Positive','Negative', 'Neutral']
pos = np.arange(len(city))
bar_width = 0.35

posit=[69,15,62,10,9]
neg=[1,65,18,30,40]
netr=[30,20,20,60,51]

plt.bar(pos,posit,bar_width,color='blue',edgecolor='black')
plt.bar(pos+bar_width,neg,bar_width,color='pink',edgecolor='black')
plt.bar(pos+bar_width,netr,bar_width,color='yellow',edgecolor='green')

plt.xticks(pos, city)
plt.xlabel('Sentiment Analysis Models', fontsize=16)
plt.ylabel('Precision (%)', fontsize=16)
plt.title('Sentiment Analysis with Normal Search Data ',fontsize=18)
plt.legend(Gender,loc=2)
plt.show()
```



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA
Denkleiers • Leading Minds • Dikgopolo tša Dihlaletsi

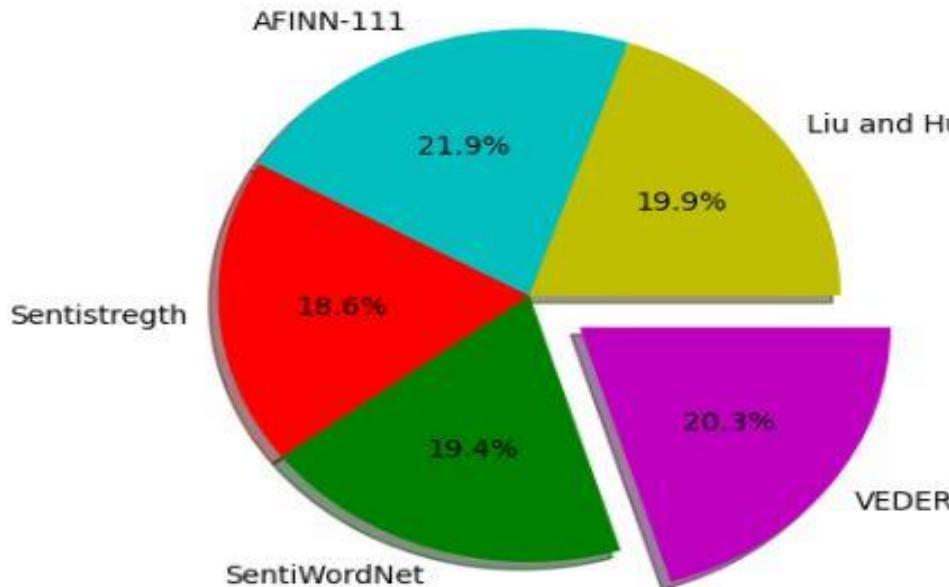
Role of visualization in understanding data patterns

In [2]:

```
import matplotlib.pyplot as plt

values = [2657, 2543, 2436, 2873, 2603]
colors = ['m', 'g', 'r', 'c', 'y']
labels = ['VEDER', 'SentiWordNet', 'Sentistrength', 'AFINN-111', 'Liu and Hu']
explode = (0.2, 0, 0, 0, 0)
plt.pie(values, colors=colors, labels=labels,
explode=explode, autopct='%1.1f%%',
counterclock=False, shadow=True)
plt.title('Unclassified Words Extracted with a Normal Search')
plt.show()
```

Unclassified Words Extracted with a Normal Search



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA
Denkleiers • Leading Minds • Dikgopolo tša Dihlaletsi

Case Study and Hands-on Practice

Real-world Data Exploration:

- Applying data exploration and visualization techniques to a real dataset.
- Exploring patterns, trends, and insights.

• Hands-on Jupyter Notebook Exercises:

- Guided exercises using Jupyter Notebook, Pandas, Matplotlib, Seaborn, and Plotly.
- Data cleaning, visualization, and interpretation.



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA
Denkleiers • Leading Minds • Dikgopolo tša Dihlatlefi

```
In [44]: from sklearn.model_selection import train_test_split # lib for ml models
         from sklearn.preprocessing import StandardScaler # normalizatn

#selects all rows from the DataFrame df and all columns except for the last column.
#This operation effectively creates a new DataFrame X containing all the data except the last column,
#which is often the target variable or label when working with machine learning datasets.
#This new DataFrame X typically contains the features that will be used for training and testing your machine learning models.

X = df.iloc[:, :-1] #in x features are adding except last column
y = df.iloc[:, -1] #saving last column of dataset in y
X_train, X_test, y_train, y_test = train_test_split(X, y, train_size = 0.8, random_state = 42) #split test and train into 4 parts
```

```
In [45]: X_train
```

```
Out[45]:
```

	Type1	Authority	Consistency	Liking	Scarcity	Reciprocity
65	3	0	0	1	0	0
67	3	0	0	1	0	0
31	6	1	0	0	1	1
12	6	0	0	1	0	0
41	6	0	0	1	0	0
...
106	3	0	0	1	0	0
14	6	1	0	1	1	1
92	3	0	0	1	0	0
179	0	0	0	1	0	1
102	3	0	0	1	0	0

147 rows × 6 columns

```
In [48]:  
from sklearn.ensemble import RandomForestClassifier  
from sklearn.svm import LinearSVC  
from sklearn.naive_bayes import GaussianNB  
from sklearn.model_selection import train_test_split  
  
from sklearn.ensemble import StackingClassifier #ensabl method of stacking classify for  
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score  
from sklearn.metrics import confusion_matrix  
from sklearn.metrics import classification_report  
  
from genetic_selection import GeneticSelectionCV #function to import gA -> cross valia  
from sklearn.tree import DecisionTreeClassifier #estimator in GA  
import numpy as np  
  
import warnings  
warnings.filterwarnings('ignore')  
  
In [49]: rf = RandomForestClassifier(n_estimators=100, random_state=42) # 100 means no. of trees  
  
In [50]: rf.fit(X_train, y_train)  
Out[50]: RandomForestClassifier(random_state=42)  
  
In [51]: rf_pred=rf.predict(X_test)  
  
In [52]: rf_accuracy = accuracy_score(rf_pred, y_test)  
rf_report = classification_report(rf_pred, y_test)  
rf_matrix = confusion_matrix(rf_pred, y_test)  
print('Accuracy of Random Forest : ', round(rf_accuracy, 3))  
print('Classification report of Random Forest : \n', rf_report)  
print('Confusion Matrix of Random Forest : \n', rf_matrix)  
  
Accuracy of Random Forest : 0.973  
Classification report of Random Forest :  
precision recall f1-score support  
0 1.00 0.96 0.98 23  
1 0.93 1.00 0.97 14  
  
accuracy 0.97 0.97 0.97 37  
macro avg 0.97 0.98 0.97 37  
weighted avg 0.97 0.97 0.97 37  
  
Confusion Matrix of Random Forest :  
[[22 1]  
[ 0 14]]
```



Case Study and Hands-on Practice

```
In [55]: # Precision, recall, and f1-score values
precision = [1.00, 0.93]
recall = [0.96, 1.00]
f1_score = [0.98, 0.97]
class_labels = ['Class 0', 'Class 1']

# Create a bar chart
fig, ax = plt.subplots(figsize=(8, 6))
x = range(len(class_labels))

# Plot precision, recall, and f1-score
ax.bar(x, precision, width=0.2, label='Precision')
ax.bar([i + 0.2 for i in x], recall, width=0.2, label='Recall')
ax.bar([i + 0.4 for i in x], f1_score, width=0.2, label='F1-Score')

# Add data labels
for i in x:
    ax.text(i, precision[i] + 0.01, f'{precision[i]:.2f}', ha='center')
    ax.text(i + 0.2, recall[i] + 0.01, f'{recall[i]:.2f}', ha='center')
    ax.text(i + 0.4, f1_score[i] + 0.01, f'{f1_score[i]:.2f}', ha='center')

# Set x-axis ticks and labels
ax.set_xticks([i + 0.2 for i in x])
ax.set_xticklabels(class_labels)
ax.set_xlabel('Class or Category')
ax.set_ylabel('Random Forest Score')
ax.set_title('Precision, Recall, and F1-Score for Each Class')
ax.legend()

plt.show()
```



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

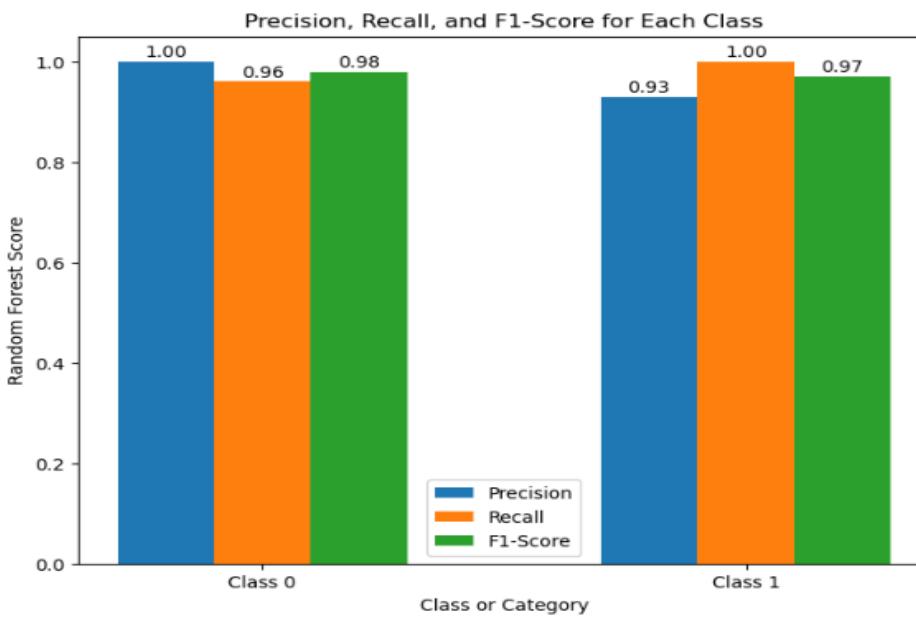
Denkleiers • Leading Minds • Dikgopolo tša Dihlatlefi

Case Study and Hands-on Practice

```
# Add data labels
for i in x:
    ax.text(i, precision[i] + 0.01, f'{precision[i]:.2f}', ha='center')
    ax.text(i + 0.2, recall[i] + 0.01, f'{recall[i]:.2f}', ha='center')
    ax.text(i + 0.4, f1_score[i] + 0.01, f'{f1_score[i]:.2f}', ha='center')

# Set x-axis ticks and labels
ax.set_xticks([i + 0.2 for i in x])
ax.set_xticklabels(class_labels)
ax.set_xlabel('Class or Category')
ax.set_ylabel('Random Forest Score')
ax.set_title('Precision, Recall, and F1-Score for Each Class')
ax.legend()

plt.show()
```



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Denkleiers • Leading Minds • Dikgopolo tsa Dihlaletsi



ENGINEERING 4.0
UNIVERSITY OF PRETORIA



INF 491/791: Applied Data Science



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Make today matter

www.up.ac.za

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

L 04: Applied Statistical Analysis



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Make today matter

www.up.ac.za

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Outline

- ❖ What is Applied Statistical Analysis?
- ❖ What are key aspects of Applied Statistical Analysis?
- ❖ Gather Data
 - ❖ Data points and features
 - ❖ Cleaning data
 - ❖ Data exploration
 - ❖ Visualising Data
 - ❖ Descriptive Statistics
 - ❖ Correlation
 - ❖ Training & Test Dataset Split
 - ❖ Multivariable Linear Regression
 - ❖ Data Transformations
 - ❖ P values & Evaluating Coefficients
- ❖ Admin



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology
Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

What is Applied Statistical Analysis?



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

**Faculty of Engineering,
Built Environment and
Information Technology**

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za

What is Applied Statistical Analysis?

- "Applied statistical analysis refers to the use of **statistical methods** and **techniques** to analyze **real-world** data in order to **make informed decisions, draw conclusions, and solve practical problems**. It involves the application of statistical principles to various fields such as science, engineering, business, healthcare, social sciences, and more. The **primary goal** of applied statistical analysis is to **extract meaningful information** from data, **uncover patterns** and trends, **test hypotheses**, and **make predictions** or recommendations based on the data." - ChatGPT



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology
Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter
www.up.ac.za

What is Applied Statistical Analysis? (continued...)

- "Applied statistics is the **root** of data analysis. The **practice of applied statistics** involves **analyzing data** to *help define and determine business needs*. Modern workplaces are overwhelmed with big data and are looking for statisticians, data analysts, data scientists, and other professionals with applied statistics knowledge who can **organize, analyze, and use data to solve real-world problems.**" - Michigan Tech



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter
www.up.ac.za

What are key aspects of Applied Statistical Analysis?



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za

What are key aspects of Applied Statistical Analysis?

- Data Collection
- Data Cleaning and Preprocessing
- Descriptive Statistics
- Exploratory Data Analysis
- Inferential Statistics
- Regression Analysis
- ...
- ...
- ...
- Data Visualization
- ...
- Interpretation and Reporting



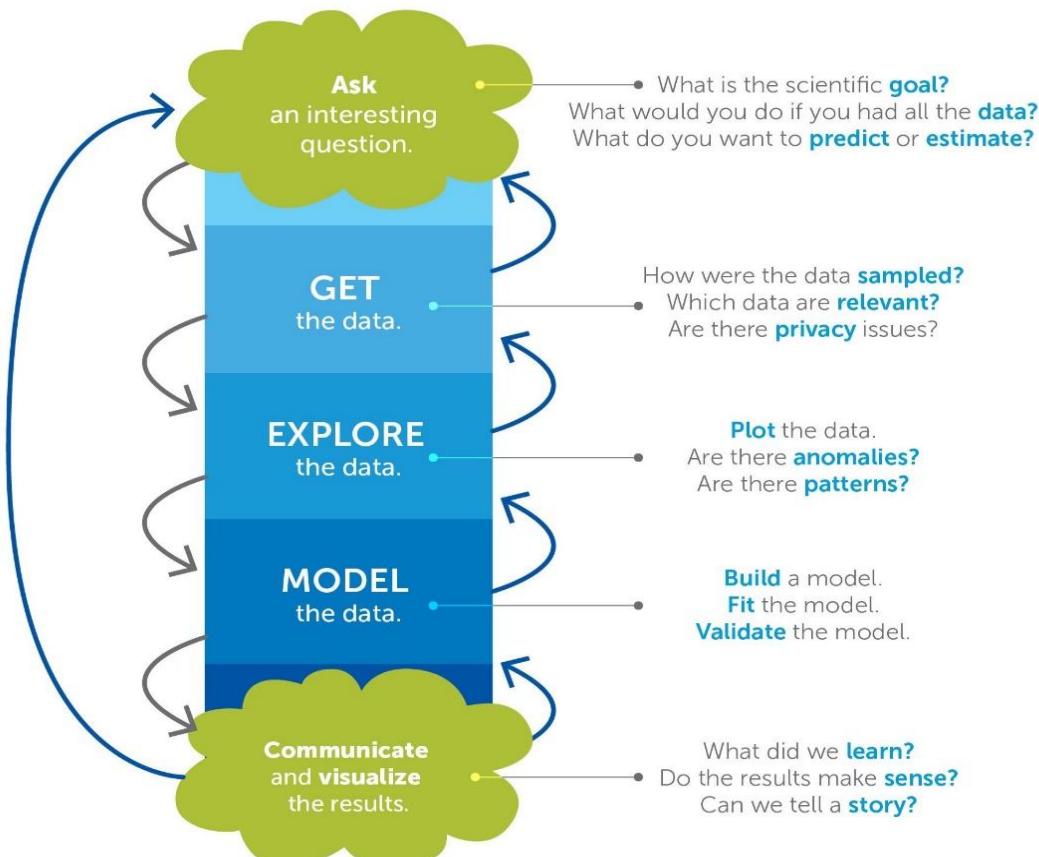
UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

What are key aspects of Applied Statistical Analysis? (continued...)

The Data Science Process



Derived from the work of Joe Blitzstein and Hanspeter Pfister,
originally created for the Harvard data science course <http://cs109.org/>.



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za

Gather Data



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

**Faculty of Engineering,
Built Environment and
Information Technology**

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za

Gather Data

Least Angle Regression

Bradley Efron, Trevor Hastie, Iain Johnstone and Robert Tibshirani
Statistics Department, Stanford University

January 9, 2003

← → C 🔒 www4.stat.ncsu.edu/~boos/var.select/diabetes.tab.txt

AGE	SEX	BMI	BP	S1	S2	S3	S4	S5	S6	Y
59	2	32.1	101	157	93.2	38	4	4.8598	87	151
48	1	21.6	87	183	103.2	70	3	3.8918	69	75
72	2	30.5	93	156	93.6	41	4	4.6728	85	141
24	1	25.3	84	198	131.4	40	5	4.8903	89	206
50	1	23	101	192	125.4	52	4	4.2905	80	135
23	1	22.6	89	139	64.8	61	2	4.1897	68	97
36	2	22	90	160	99.6	50	3	3.9512	82	138
66	2	26.2	114	255	185	56	4.55	4.2485	92	63
60	2	32.1	83	179	119.4	42	4	4.4773	94	110
29	1	30	85	180	93.4	43	4	5.3845	88	310
22	1	18.6	97	114	57.6	46	2	3.9512	83	101
56	2	28	85	184	144.8	32	6	3.5835	77	69
53	1	23.7	92	186	109.2	62	3	4.3041	81	179
50	2	26.2	97	186	105.4	49	4	5.0626	88	185
61	1	24	91	202	115.4	72	3	4.2905	73	118
34	2	24.7	118	254	184.2	39	7	5.037	81	171
47	1	30.3	109	207	100.2	70	3	5.2149	98	166
68	2	27.5	111	214	147	39	5	4.9416	91	144
38	1	25.4	84	162	103	42	4	4.4427	87	97
41	1	24.7	83	187	108.2	60	3	4.5433	78	168
35	1	21.1	82	156	87.8	50	3	4.5109	95	68
25	2	24.3	95	162	98.6	54	3	3.8501	87	49
25	1	26	92	187	120.4	56	3	3.9703	88	68
61	2	22	102	67	210	95	2	6.107	124	245



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za

Gather Data (continued...)

The screenshot shows a Jupyter Notebook interface with two tabs: 'diabetes_dataset.csv' and 'Diabetes_study.ipynb'. The 'diabetes_dataset.csv' tab displays a CSV file with 14 rows of data. The columns are labeled AGE, SEX, BMI, BP, S1, S2, S3, S4, and S5. The data includes various numerical values such as age (e.g., 59, 48, 72), sex (0 or 1), BMI (e.g., 32.1, 21.6, 30.5), blood pressure (e.g., 101, 87, 93), and other physiological measurements. The 'S3' column for row 3 is highlighted with a blue border.

	AGE	SEX	BMI	BP	S1	S2	S3	S4	S5
1	59	2	32.1	101	157	93.2	38	4	4.8598
2	48	1	21.6	87	183	103.2	70	3	3.8918
3	72	2	30.5	93	156	93.6	41	4	4.6728
4	24	1	25.3	84	198	131.4	40	5	4.8903
5	50	1	23	101	192	125.4	52	4	4.2905
6	23	1	22.6	89	139	64.8	61	2	4.1897
7	36	2	22	90	160	99.6	50	3	3.9512
8	66	2	26.2	114	255	185	56	4.55	4.2485
9	60	2	32.1	83	179	119.4	42	4	4.4773
10	29	1	30	85	180	93.4	43	4	5.3845
11	22	1	18.6	97	114	57.6	46	2	3.9512
12	56	2	28	85	184	144.8	32	6	3.5835
13	53	1	23.7	92	186	109.2	62	3	4.3041
14	60	2	26.2	97	186	105.4	40	4	5.0626

```
import pandas as pd  
data = pd.read_csv('diabetes_dataset.csv')
```



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter
www.up.ac.za

Data points and features



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

**Faculty of Engineering,
Built Environment and
Information Technology**

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za

Data points and features

```
type(data)
```

```
pandas.core.frame.DataFrame
```

```
data.shape
```

```
(442, 11)
```

10 Features in order:

- AGE age in years
- SEX sex
- BMI body mass index
- BP average blood pressure
- S1 tc, total serum cholesterol
- S2 ldl, low-density lipoproteins
- S3 hdl, high-density lipoproteins
- S4 tch, total cholesterol / HDL
- S5 ltg, possibly log of serum triglycerides level
- S6 glu, blood sugar level

Target variable:

- Y response



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za

Cleaning data



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

**Faculty of Engineering,
Built Environment and
Information Technology**

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za

Cleaning data

```
pd.isnull(data).any()
```

```
AGE    False
SEX    False
BMI    False
BP     False
S1     False
S2     False
S3     False
S4     False
S5     False
S6     False
Y      False
dtype: bool
```

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 442 entries, 0 to 441
Data columns (total 11 columns):
 #   Column   Non-Null Count  Dtype  
--- 
 0   AGE       442 non-null    int64  
 1   SEX       442 non-null    int64  
 2   BMI       442 non-null    float64 
 3   BP        442 non-null    float64 
 4   S1        442 non-null    int64  
 5   S2        442 non-null    float64 
 6   S3        442 non-null    float64 
 7   S4        442 non-null    float64 
 8   S5        442 non-null    float64 
 9   S6        442 non-null    int64  
 10  Y         442 non-null    int64  
dtypes: float64(6), int64(5)
memory usage: 38.1 KB
```



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Data exploration



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

**Faculty of Engineering,
Built Environment and
Information Technology**

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za

Data exploration

data.head()												
	AGE	SEX	BMI	BP	S1	S2	S3	S4	S5	S6	Y	
0	59	2	32.1	101.0	157	93.2	38.0	4.0	4.8598	87	151	
1	48	1	21.6	87.0	183	103.2	70.0	3.0	3.8918	69	75	
2	72	2	30.5	93.0	156	93.6	41.0	4.0	4.6728	85	141	
3	24	1	25.3	84.0	198	131.4	40.0	5.0	4.8903	89	206	
4	50	1	23.0	101.0	192	125.4	52.0	4.0	4.2905	80	135	

data.tail()												
	AGE	SEX	BMI	BP	S1	S2	S3	S4	S5	S6	Y	
437	60	2	28.2	112.00	185	113.8	42.0	4.00	4.9836	93	178	
438	47	2	24.9	75.00	225	166.0	42.0	5.00	4.4427	102	104	
439	60	2	24.9	99.67	162	106.6	43.0	3.77	4.1271	95	132	
440	36	1	30.0	95.00	201	125.2	42.0	4.79	5.1299	85	220	
441	36	1	19.6	71.00	250	133.2	97.0	3.00	4.5951	92	57	

data.count()	
AGE	442
SEX	442
BMI	442
BP	442
S1	442
S2	442
S3	442
S4	442
S5	442
S6	442
Y	442
dtype:	int64

- Data visualization also help us make sense of the data at the Exploration Stage



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Visualising Data



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

**Faculty of Engineering,
Built Environment and
Information Technology**

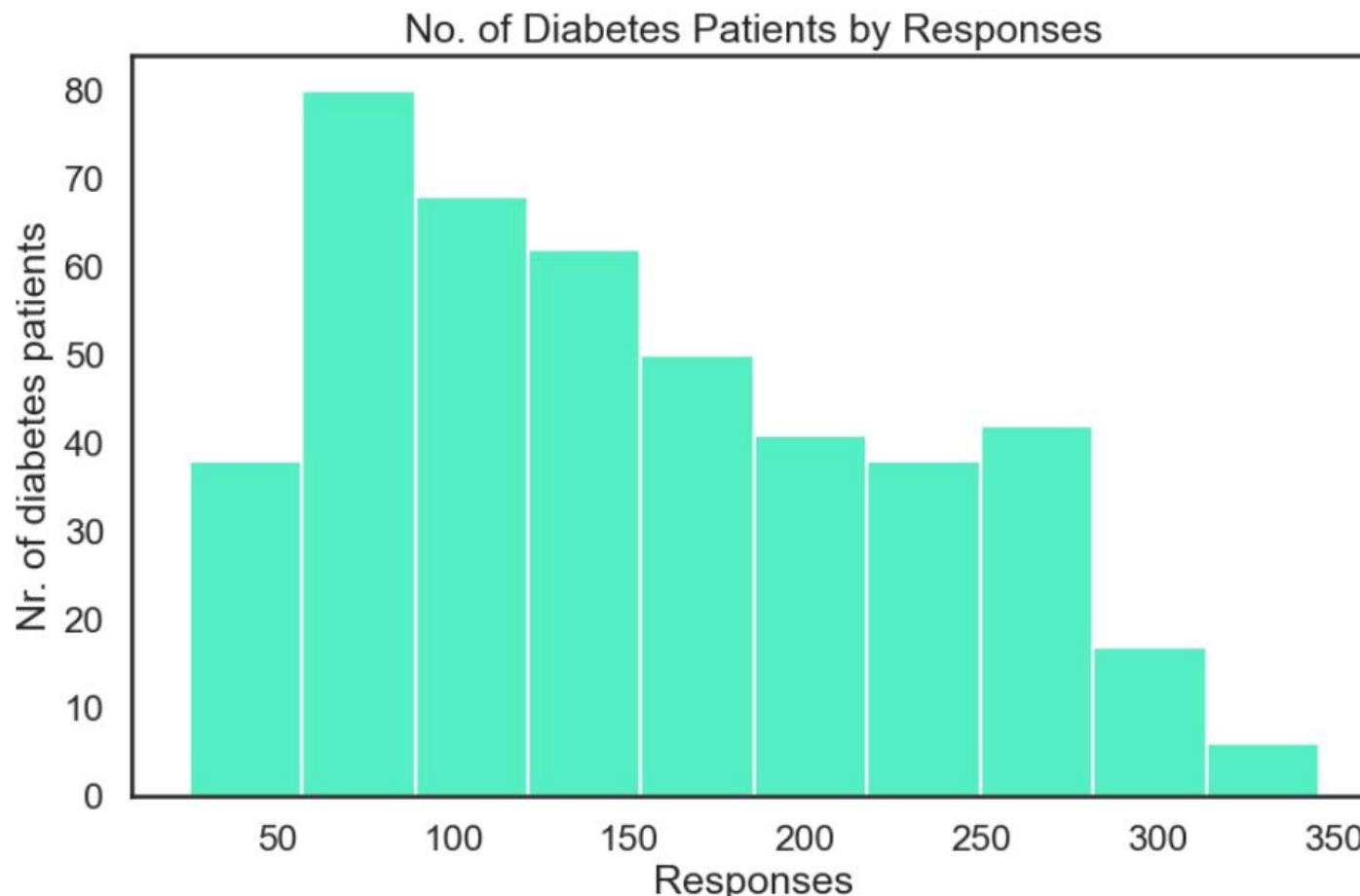
Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za

Visualising Data

```
plt.figure(figsize=(10, 6))
plt.hist(data['Y'], ec='#ffffff', color='#55efc4')
plt.xlabel('Responses')
plt.ylabel('Nr. of diabetes patients')
plt.title('No. of Diabetes Patients by Responses')
plt.show()
```



- Data visualization also help us make sense of the data at the Exploration Stage
- Used for data distribution & outliers
- Useful for e.g., identifying Normal vs Skewed Distributions



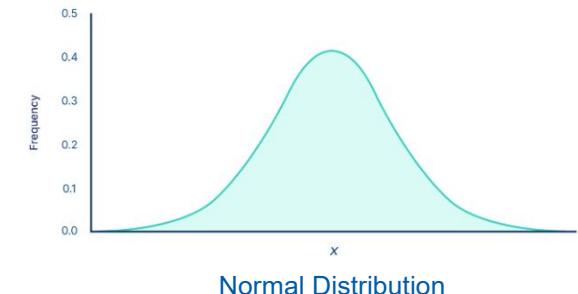
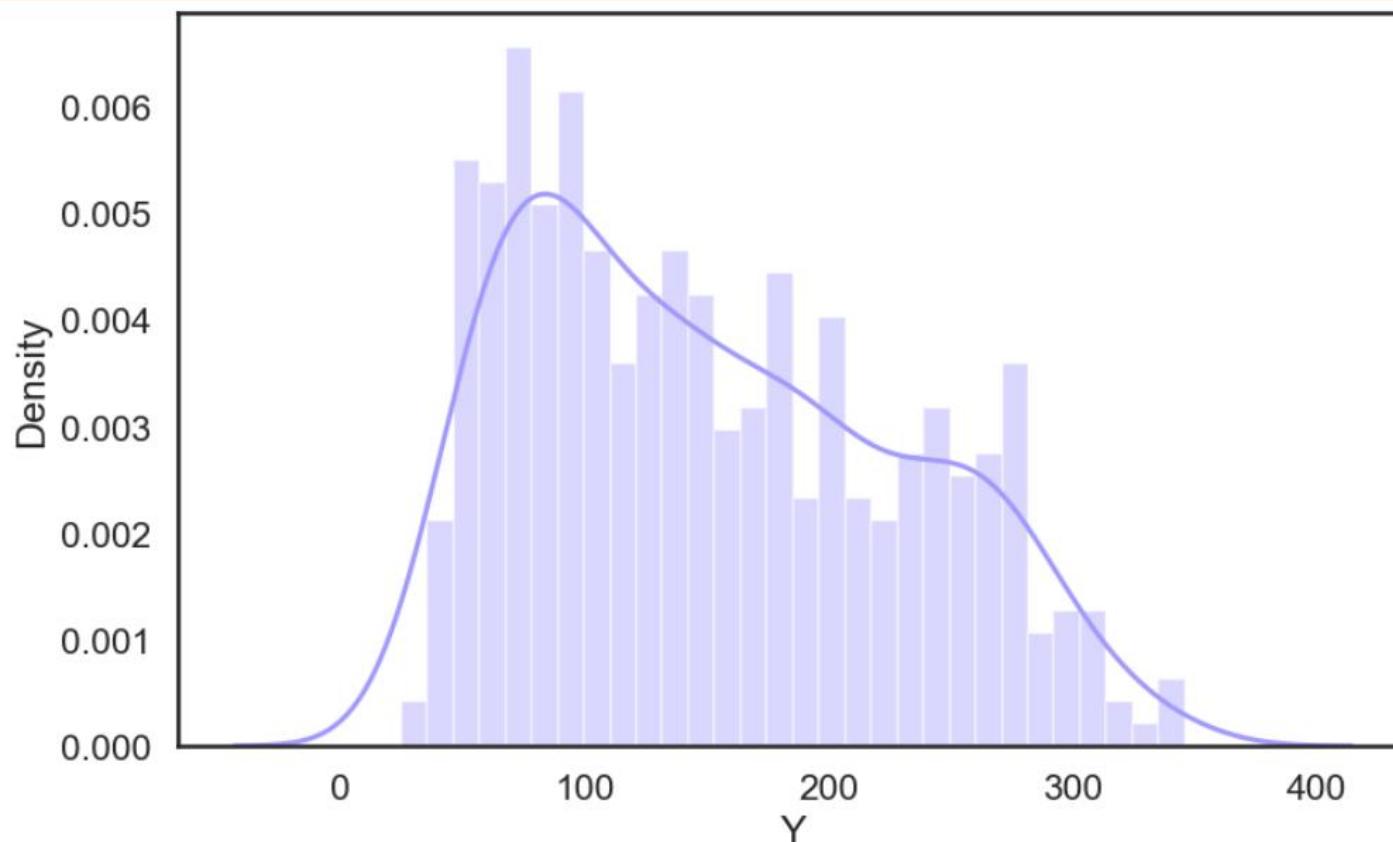
UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Visualising Data (continued...)

```
plt.figure(figsize=(10, 6))
sns.distplot(data['Y'], bins=30, color='#a29bfe')
plt.show()
```



Normal Distribution



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za

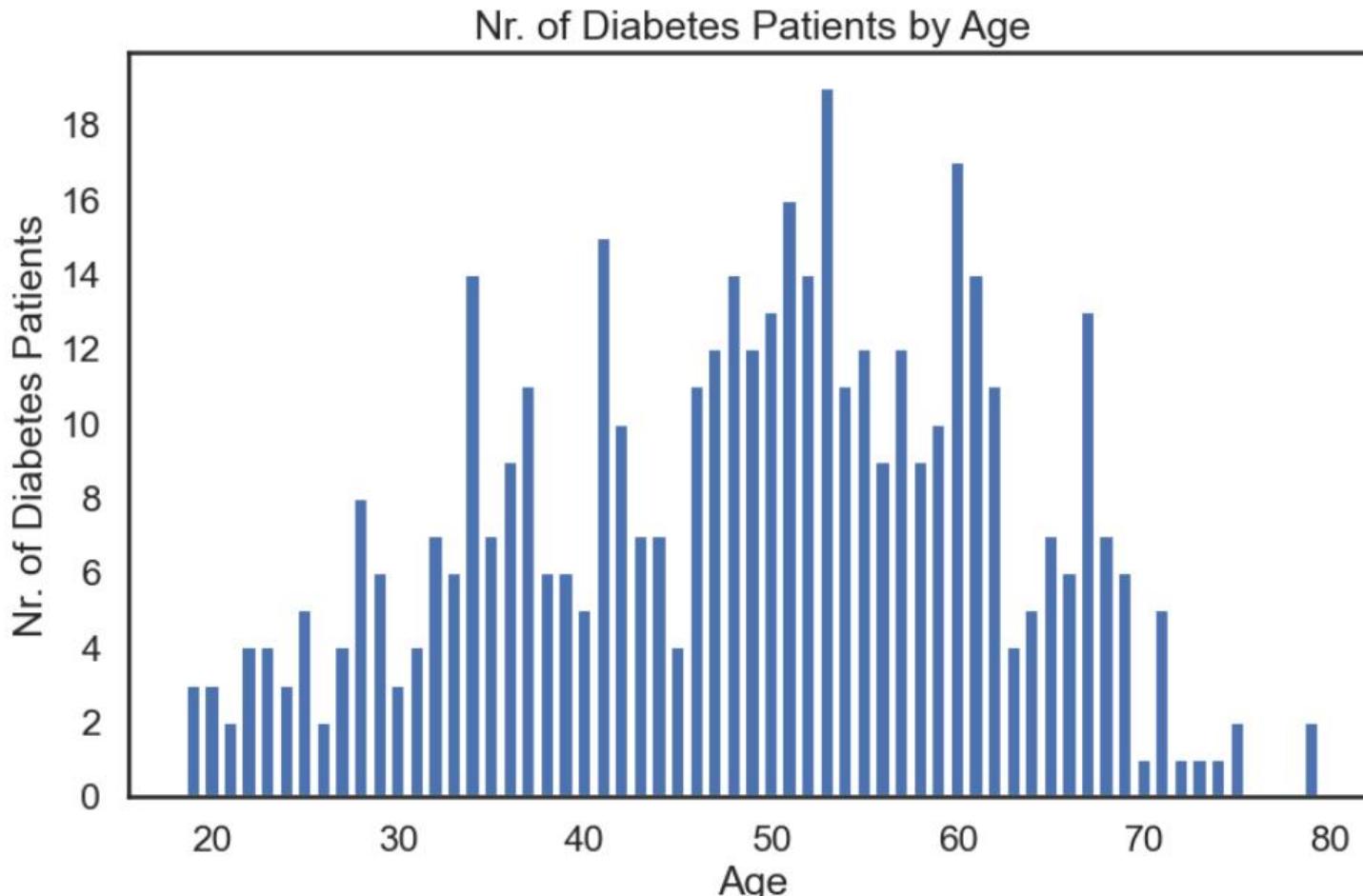
Visualising Data (continued...)

```
data['AGE'].mean()
```

```
48.51809954751131
```

```
data['AGE'].median()
```

```
50.0
```



```
frequency = data['AGE'].value_counts()  
  
plt.figure(figsize=(10, 6))  
plt.xlabel('Age')  
plt.ylabel('Nr. of Diabetes Patients')  
custom_y_ticks = [0, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20]  
plt.yticks(custom_y_ticks)  
plt.title('Nr. of Diabetes Patients by Age')  
plt.bar(frequency.index, height=frequency)  
plt.show()
```



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Descriptive Statistics



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

**Faculty of Engineering,
Built Environment and
Information Technology**

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za

Descriptive Statistics

```
data['Y'].min()
```

25

```
data['Y'].max()
```

346

```
data.min()
```

AGE 19.0000

SEX 1.0000

BMI 18.0000

BP 62.0000

S1 97.0000

S2 41.6000

S3 22.0000

S4 2.0000

S5 3.2581

S6 58.0000

Y 25.0000

dtype: float64

```
data.max()
```

AGE 79.000

SEX 2.000

BMI 42.200

BP 133.000

S1 301.000

S2 242.400

S3 99.000

S4 9.090

S5 6.107

S6 124.000

Y 346.000

dtype: float64

```
data.head()
```

	AGE	SEX	BMI	BP	S1	S2	S3	S4	S5	S6	Y
--	-----	-----	-----	----	----	----	----	----	----	----	---

0	59	2	32.1	101.0	157	93.2	38.0	4.0	4.8598	87	151
---	----	---	------	-------	-----	------	------	-----	--------	----	-----

1	48	1	21.6	87.0	183	103.2	70.0	3.0	3.8918	69	75
---	----	---	------	------	-----	-------	------	-----	--------	----	----

2	72	2	30.5	93.0	156	93.6	41.0	4.0	4.6728	85	141
---	----	---	------	------	-----	------	------	-----	--------	----	-----

3	24	1	25.3	84.0	198	131.4	40.0	5.0	4.8903	89	206
---	----	---	------	------	-----	-------	------	-----	--------	----	-----

4	50	1	23.0	101.0	192	125.4	52.0	4.0	4.2905	80	135
---	----	---	------	-------	-----	-------	------	-----	--------	----	-----

```
data.tail()
```

	AGE	SEX	BMI	BP	S1	S2	S3	S4	S5	S6	Y
--	-----	-----	-----	----	----	----	----	----	----	----	---

437	60	2	28.2	112.00	185	113.8	42.0	4.00	4.9836	93	178
-----	----	---	------	--------	-----	-------	------	------	--------	----	-----

438	47	2	24.9	75.00	225	166.0	42.0	5.00	4.4427	102	104
-----	----	---	------	-------	-----	-------	------	------	--------	-----	-----

439	60	2	24.9	99.67	162	106.6	43.0	3.77	4.1271	95	132
-----	----	---	------	-------	-----	-------	------	------	--------	----	-----

440	36	1	30.0	95.00	201	125.2	42.0	4.79	5.1299	85	220
-----	----	---	------	-------	-----	-------	------	------	--------	----	-----

441	36	1	19.6	71.00	250	133.2	97.0	3.00	4.5951	92	57
-----	----	---	------	-------	-----	-------	------	------	--------	----	----



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

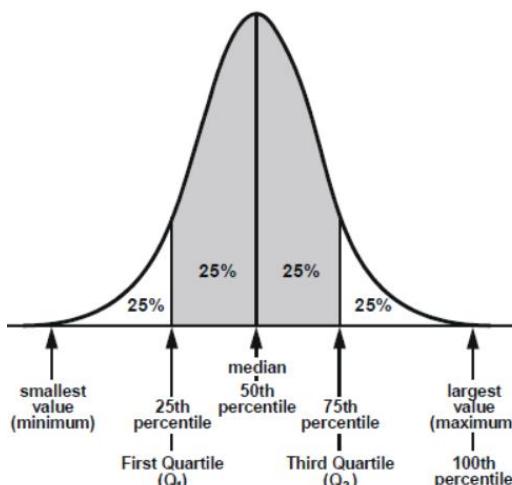
Make today matter

www.up.ac.za

Descriptive Statistics (continued...)

data.describe()													data.mean()		data.median()	
	AGE	SEX	BMI	BP	S1	S2	S3	S4	S5	S6	Y	AGE	48.518100	AGE	50.00000	
count	442.000000	442.000000	442.000000	442.000000	442.000000	442.000000	442.000000	442.000000	442.000000	442.000000	442.000000	SEX	1.468326	SEX	1.00000	
mean	48.518100	1.468326	26.375792	94.647014	189.140271	115.439140	49.788462	4.070249	4.641411	91.260181	152.133484	BMI	26.375792	BMI	25.70000	
std	13.109028	0.499561	4.418122	13.831283	34.608052	30.413081	12.934202	1.290450	0.522391	11.496335	77.093005	BP	94.647014	BP	93.00000	
min	19.000000	1.000000	18.000000	62.000000	97.000000	41.600000	22.000000	2.000000	3.258100	58.000000	25.000000	S1	189.140271	S1	186.00000	
25%	38.250000	1.000000	23.200000	84.000000	164.250000	96.050000	40.250000	3.000000	4.276700	83.250000	87.000000	S2	115.439140	S2	113.00000	
50%	50.000000	1.000000	25.700000	93.000000	186.000000	113.000000	48.000000	4.000000	4.620050	91.000000	140.500000	S3	49.788462	S3	48.00000	
75%	59.000000	2.000000	29.275000	105.000000	209.750000	134.500000	57.750000	5.000000	4.997200	98.000000	211.500000	S4	4.070249	S4	4.00000	
max	79.000000	2.000000	42.200000	133.000000	301.000000	242.400000	99.000000	9.090000	6.107000	124.000000	346.000000	S5	4.641411	S5	4.62005	
												S6	91.260181	S6	91.00000	
												Y	152.133484	Y	140.50000	
												dtype:	float64	dtype:	float64	

- Mean: The average
- Median: The midpoint of the distribution. I.e., the middle value



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Correlation



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

**Faculty of Engineering,
Built Environment and
Information Technology**

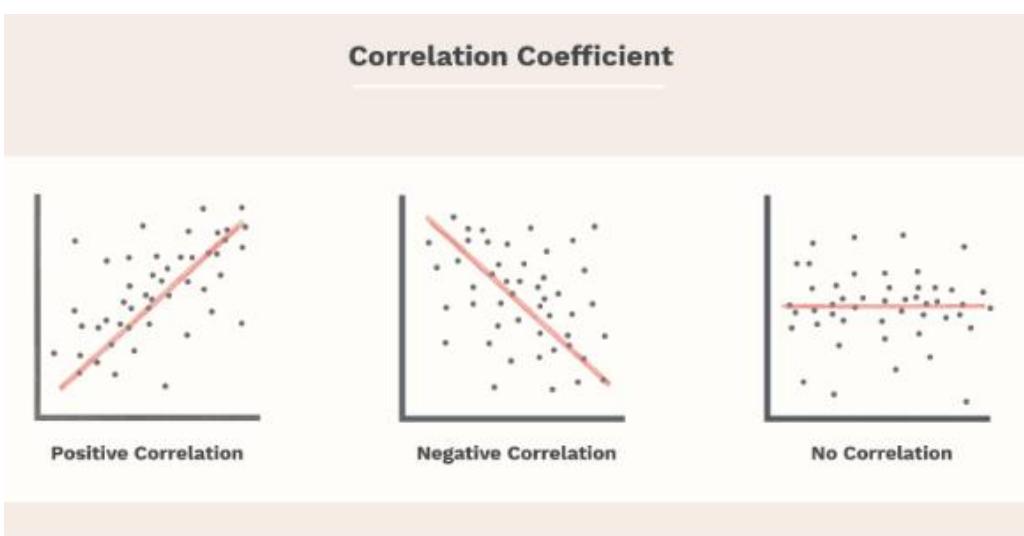
Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za

Correlation

- Correlation is a statistical measure that describes the extent to which two variables change together (i.e., their relationship).
 - 1 is a perfect positive correlation
 - 0 means there is no correlation
 - -1 is a perfect negative correlation
- Correlation (a linear relationship): is important because we want to include features that have the right **strength** and **direction** (i.e., features that are correlated with the target)



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

**Faculty of Engineering,
Built Environment and
Information Technology**

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za

Correlation

```
data['Y'].corr(data['BMI'])
```

```
0.5864501344746885
```

```
data['BMI'].corr(data['AGE'])
```

```
0.18508466614655544
```

```
data.corr()
```

	AGE	SEX	BMI	BP	S1	S2	S3	S4	S5	S6	Y
AGE	1.000000	0.173737	0.185085	0.335428	0.260061	0.219243	-0.075181	0.203841	0.270774	0.301731	0.187889
SEX	0.173737	1.000000	0.088161	0.241010	0.035277	0.142637	-0.379090	0.332115	0.149916	0.208133	0.043062
BMI	0.185085	0.088161	1.000000	0.395411	0.249777	0.261170	-0.366811	0.413807	0.446157	0.388680	0.586450
BP	0.335428	0.241010	0.395411	1.000000	0.242464	0.185548	-0.178762	0.257650	0.393480	0.390430	0.441482
S1	0.260061	0.035277	0.249777	0.242464	1.000000	0.896663	0.051519	0.542207	0.515503	0.325717	0.212022
S2	0.219243	0.142637	0.261170	0.185548	0.896663	1.000000	-0.196455	0.659817	0.318357	0.290600	0.174054
S3	-0.075181	-0.379090	-0.366811	-0.178762	0.051519	-0.196455	1.000000	-0.738493	-0.398577	-0.273697	-0.394789
S4	0.203841	0.332115	0.413807	0.257650	0.542207	0.659817	-0.738493	1.000000	0.617859	0.417212	0.430453
S5	0.270774	0.149916	0.446157	0.393480	0.515503	0.318357	-0.398577	0.617859	1.000000	0.464669	0.565883
S6	0.301731	0.208133	0.388680	0.390430	0.325717	0.290600	-0.273697	0.417212	0.464669	1.000000	0.382483
Y	0.187889	0.043062	0.586450	0.441482	0.212022	0.174054	-0.394789	0.430453	0.565883	0.382483	1.000000

10 Features in order:

- AGE age in years
- SEX sex
- BMI body mass index
- BP average blood pressure
- S1 tc, total serum cholesterol
- S2 ldl, low-density lipoproteins
- S3 hdl, high-density lipoproteins
- S4 tch, total cholesterol / HDL
- S5 ltg, possibly log of serum triglycerides level
- S6 glu, blood sugar level

Target variable:

- Y response



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Correlation (continued...)

```
mask = np.zeros_like(data.corr())
triangle_indices = np.triu_indices_from(mask)
mask[triangle_indices] = True
mask

array([[1., 1., 1., 1., 1., 1., 1., 1., 1., 1.],
       [0., 1., 1., 1., 1., 1., 1., 1., 1., 1.],
       [0., 0., 1., 1., 1., 1., 1., 1., 1., 1.],
       [0., 0., 0., 1., 1., 1., 1., 1., 1., 1.],
       [0., 0., 0., 0., 1., 1., 1., 1., 1., 1.],
       [0., 0., 0., 0., 0., 1., 1., 1., 1., 1.],
       [0., 0., 0., 0., 0., 0., 1., 1., 1., 1.],
       [0., 0., 0., 0., 0., 0., 0., 1., 1., 1.],
       [0., 0., 0., 0., 0., 0., 0., 0., 1., 1.],
       [0., 0., 0., 0., 0., 0., 0., 0., 0., 1.]])
```



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Correlation (continued...)

```
plt.figure(figsize=(16, 10))
sns.heatmap(data.corr(), mask=mask, annot=True, annot_kws={'size': 14})
plt.xticks(fontsize=10)
plt.yticks(fontsize=10)
plt.show()
```

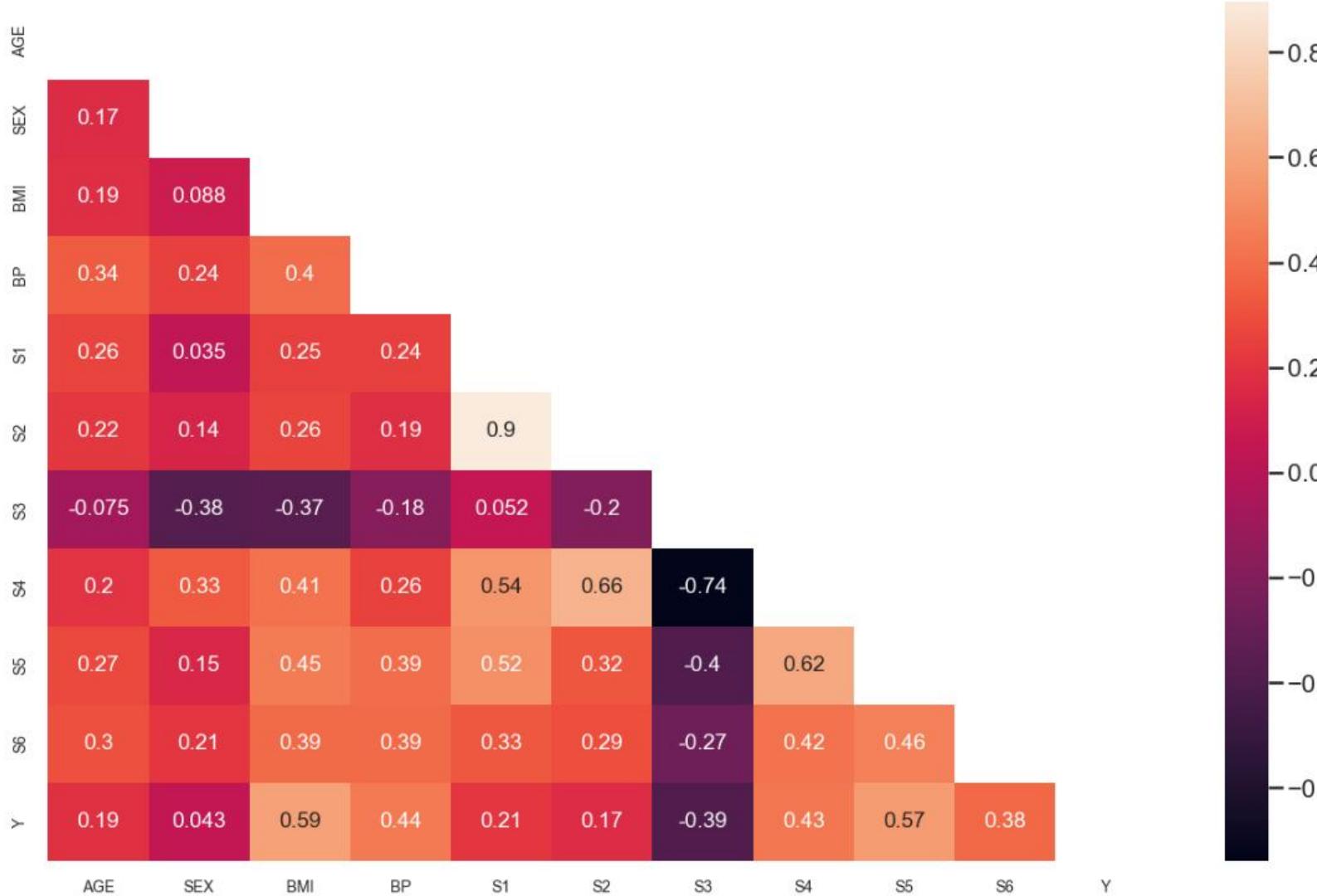


UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Correlation (continued...)



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

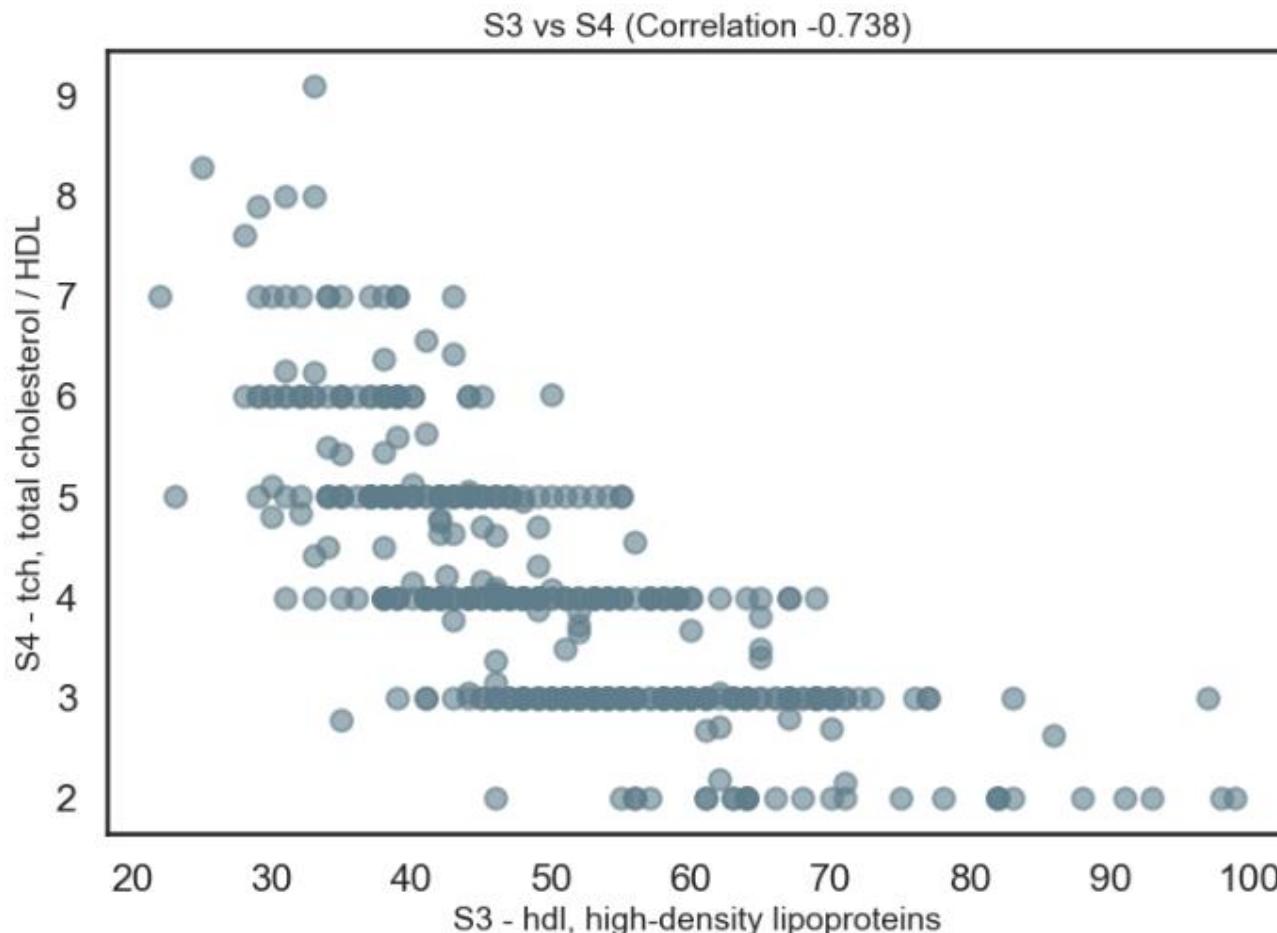
Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za

Correlation (continued...)

```
# Scatter plot between S4 and S3 Correlation
s3_s4_corr = round(data['S3'].corr(data['S4']), 3)
plt.figure(figsize=(9, 6))
plt.scatter(x=data['S3'], y=data['S4'], alpha=0.6, s=80, color="#607D8B")
plt.title(f'S3 vs S4 (Correlation {s3_s4_corr})', fontsize=14)
plt.xlabel('S3 - hdl, high-density lipoproteins', fontsize=14)
plt.ylabel('S4 - tch, total cholesterol / HDL', fontsize=14)
plt.show()
```



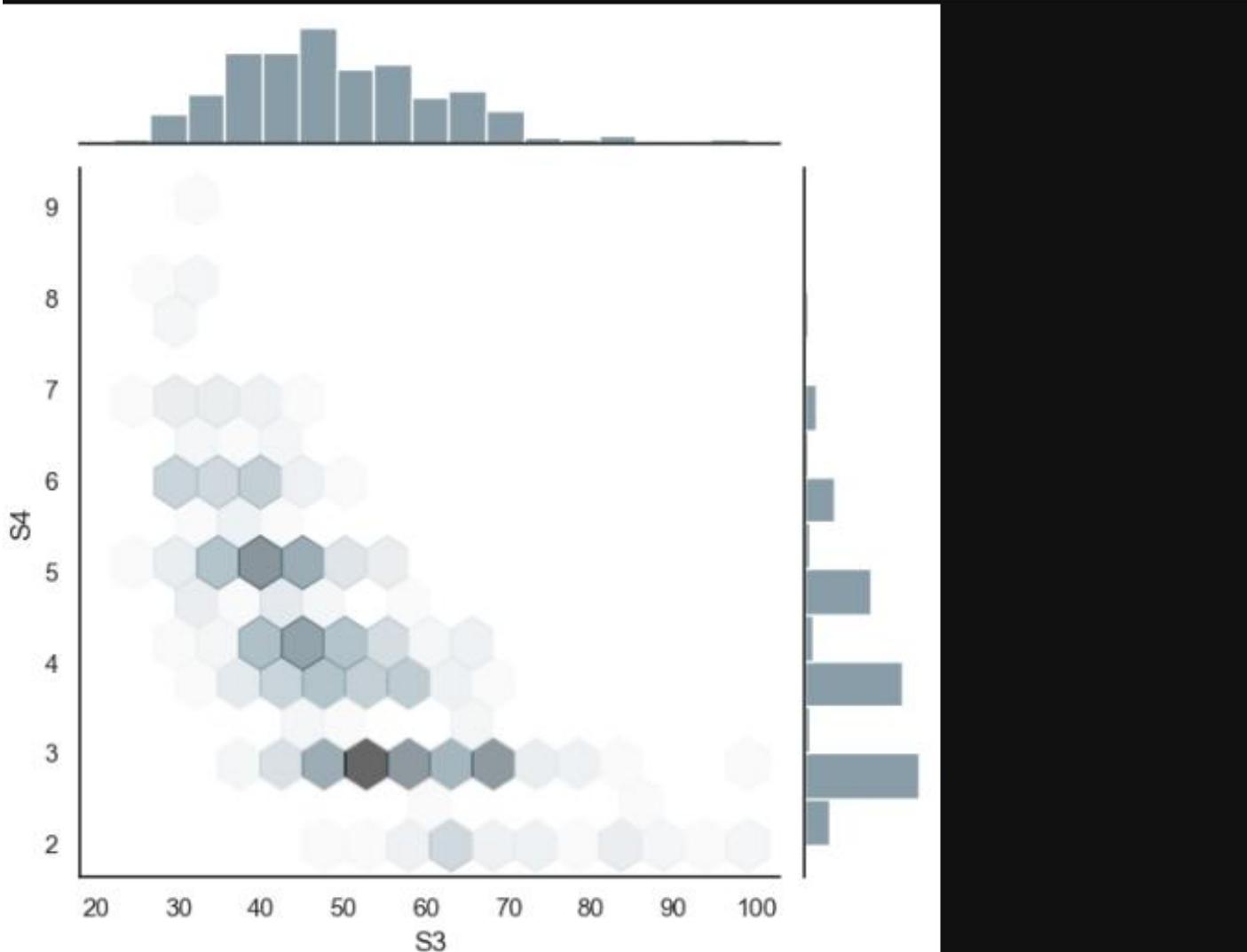
UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Correlation (continued...)

```
sns.set()  
sns.set_context('notebook')  
sns.set_style('white')  
sns.jointplot(x=data['S3'], y=data['S4'], height=6, color='#607D8B', kind='hex', joint_kws={'alpha': 0.6})  
plt.show()
```



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

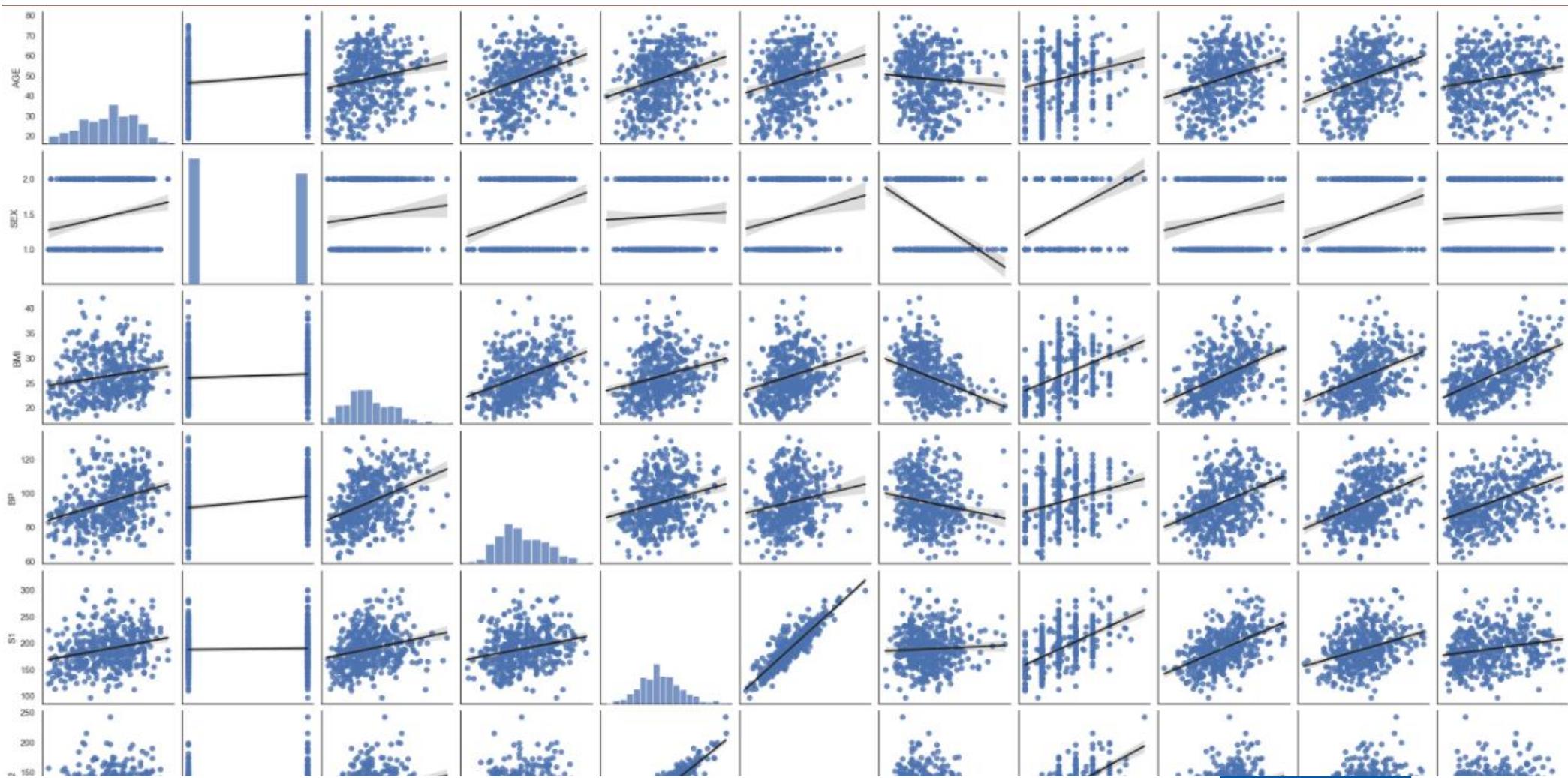
Make today matter
www.up.ac.za

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Correlation (continued...)

```
%%time  
  
sns.pairplot(data, kind='reg', plot_kws={'line_kws':{'color':'#212121'}})  
plt.show()
```



Training & Test Dataset Split



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

**Faculty of Engineering,
Built Environment and
Information Technology**

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter
www.up.ac.za

Training & Test Dataset Split

```
responses = data['Y']
features = data.drop('Y', axis=1)

X_train, X_test, y_train, y_test = train_test_split(features, responses,
                                                    test_size=0.2, random_state=30)

len(X_train)/len(features)
```

```
len(X_test)/len(features)
0.20135746606334842
```



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter
www.up.ac.za

Multivariable Linear Regression



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

**Faculty of Engineering,
Built Environment and
Information Technology**

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter
www.up.ac.za

Multivariable Linear Regression

```
regr = LinearRegression()
regr.fit(X_train, y_train)

print('Training data r-squared:', regr.score(X_train, y_train))
print('Test data r-squared:', regr.score(X_test, y_test))

print('Intercept', regr.intercept_)
pd.DataFrame(data=regr.coef_, index=X_train.columns, columns=['coef'])
```

- Linear regression helps us understand how changes in one or more variables are associated with changes in another variable.



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Multivariable Linear Regression (continued...)

```
Training data r-squared: 0.5264783626678893
Test data r-squared: 0.4653044632644133
Intercept -368.2185304349134

      coef
AGE    0.029063
SEX   -23.059093
BMI    5.602677
BP     1.254404
S1    -1.334738
S2     0.955752
S3     0.588029
S4     3.158733
S5    78.918445
S6     0.226053
```



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Data Transformations



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

**Faculty of Engineering,
Built Environment and
Information Technology**

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za

Data Transformation

```
data['Y'].skew()
```

```
0.44056293407014124
```

```
y_log = np.log(data['Y'])  
y_log.tail()
```

```
437    5.181784  
438    4.644391  
439    4.882802  
440    5.393628  
441    4.043051  
Name: Y, dtype: float64
```

```
y_log.skew()
```

```
-0.3325670604728491
```

- Logarithms (log) helps transform the data before you fit the linear regression line - useful if the data is skew



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

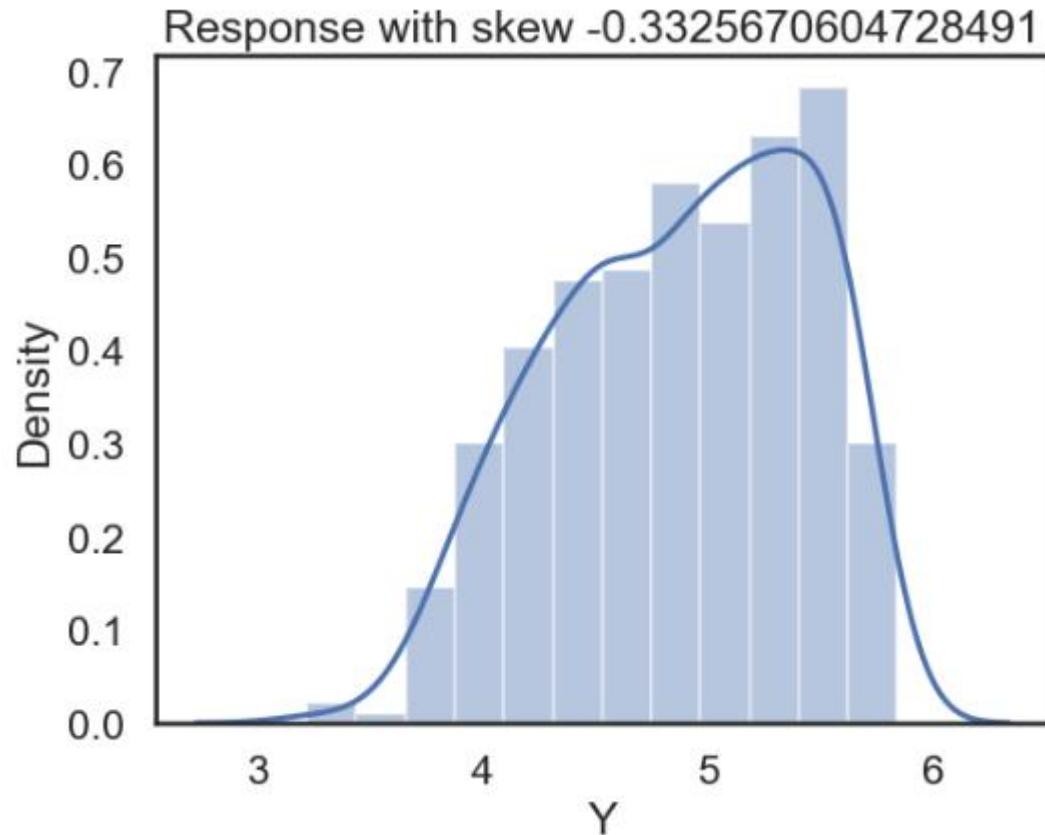
Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

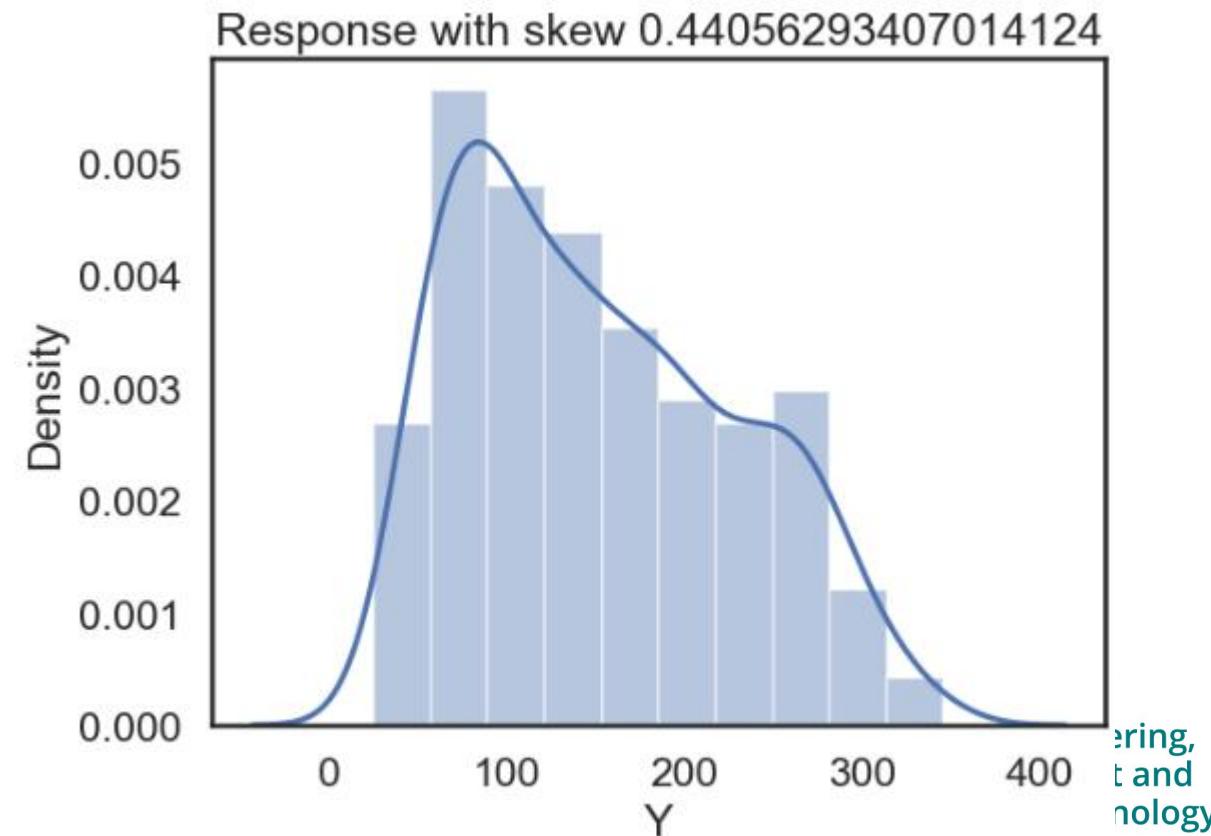
www.up.ac.za

Data Transformation (continued...)

```
sns.distplot(y_log)
plt.title(f'Response with skew {y_log.skew():.15f}')
plt.show()
```



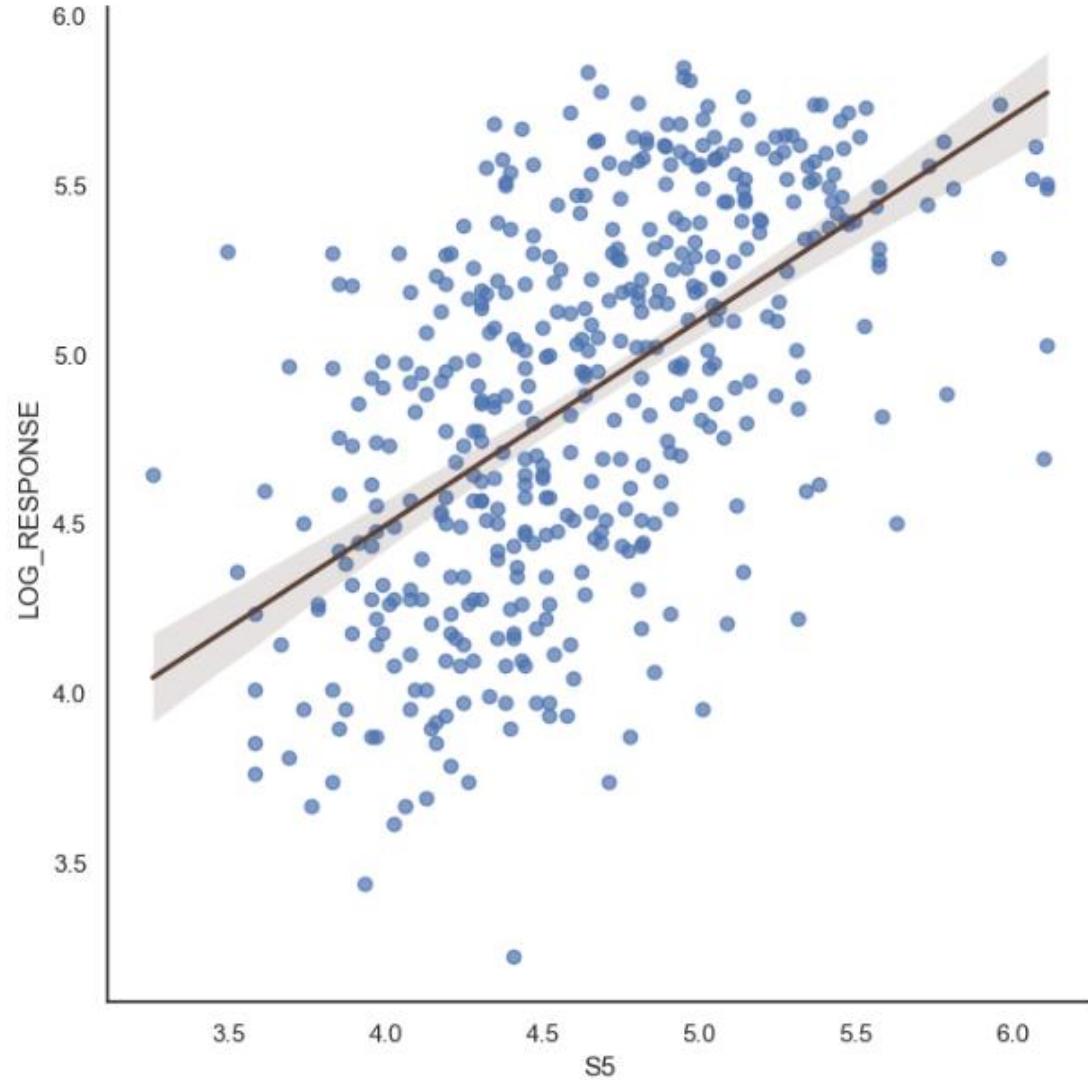
```
tmpdata = data['Y']
sns.distplot(tmpdata)
plt.title(f'Response with skew {tmpdata.skew():.15f}')
plt.show()
```



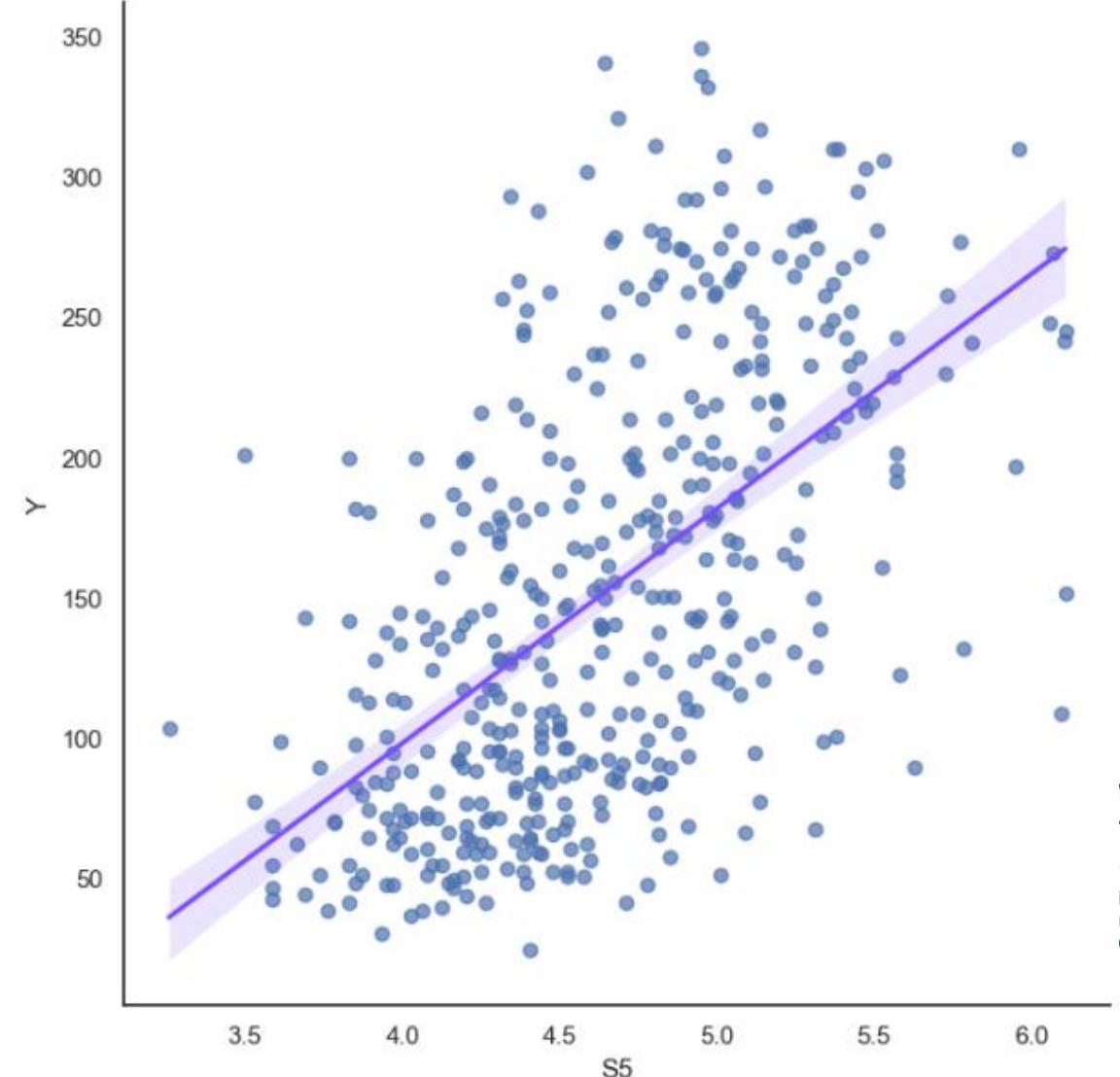
Data Transformation (continued...)

```
# after transformed Log Data
transformed_data = features
transformed_data['LOG_RESPONSE'] = y_log

sns.lmplot(x='S5', y='LOG_RESPONSE', data=transformed_data, height=7, scatter_kws={'alpha': 0.7}, line_kws={'color': '#5D4037'})
plt.show()
```



```
# before transformed Log Data
sns.lmplot(x='S5', y='Y', data=data, height=7, scatter_kws={'alpha': 0.7}, line_kws={'color': '#7C4dff'})
plt.show()
```



P values & Evaluating Coefficients



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za

P values & Evaluating Coefficients

```
X_incl_const = sm.add_constant(X_train)

model = sm.OLS(y_train, X_incl_const)
results = model.fit()

pd.DataFrame({'coef': results.params, 'p-value': round(results.pvalues, 3)})
```

	coef	p-value
const	-368.218530	0.000
AGE	0.029063	0.905
SEX	-23.059093	0.000
BMI	5.602677	0.000
BP	1.254404	0.000
S1	-1.334738	0.031
S2	0.955752	0.096
S3	0.588029	0.487
S4	3.158733	0.624
S5	78.918445	0.000
S6	0.226053	0.457

- $P \leq 0,05$ is considered statistically significant.



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Assignments



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

**Faculty of Engineering,
Built Environment and
Information Technology**

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter
www.up.ac.za

Admin

- Assignment 1: **Due – Monday 11 September @ 5 PM.**
- Assignment 1: marks to be available by the start of Q4.
- Assignment 2: to be available by the start of Q4.
- Semester test: Scheduled for **Friday 06 October 3 PM to 6 PM** (**see Study Guide on ClickUP**)
- Semester test: Written in Labs (Brown, Blue Lab 1, Blue Lab 2)
- Semester test: Open-book, internet availability
- Semester test: test method (prelim)
 - using python and Jupyter Lab



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Enjoy recess!



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

**Faculty of Engineering,
Built Environment and
Information Technology**

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za

Using the Iris Dataset with K-means

The Iris dataset is a classic dataset in machine learning, ideal for clustering since it contains numeric features and well-defined clusters (species of iris flowers).

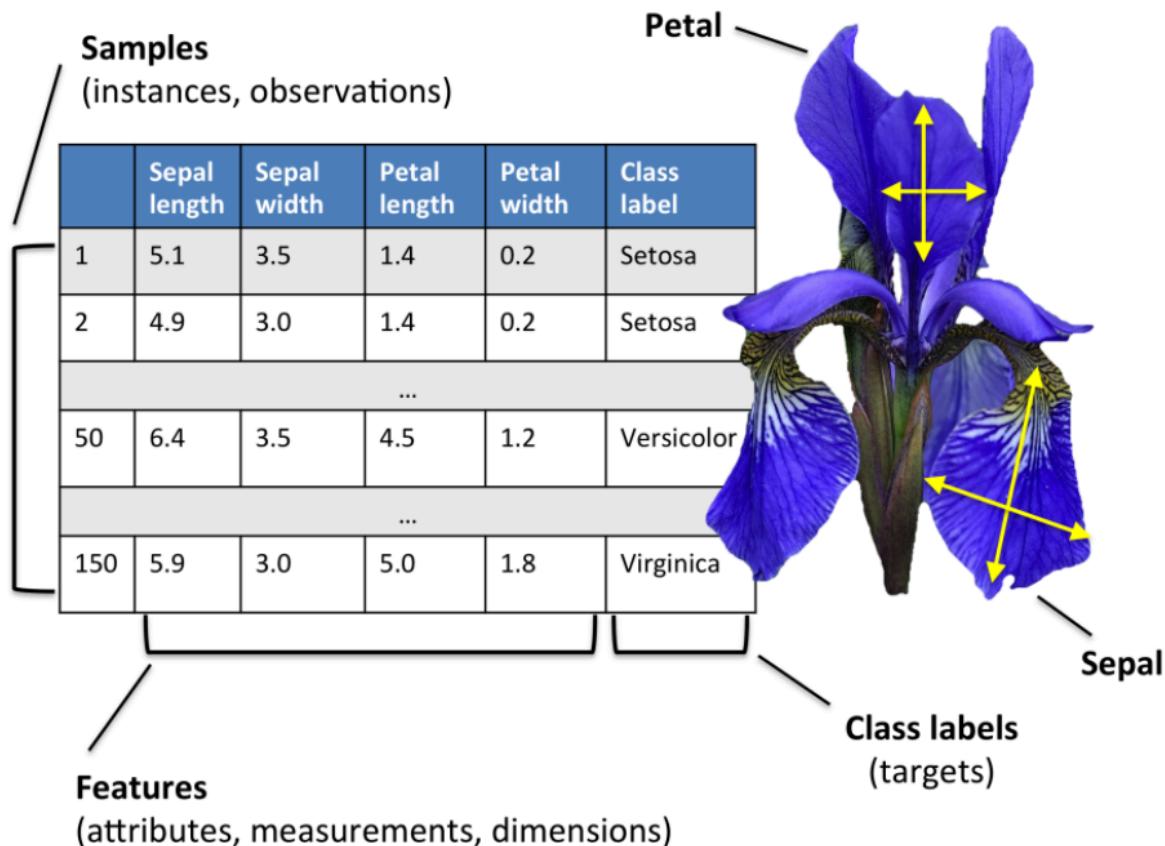


Figure 1. Iris dataset.

```
import pandas as pd

from sklearn.datasets import load_iris

from sklearn.preprocessing import StandardScaler

from sklearn.cluster import KMeans

import seaborn as sns

import matplotlib.pyplot as plt

# Load the Iris dataset
```

```

iris = load_iris()

iris_df = pd.DataFrame(iris.data, columns=iris.feature_names)

# Standardize the data
scaler = StandardScaler()

iris_scaled = scaler.fit_transform(iris_df)

# Run K-means clustering
kmeans = KMeans(n_clusters=3, random_state=42)

iris_df['Cluster'] = kmeans.fit_predict(iris_scaled)

# Visualize the clusters
sns.pairplot(iris_df, hue='Cluster', palette='viridis')

plt.show()

# Inspect cluster centers
print("Cluster Centers:")
print(kmeans.cluster_centers_)

```

Explanation:

- **Iris Dataset:** The Iris dataset is loaded using `sklearn.datasets.load_iris()`. It contains four features (sepal length, sepal width, petal length, and petal width) and is often used to demonstrate clustering.
- **Scaling:** The features are standardised to ensure each feature contributes equally to the clustering.
- **K-means Clustering:** K-means is applied with 3 clusters (since there are three species of iris in the dataset).
- **Visualisation:** A pairplot is used to visualise the clusters across different feature combinations.
- **Cluster Centers:** The centres of the clusters are printed for analysis.

Running K-means on the Iris Dataset

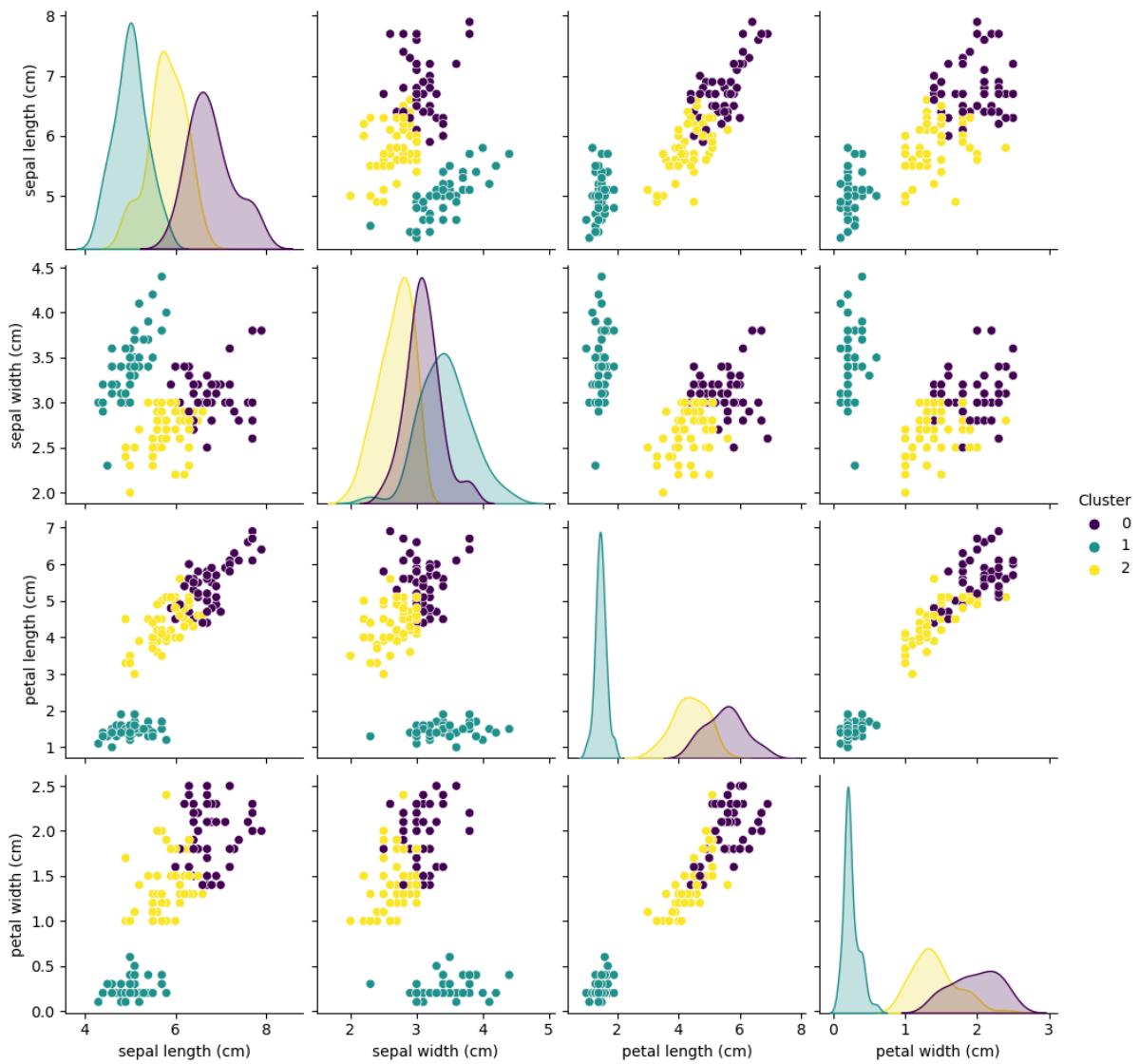


Figure 2. K-Means Clustering Outputs.



ENGINEERING 4.0
UNIVERSITY OF PRETORIA



INF 491/791: Applied Data Science



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter
www.up.ac.za

L 05: Machine Learning



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za

Outline

- ❖ What is Machine Learning?
- ❖ What are key aspects of Machine Learning?
- ❖ Correlation (Recap)
- ❖ Training & Test Dataset Split (Recap)
- ❖ Multivariable Linear Regression (Recap)
- ❖ Data Transformations (Recap)
- ❖ P values & Evaluating Coefficients
- ❖ Bayesian Information Criteria (BIC)
- ❖ Variance Inflation Factor (VIF)
- ❖ Mean Square Error (MSE)
- ❖ Other Mentions
- ❖ Assignment 2 Questions
- ❖ Semester Test Questions



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology
Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

What is Machine Learning?



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

**Faculty of Engineering,
Built Environment and
Information Technology**

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za

What is Machine Learning?

- "Machine learning is a **subset of artificial intelligence** (AI) that focuses on developing algorithms and statistical **models** that enable computer systems to improve their performance on a specific task through learning from **data**, without being explicitly programmed. In essence, it is a way for computers to automatically learn and make predictions or decisions based on patterns and information within the data they are provided." - ChatGPT



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

What is Machine Learning? (continued...)

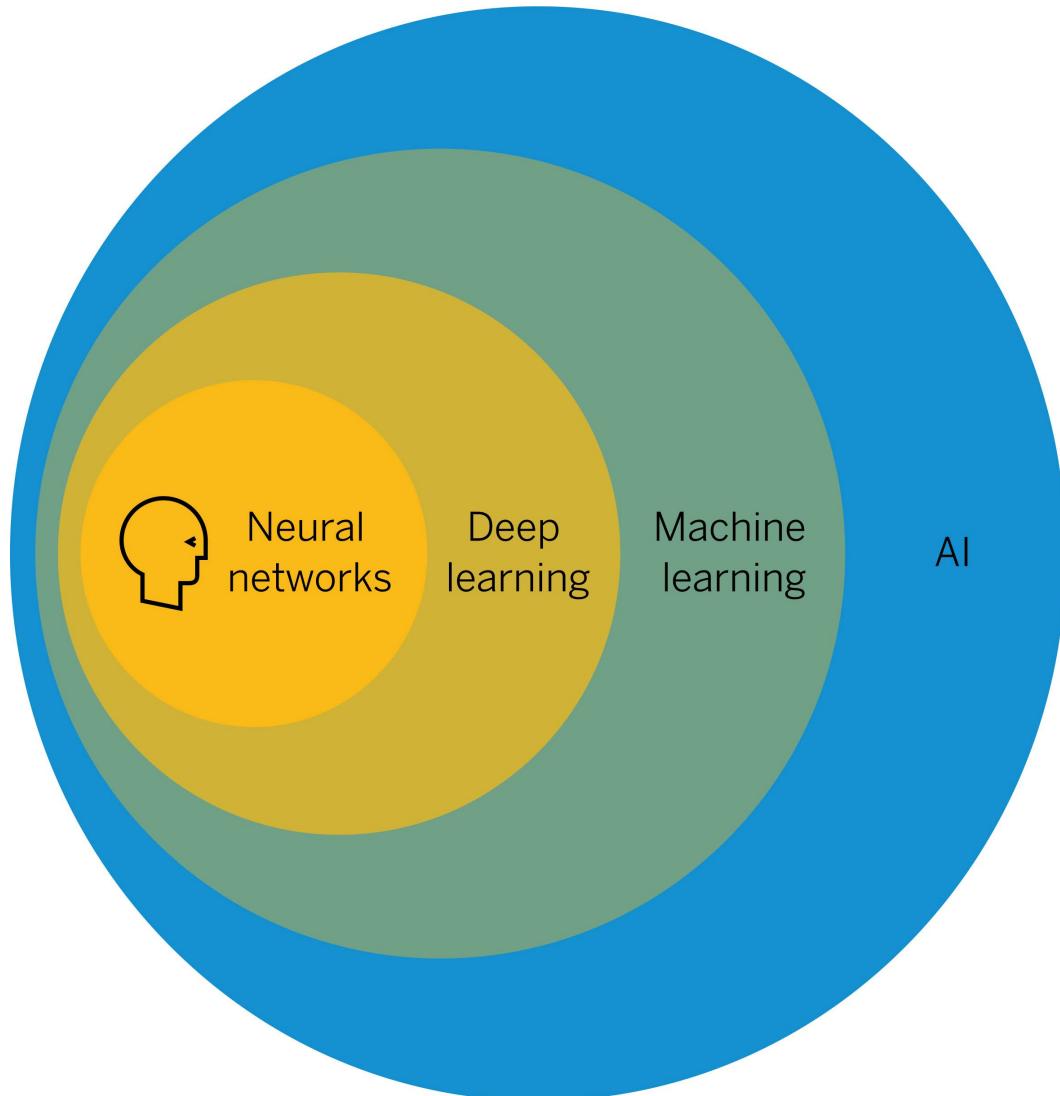
- "*Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of **data** and algorithms to imitate the way that humans learn, gradually improving its accuracy.*" - IBM



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology
Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

What is Machine Learning? (continued...)



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

**Faculty of Engineering,
Built Environment and
Information Technology**

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za

What are key aspects of Machine Learning?



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za

What are key aspects of Machine Learning?

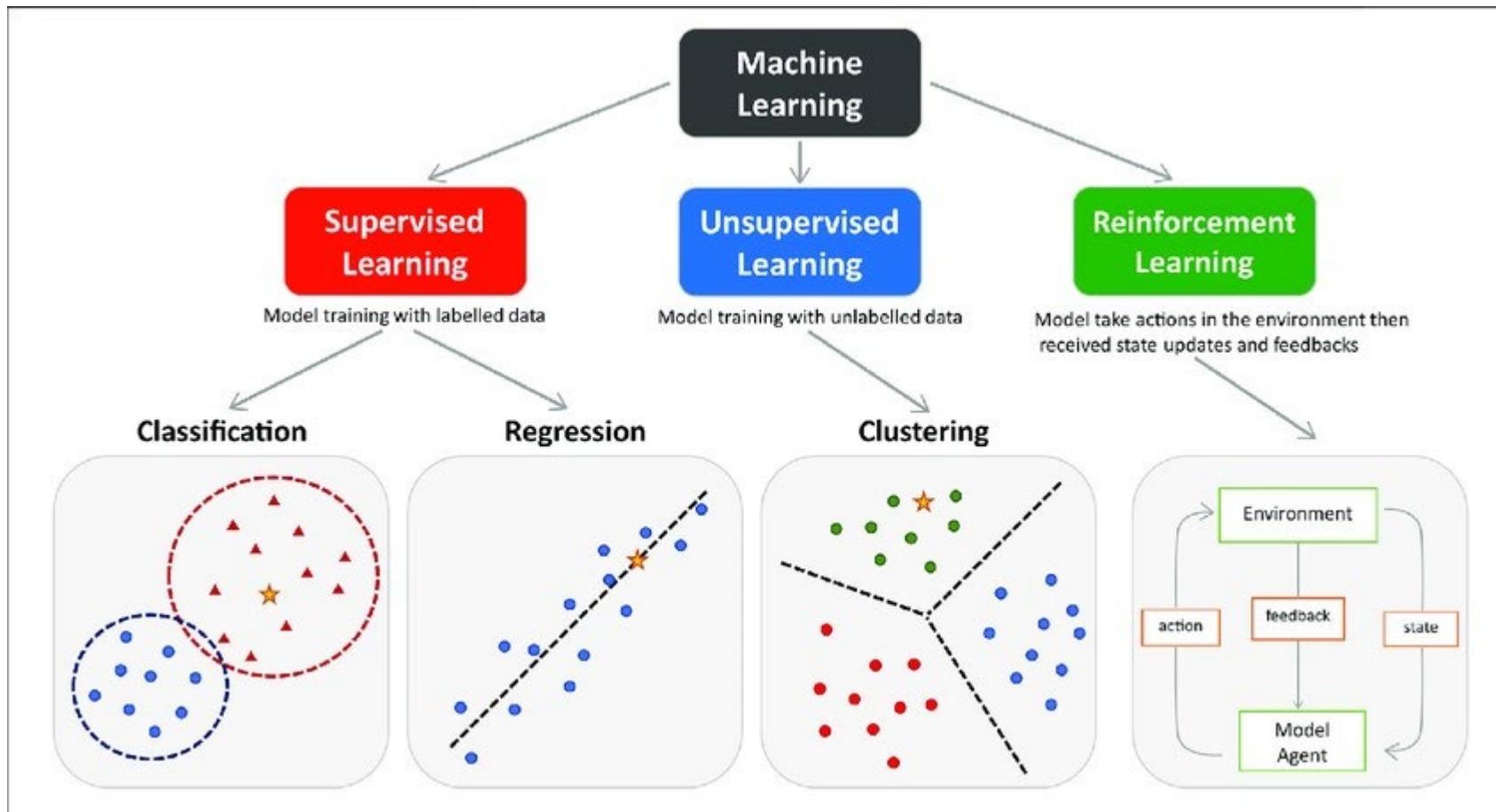
- Data
- Features
- Algorithms
- Learning types (Supervised, Unsupervised, Reinforcement, etc...)
- Model Training
- Evaluation Metrics
- ...
- Overfitting and Underfitting
- ...
- Model Selection
- ...
- Deployment
- ...



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology
Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

What are key aspects of Machine Learning?(continued...)



[1]



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology
Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Correlation (Recap)



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

**Faculty of Engineering,
Built Environment and
Information Technology**

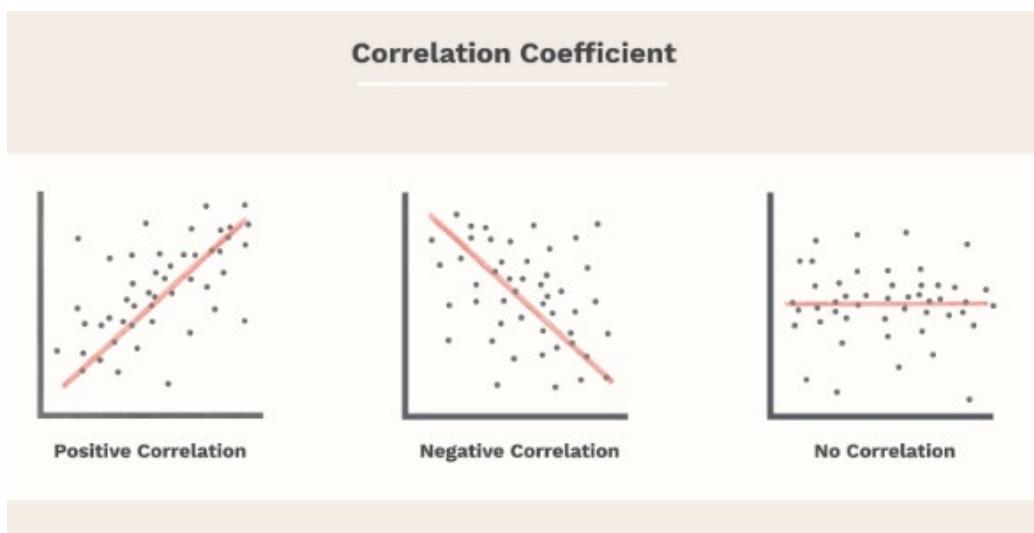
Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za

Correlation

- Correlation is a statistical measure that describes the extent to which two variables change together (i.e., their relationship).
 - 1 is a perfect positive correlation
 - 0 means there is no correlation
 - -1 is a perfect negative correlation
- Correlation (a linear relationship): is important because we want to include features that have the right **strength** and **direction** (i.e., features that are correlated with the target)



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za

Correlation

```
data['Y'].corr(data['BMI'])
```

```
0.5864501344746885
```

```
data['BMI'].corr(data['AGE'])
```

```
0.18508466614655544
```

data.corr()												
	AGE	SEX	BMI	BP	S1	S2	S3	S4	S5	S6	Y	
AGE	1.000000	0.173737	0.185085	0.335428	0.260061	0.219243	-0.075181	0.203841	0.270774	0.301731	0.187889	
SEX	0.173737	1.000000	0.088161	0.241010	0.035277	0.142637	-0.379090	0.332115	0.149916	0.208133	0.043062	
BMI	0.185085	0.088161	1.000000	0.395411	0.249777	0.261170	-0.366811	0.413807	0.446157	0.388680	0.586450	
BP	0.335428	0.241010	0.395411	1.000000	0.242464	0.185548	-0.178762	0.257650	0.393480	0.390430	0.441482	
S1	0.260061	0.035277	0.249777	0.242464	1.000000	0.896663	0.051519	0.542207	0.515503	0.325717	0.212022	
S2	0.219243	0.142637	0.261170	0.185548	0.896663	1.000000	-0.196455	0.659817	0.318357	0.290600	0.174054	
S3	-0.075181	-0.379090	-0.366811	-0.178762	0.051519	-0.196455	1.000000	-0.738493	-0.398577	-0.273697	-0.394789	
S4	0.203841	0.332115	0.413807	0.257650	0.542207	0.659817	-0.738493	1.000000	0.617859	0.417212	0.430453	
S5	0.270774	0.149916	0.446157	0.393480	0.515503	0.318357	-0.398577	0.617859	1.000000	0.464669	0.565883	
S6	0.301731	0.208133	0.388680	0.390430	0.325717	0.290600	-0.273697	0.417212	0.464669	1.000000	0.382483	
Y	0.187889	0.043062	0.586450	0.441482	0.212022	0.174054	-0.394789	0.430453	0.565883	0.382483	1.000000	

10 Features in order:

- AGE age in years
- SEX sex
- BMI body mass index
- BP average blood pressure
- S1 tc, total serum cholesterol
- S2 ldl, low-density lipoproteins
- S3 hdl, high-density lipoproteins
- S4 tch, total cholesterol / HDL
- S5 ltg, possibly log of serum triglycerides level
- S6 glu, blood sugar level

Target variable:

- Y response



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Correlation (continued...)

```
mask = np.zeros_like(data.corr())
triangle_indices = np.triu_indices_from(mask)
mask[triangle_indices] = True
mask

array([[1., 1., 1., 1., 1., 1., 1., 1., 1., 1.],
       [0., 1., 1., 1., 1., 1., 1., 1., 1., 1.],
       [0., 0., 1., 1., 1., 1., 1., 1., 1., 1.],
       [0., 0., 0., 1., 1., 1., 1., 1., 1., 1.],
       [0., 0., 0., 0., 1., 1., 1., 1., 1., 1.],
       [0., 0., 0., 0., 0., 1., 1., 1., 1., 1.],
       [0., 0., 0., 0., 0., 0., 1., 1., 1., 1.],
       [0., 0., 0., 0., 0., 0., 0., 1., 1., 1.],
       [0., 0., 0., 0., 0., 0., 0., 0., 1., 1.],
       [0., 0., 0., 0., 0., 0., 0., 0., 0., 1.],
       [0., 0., 0., 0., 0., 0., 0., 0., 0., 0.]])
```



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Correlation (continued...)

```
plt.figure(figsize=(16, 10))
sns.heatmap(data.corr(), mask=mask, annot=True, annot_kws={'size': 14})
plt.xticks(fontsize=10)
plt.yticks(fontsize=10)
plt.show()
```

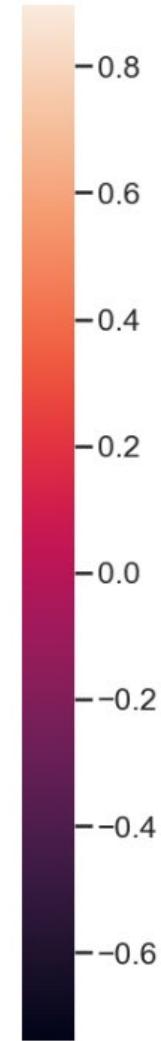
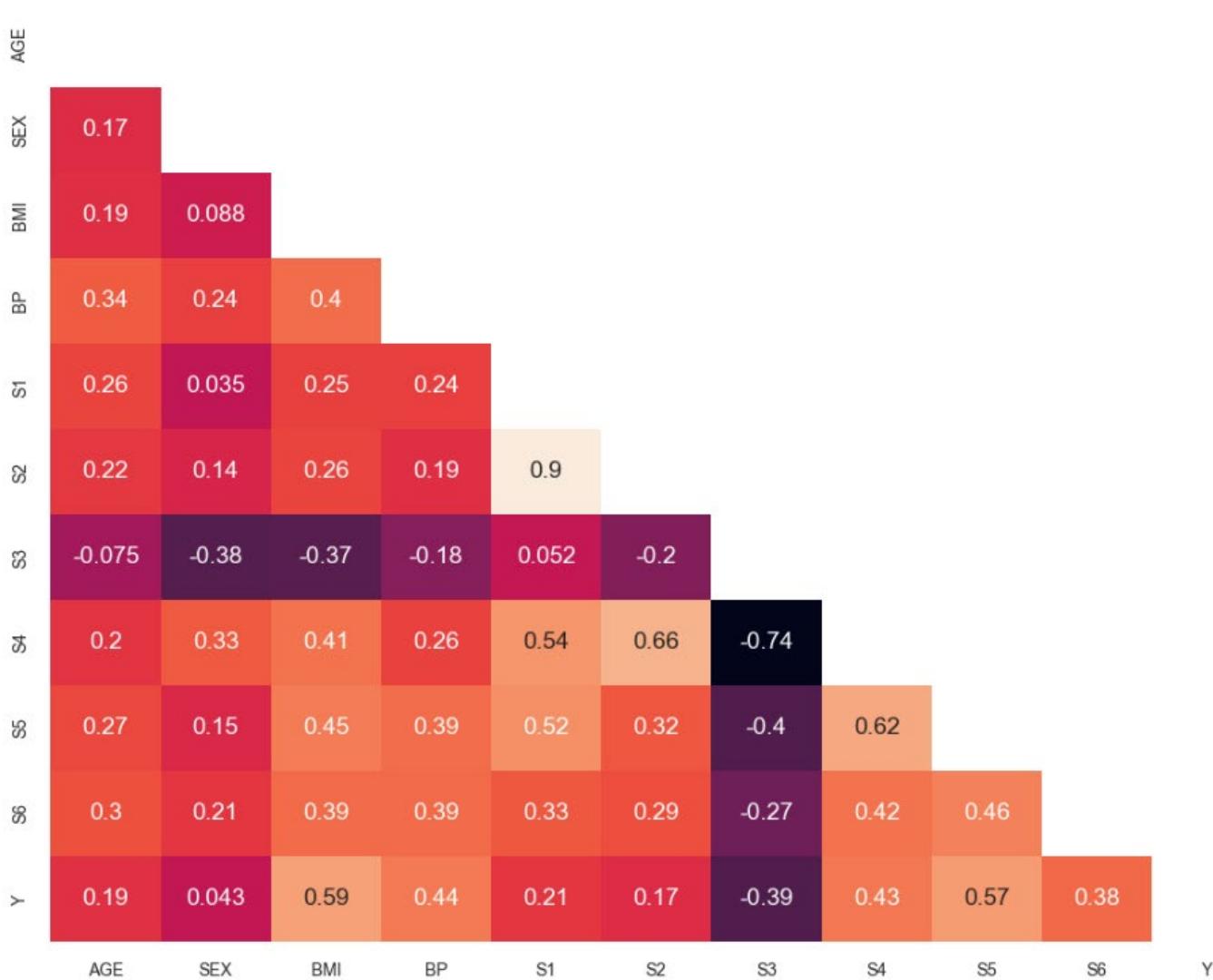


UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Correlation (continued...)



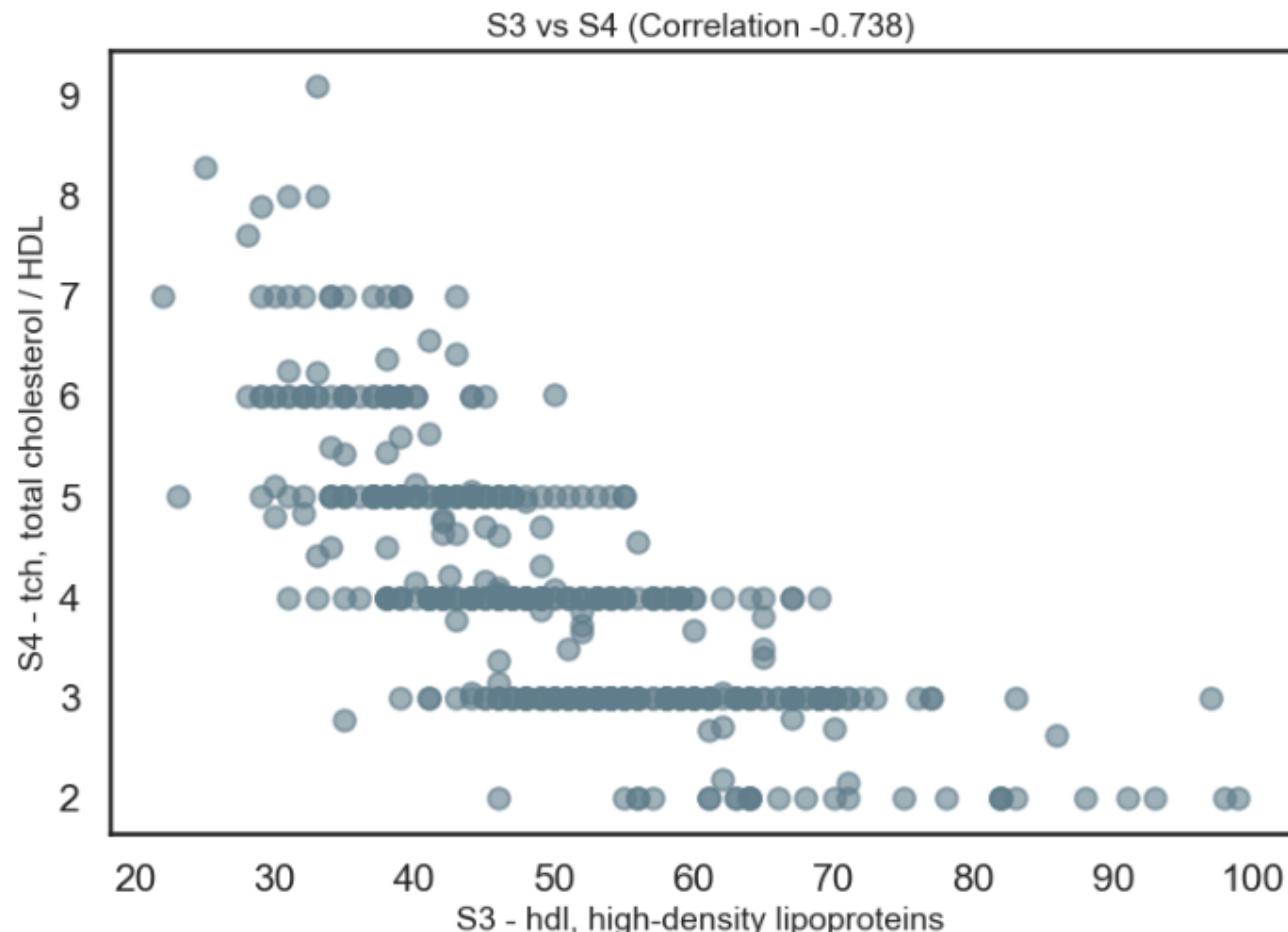
UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Correlation (continued...)

```
# Scatter plot between S4 and S3 Correlation
s3_s4_corr = round(data['S3'].corr(data['S4']), 3)
plt.figure(figsize=(9, 6))
plt.scatter(x=data['S3'], y=data['S4'], alpha=0.6, s=80, color='#607D8B')
plt.title(f'S3 vs S4 (Correlation {s3_s4_corr})', fontsize=14)
plt.xlabel('S3 - hdl, high-density lipoproteins', fontsize=14)
plt.ylabel('S4 - tch, total cholesterol / HDL', fontsize=14)
plt.show()
```



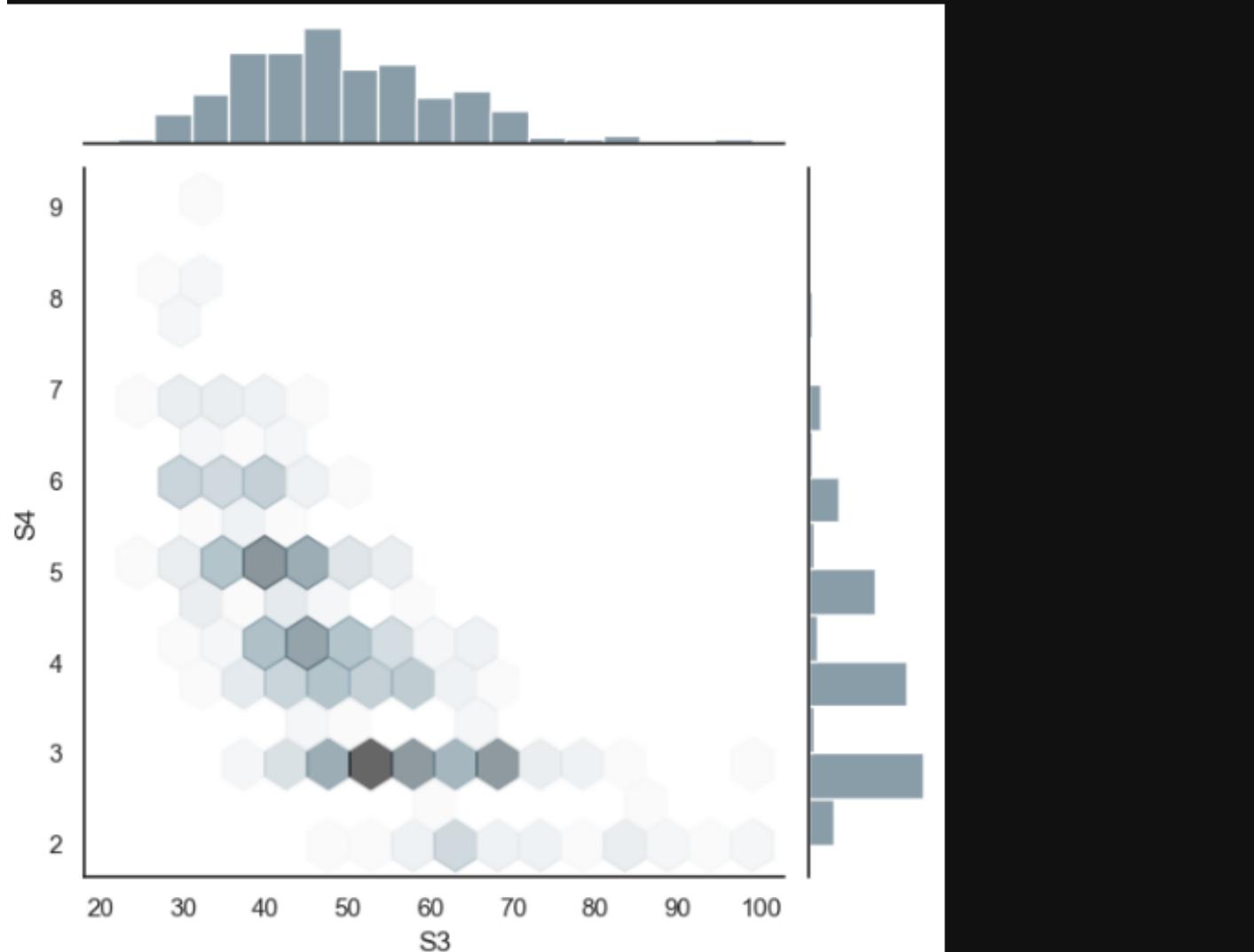
UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Correlation (continued...)

```
sns.set()  
sns.set_context('notebook')  
sns.set_style('white')  
sns.jointplot(x=data['S3'], y=data['S4'], height=6, color="#607D8B", kind='hex', joint_kws={'alpha': 0.6})  
plt.show()
```



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

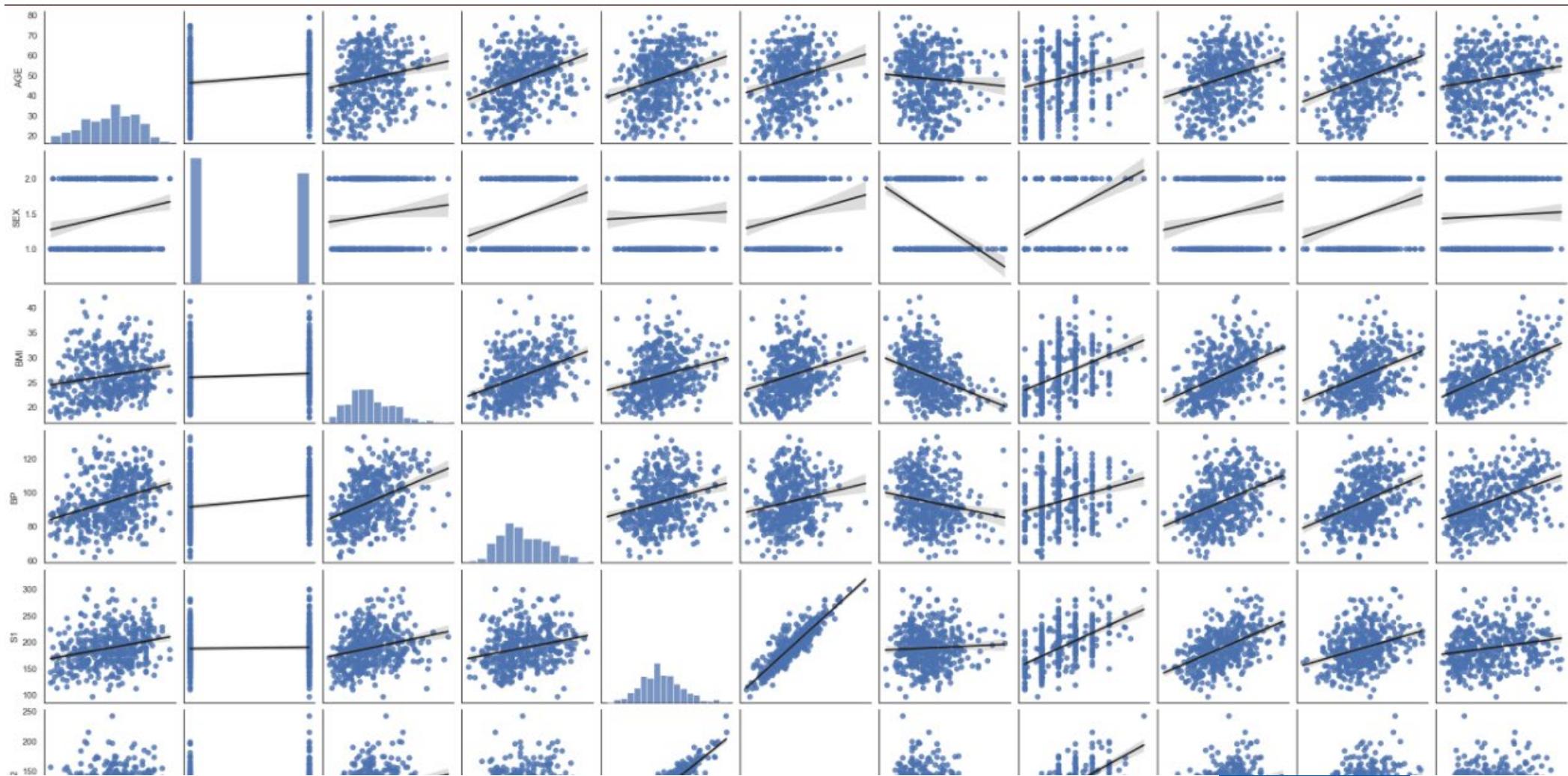
Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter
www.up.ac.za

Correlation (continued...)

```
%time  
  
sns.pairplot(data, kind='reg', plot_kws={'line_kws':{'color':'#212121'}})  
plt.show()
```



Training & Test Dataset Split (Recap)



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

**Faculty of Engineering,
Built Environment and
Information Technology**

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za

Training & Test Dataset Split

```
responses = data['Y']
features = data.drop('Y', axis=1)

X_train, X_test, y_train, y_test = train_test_split(features, responses,
                                                    test_size=0.2, random_state=30)

len(X_train)/len(features)
```

```
len(X_test)/len(features)
0.20135746606334842
```

- Shuffling randomizes the order of your data.
- For data split there is no one size fits all approach.



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology
Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Multivariable Linear Regression (Recap)



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

**Faculty of Engineering,
Built Environment and
Information Technology**

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za

Multivariable Linear Regression

```
regr = LinearRegression()
regr.fit(X_train, y_train)

print('Training data r-squared:', regr.score(X_train, y_train))
print('Test data r-squared:', regr.score(X_test, y_test))

print('Intercept', regr.intercept_)
pd.DataFrame(data=regr.coef_, index=X_train.columns, columns=['coef'])
```

- Linear regression helps us understand how changes in one or more variables are associated with changes in another variable.
- Simple linear regression: between 2 variables (e.g., target (y) and feature (x)).
- Multivariate linear regression: between multiple variables (e.g., target (y) and features (x₁, x₂, x₃ ...x_n))



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology
Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Multivariable Linear Regression (continued...)

Training data r-squared:	0.548373016526865
Test data r-squared:	0.41370714229360894
Intercept	-10.534842392647715
<hr/>	
coef	
AGE	0.002635
SEX	-0.788625
BMI	0.227139
BP	0.056038
S1	-0.087029
S2	0.071969
S3	0.048395
S4	0.268017
S5	3.932265
S6	-0.014507

- To evaluate your model you use statistics such as:
 - r-squared
 - p-values
 - MSE
 - VIF
 - BIC
 - ...

- Note: we were able to explain approximately 53% (r-squared) of the variance of the response with just 10 features.
- The performance of the model on the test data is going to be worse than that of the training data.
- Negative coefficients means that there is an inverse relationship between that feature and the dependent variable. You interpret coefficients in the context of your specific dataset and problem domain.



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Data Transformations *(Recap)*



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

**Faculty of Engineering,
Built Environment and
Information Technology**

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za

Data Transformation

```
data['Y'].skew()
```

```
0.44056293437814124
```

```
y_log = np.log(data['Y'])  
y_log.tail()
```

```
437    5.181784  
438    4.644391  
439    4.882802  
440    5.393628  
441    4.043051  
Name: Y, dtype: float64
```

```
y_log.skew()
```

```
-0.3325670604728491
```

- Logarithms (log) helps transform the data before you fit the linear regression line - useful if the data is skew



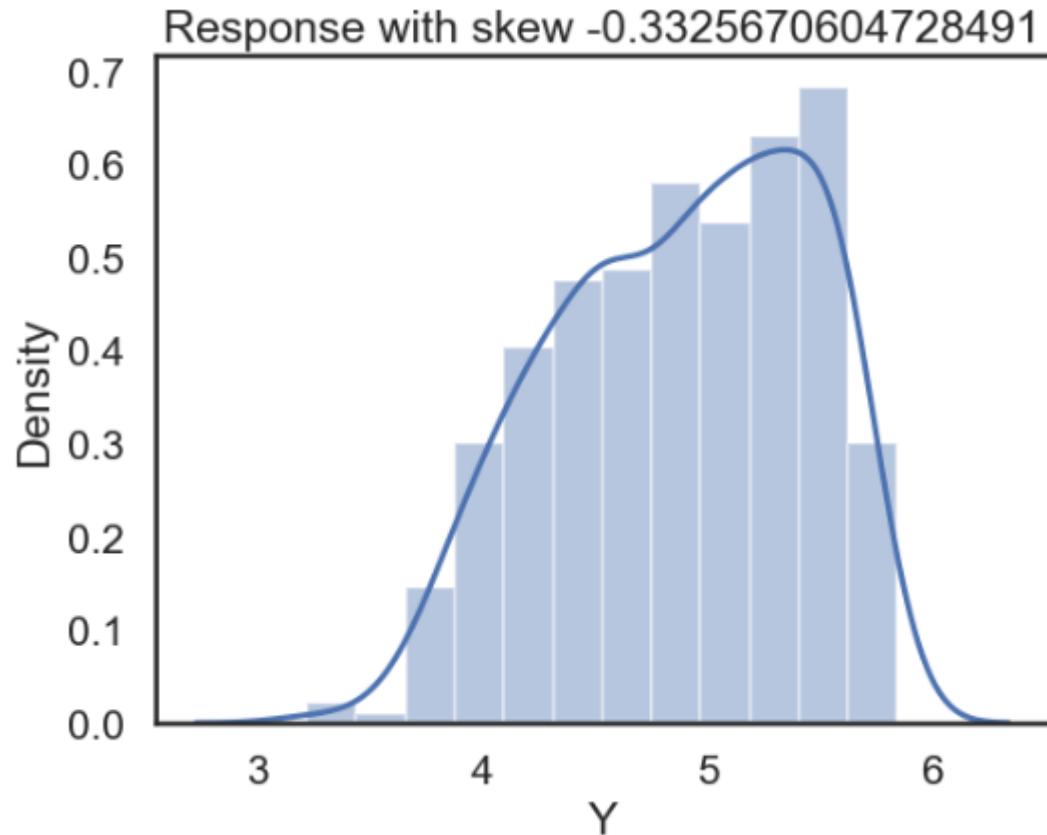
UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

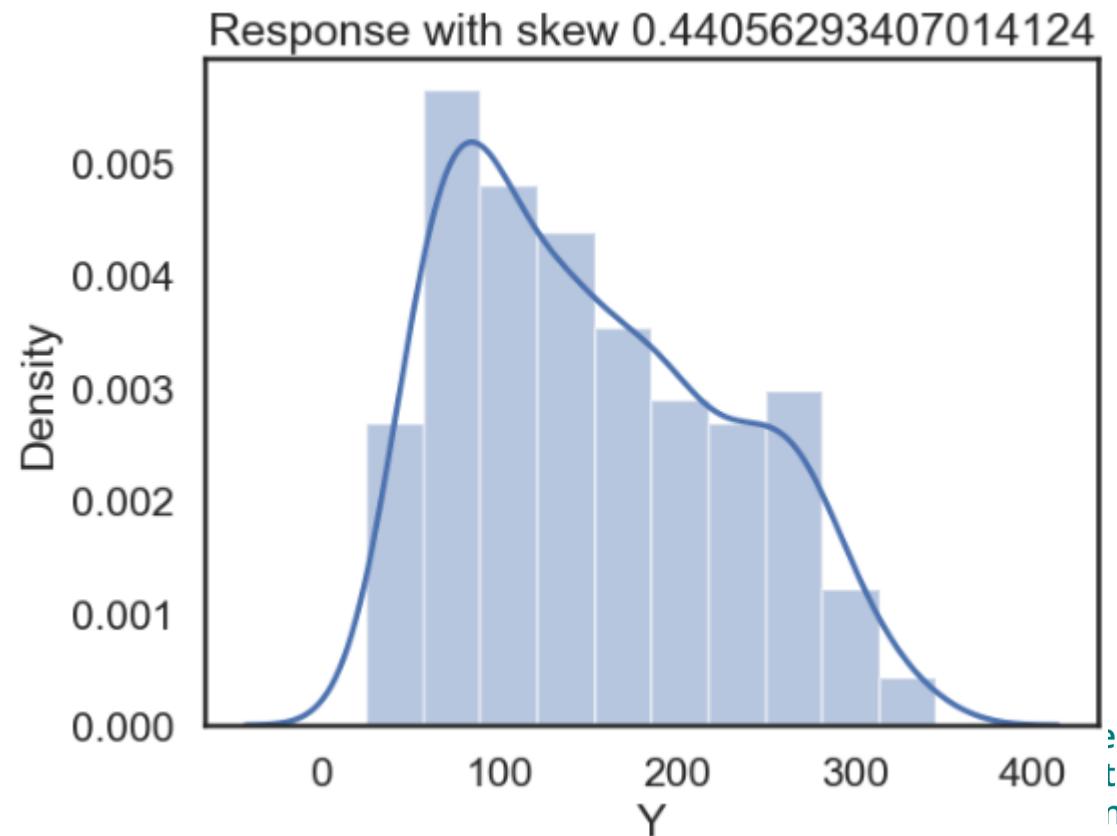
Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Data Transformation (continued...)

```
sns.distplot(y_log)
plt.title(f'Response with skew {y_log.skew():.15f}')
plt.show()
```



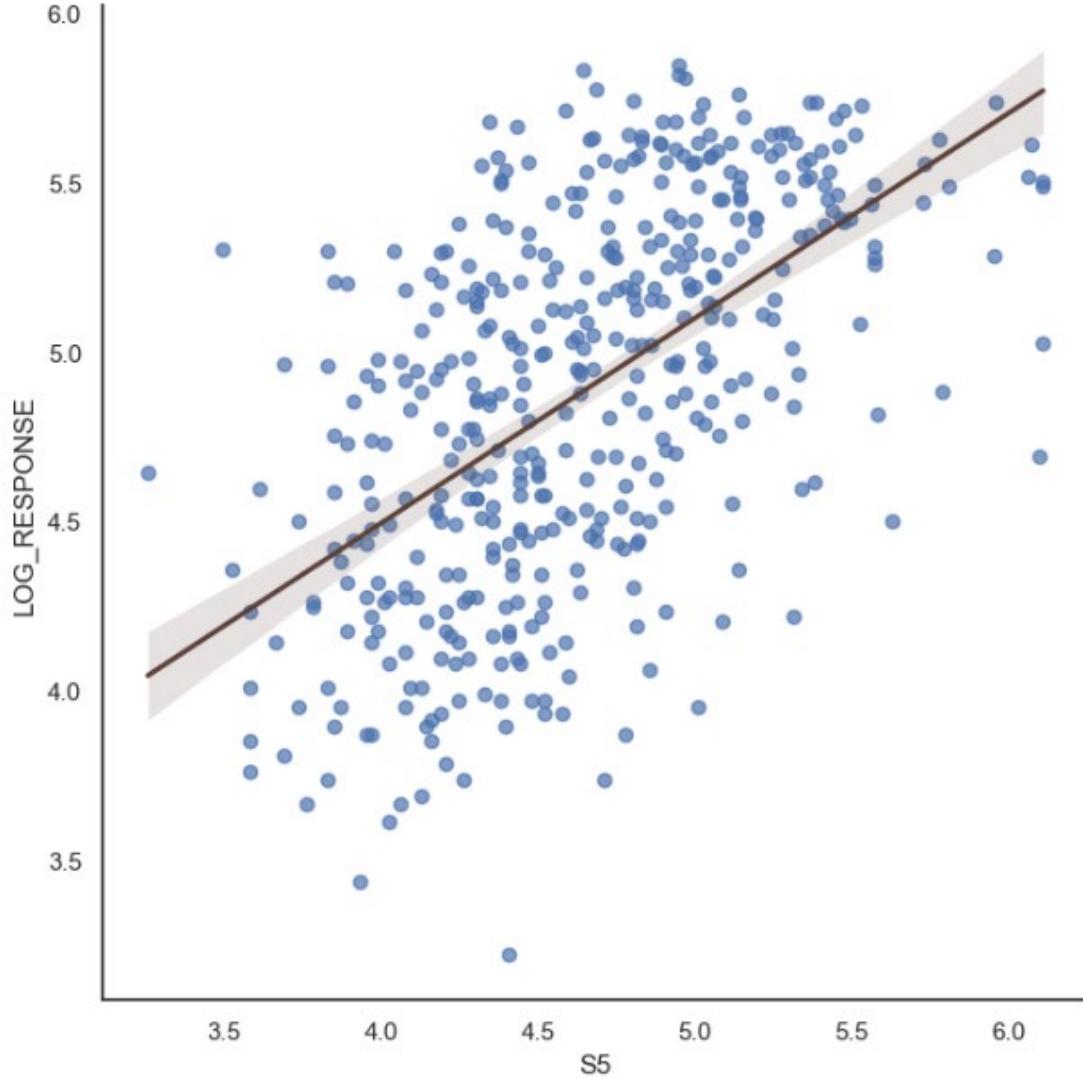
```
tmpdata = data[ "Y" ]
sns.distplot(tmpdata)
plt.title(f'Response with skew {tmpdata.skew():.15f}')
plt.show()
```



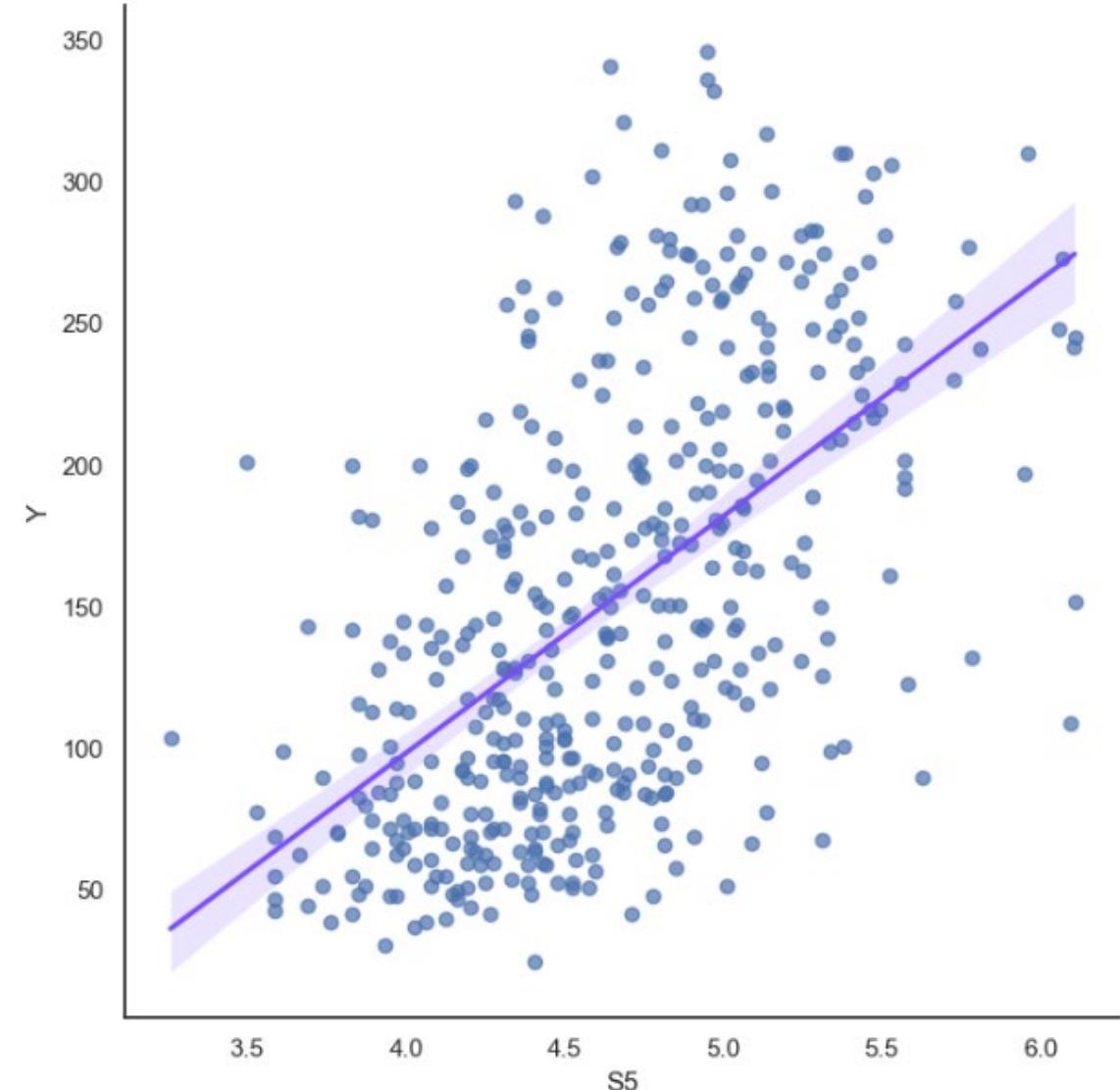
Data Transformation (continued...)

```
# after transformed Log Data
transformed_data = features
transformed_data['LOG_RESPONSE'] = y_log

sns.lmplot(x='S5', y='LOG_RESPONSE', data=transformed_data, height=7, scatter_kws={'alpha': 0.7}, line_kws={'color': '#5D4037'})
plt.show()
```



```
# before transformed Log Data
sns.lmplot(x='S5', y='Y', data=data, height=7, scatter_kws={'alpha': 0.7}, line_kws={'color': '#7C4dff'})
plt.show()
```



Data Transformation (continued...)

- Other data transformations besides log transformation for linear regression include:
 - Square Root Transformation
 - Exponential Transformation
 - Box-Cox Transformation
 - Yeo-Johnson Transformation
 - Etc.

```
response = np.log(data['Y'])
features = data.drop(['Y'], axis=1)

X_train, X_test, y_train, y_test = train_test_split(features, response,
                                                    test_size=0.4, random_state=20)

regr = LinearRegression()
regr.fit(X_train, y_train)

print('Training data r-squared:', regr.score(X_train, y_train))
print('Test data r-squared:', regr.score(X_test, y_test))

print('Intercept', regr.intercept_)
pd.DataFrame(data=regr.coef_, index=X_train.columns, columns=['coef'])

Training data r-squared: 0.5250952185956388
Test data r-squared: 0.38530854757235067
Intercept 1.0806938826594847
```

```
response = np.sqrt(data['Y'])
features = data.drop('Y', axis=1)

X_train, X_test, y_train, y_test = train_test_split(features, response,
                                                    test_size=0.4, random_state=20)

regr = LinearRegression()
regr.fit(X_train, y_train)

print('Training data r-squared:', regr.score(X_train, y_train))
print('Test data r-squared:', regr.score(X_test, y_test))

print('Intercept', regr.intercept_)
pd.DataFrame(data=regr.coef_, index=X_train.columns, columns=['coef'])

Training data r-squared: 0.548373016526865
Test data r-squared: 0.41370714229360894
Intercept -10.534842392647715
```

P values & Evaluating Coefficients (Recap)



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za

P values & Evaluating Coefficients

	coef	p-value
const	-10.534842	0.003
AGE	0.002635	0.817
SEX	-0.788625	0.009
BMI	0.227139	0.000
BP	0.056038	0.000
S1	-0.087029	0.003
S2	0.071969	0.008
S3	0.048395	0.233
S4	0.268017	0.397
S5	3.932265	0.000
S6	-0.014507	0.312

- $P \leq 0,05$ is considered statistically significant.
- Nb: It is preferred to have a more simple model to explain things, therefore, simpler models are considered better than complex models.
- So, to simplify your model you can drop features that are insignificant. But, this should not be done “**willy-nilly**”. Because, even a feature with a high/bad p-value can add value to the model as a whole. **To drop features requires some contemplation.**



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Bayesian Information Criteria (BIC)



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

**Faculty of Engineering,
Built Environment and
Information Technology**

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za

Bayesian Information Criteria (BIC)

```
print('BIC is', results.bic)
print('r-squared is', results.rsquared)
pd.DataFrame({'coef': results.params, 'p-value': round(results.pvalues, 3)})
BIC is 2900.7543929935687
```

```
print('BIC is', results.bic)
pd.DataFrame({'coef': results.params, 'p-value': round(results.pvalues, 3)})
BIC is 305.1614591304148
```

```
print('BIC is', results.bic)
pd.DataFrame({'coef': results.params, 'p-value': round(results.pvalues, 3)})
BIC is 1214.431162240603
```

- BIC is a model selection criterion for a finite list of models.
- A way to measure complexity between models.
- The lower the BIC value the better the model (a lower value is better).



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Variance Inflation Factor (VIF)



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

**Faculty of Engineering,
Built Environment and
Information Technology**

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za

Variance Inflation Factor (VIF)

```
x_incl_const.columns  
Index(['const', 'AGE', 'SEX', 'BMI', 'BP', 'S1', 'S2', 'S3', 'S4', 'S5', 'S6'], dtype='object')  
  
for i in range(X_incl_const.shape[1]):  
    print(variance_inflation_factor(exog=X_incl_const, exog_idx=i))  
  
671.8771190666284  
1.1487198884621623  
1.2553595575721868  
1.4774401394787282  
1.4332140437255596  
57.98699505658825  
39.391753886328544  
14.313441346908945  
9.783714465306861  
10.571305534829273  
1.4323006290342313
```

- We are testing for Multicollinearity.
- VIF is a measure of collinearity amongst the features within a regression model.
- A high VIF value indicates that a feature is highly correlated with other features, potentially leading to unstable coefficient estimates.



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Variance Inflation Factor (VIF) continued...

- **VIF = 1:** A VIF of 1 indicates no multicollinearity. It means that the feature is not correlated with any other feature in the model.
- **VIF < 5:** A VIF below 5 generally suggests low to moderate multicollinearity. In this range, you typically don't need to be overly concerned about multicollinearity.
- **VIF >= 5:** A VIF of 5 or greater suggests high multicollinearity. This indicates that the feature is highly correlated with other features in the model, and it may be problematic.
- **VIF > 10:** A VIF greater than 10 is a strong indication of severe multicollinearity. In such cases, it's essential to address multicollinearity, such as by removing one or more correlated features, through feature selection, or by applying dimensionality reduction techniques.



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology
Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Mean Square Error (MSE)



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za

Mean Square Error (MSE)

```
# Mean Squared Error & R-Squared
full_normal_mse = round(results.mse_resid, 3)
full_normal_rsquared = round(results.rsquared, 3)
```

```
pd.DataFrame({'R-Squared': [full_normal_rsquared],
              'MSE': [full_normal_mse],
              'RMSE': np.sqrt([full_normal_mse])},
              index=['Full Normal Response Model'])
```

	R-Squared	MSE	RMSE
Full Normal Response Model	0.475	3217.196	56.720331

```
pd.DataFrame({'R-Squared': [reduced_log_rsquared, full_normal_rsquared, omitted_var_rsquared],
              'MSE': [reduced_log_mse, full_normal_mse, omitted_var_mse],
              'RMSE': np.sqrt([reduced_log_mse, full_normal_mse, omitted_var_mse])},
              index=['Reduced Log Model', 'Full Feature Set Model', 'Reduced Log with Omitted Features Model'])
```

	R-Squared	MSE	RMSE
Reduced Log Model	0.527	0.039	0.197484
Full Feature Set Model	0.564	0.507	0.712039
Reduced Log with Omitted Features Model	0.527	0.038	0.194936

- MSE calculates how close the regression line is to the data points by taking into account the predicted value of the observation and eliminates the arbitrariness associated with the residual sum of the squares.
- MSE equation measures the average squared error of our predictions between the actual output and the model's prediction.
- The lower the MSE value the lower the variance of error.



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Other Mentions



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

**Faculty of Engineering,
Built Environment and
Information Technology**

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter
www.up.ac.za

Other Mentions

- We need to diagnose regression models on a case by case basis.
- We can check for missing features, consider transforming the data, etc.
- We use Residuals to check if model and our assumptions are valid.
- Residuals should be normally distributed with no identifiable pattern (read up).



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology
Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Assignment 2 Questions?



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

**Faculty of Engineering,
Built Environment and
Information Technology**

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter
www.up.ac.za

Semester Test Questions?



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za

Awesomeness!



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

**Faculty of Engineering,
Built Environment and
Information Technology**

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za



ENGINEERING 4.0
UNIVERSITY OF PRETORIA



INF 491/791: Applied Data Science



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Make today matter

www.up.ac.za

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

L 05: Natural Language Processing and Computational Lexicography

by Dr. Mike Wa Nkongolo



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Make today matter
www.up.ac.za

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Outline

- ❖ Introduction to Natural Language Processing (NLP)
- ❖ Computational Lexicography Overview
- ❖ Key Techniques in NLP
- ❖ Role of Lexicons in NLP
- ❖ Text Preprocessing
- ❖ NLP and Machine Translation
- ❖ Tools and Libraries for NLP and Lexicography
- ❖ Computational Lexicography in Practice
- ❖ Applications of NLP and Computational Lexicography
- ❖ Challenges and Future Directions
- ❖ Conclusion
- ❖ Code Demonstration
- ❖ **Assignment Discussion**



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

What is Natural Language Processing (NLP)?



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za

What is NLP?

- **Natural Language Processing (NLP)** is a branch of artificial intelligence that enables machines to understand, interpret, and respond to human language in a valuable way. It involves the interaction between computers and humans using natural language, making it crucial for applications like voice assistants, chatbots, translation tools, and sentiment analysis. The importance of NLP lies in its ability to bridge the gap between human communication and machine understanding, allowing for more intuitive interactions with AI systems in various industries such as healthcare, customer service, and education.



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology
Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Key NLP Tasks

Tokenization

- **Definition:** Tokenization is the **process of breaking down text into smaller units such as words, phrases, or sentences**. These tokens are the building blocks for further analysis and processing.
- **Importance:** Tokenization is essential because it **helps in structuring unstructured text for various NLP tasks**, such as sentiment analysis or machine translation.

Part-of-Speech Tagging (POS Tagging)

- **Definition:** This process involves **assigning grammatical tags (e.g., nouns, verbs, adjectives) to each word in a sentence**.
- **Importance:** POS tagging helps in understanding the **syntactic structure** of a sentence, which is critical for tasks like **parsing, machine translation, and text generation**.



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology
Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Key NLP Tasks

Named Entity Recognition (NER)

- **Definition:** NER **identifies and classifies entities in text**, such as names of people, organizations, locations, and dates.
- **Importance:** NER is used in tasks like **information extraction, question answering, and knowledge graph construction**, helping machines identify and understand real-world entities in text.

Text Classification

- **Definition:** Text classification involves **categorizing text into predefined labels or categories**, such as spam detection, topic classification, or sentiment categorization.
- **Importance:** This task is vital for **filtering information and automating processes**, such as sorting emails or analyzing customer feedback.

Sentiment Analysis

- **Definition:** Sentiment analysis **detects emotions or opinions within text**, determining whether the sentiment expressed is **positive, negative, or neutral**.
- **Importance:** Sentiment analysis is widely used in areas like **social media monitoring, customer reviews, and market research**, where understanding public opinion is crucial.



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology
Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

These NLP tasks are foundational in modern AI applications, enabling machines to process and respond to human language more naturally and effectively.



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za

Computational Lexicography

- Computational Lexicography is the study and development of **electronic dictionaries and lexicons using computational methods**. It involves leveraging **algorithms and digital tools to create, organize, and manage large-scale dictionaries and lexical resources**. **These lexicons are then used in various natural language processing (NLP) applications** to enhance machine understanding of language.

Importance in NLP

Dictionaries and lexicons are **foundational resources** in NLP as they help machines comprehend word meanings, relationships, synonyms, and grammatical structures.

They are essential for tasks such as **machine translation**, **text mining**, and **semantic analysis**, where accurate word definitions and contexts are critical for language understanding.

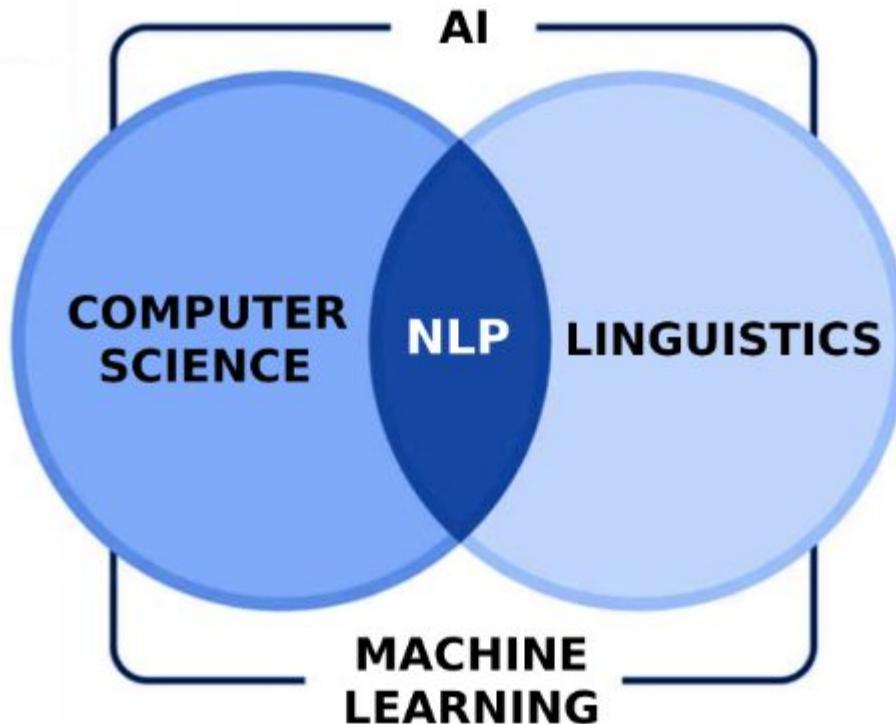
Computational lexicography enables the **creation of rich, structured data that can be processed** by algorithms, making it easier to develop more efficient and accurate NLP systems.



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology
Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

What are key aspects of NLP? (continued...)



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Key Techniques in NLP



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

**Faculty of Engineering,
Built Environment and
Information Technology**

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za

Bag of Words (BoW)

Definition: Bag of Words is a simple and widely used technique in NLP that **represents a text as an unordered collection of words, disregarding grammar and word order but keeping track of word frequencies**. In this model, a **document is represented as a vector, where each word corresponds to a dimension**, and the value represents the frequency of the word in that document.

Use Case: BoW is **commonly used in text classification tasks**, such as spam detection or sentiment analysis, where understanding the presence or absence of certain words is more important than their order.



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology
Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

BoW (continued...)

The Bag of Words Representation

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!



it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

TF-IDF (Term Frequency-Inverse Document Frequency)



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za

TF-IDF

Definition: TF-IDF is **a numerical statistic** that reflects **how important a word is to a document in a collection or corpus**. It combines two factors: **term frequency (TF)**, which measures how often a word appears in a document, and **inverse document frequency (IDF)**, which measures how common or rare the word is across the entire corpus. A **higher TF-IDF score means a word is more significant** in distinguishing the document from others.

Use Case: TF-IDF is useful for **information retrieval, ranking the importance of words in documents, and document similarity tasks**, such as recommending articles based on content.



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology
Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

```
Word indexes:  
{'geeks': 1, 'for': 0, 'r2j': 2}
```

```
tf-idf value:  
(0, 0) 0.5493512310263033  
(0, 1) 0.8355915419449176  
(1, 1) 1.0  
(2, 2) 1.0
```



tf-idf value of word having index 1 i.e. geeks in
document index 0 i.e. d0

From the above image the below table can be generated:

Document	Word	Document Index	Word Index	tf-idf value
d0	for	0	0	0.549
d0	geeks	0	1	0.8355
d1	geeks	1	1	1.000
d2	r2j	2	2	1.000



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

**Faculty of Engineering,
Built Environment and
Information Technology**

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za

BM25

Definition: BM25 is an **advanced ranking** function used in information retrieval systems like **search engines**. It builds on the **TF-IDF model** but **improves the ranking of documents** by considering term saturation (**diminishing returns as terms are repeated**) and **document length normalization**. BM25 ranks documents based on the relevance of their content to a query.

Use Case: BM25 is widely used in search engines to assess the relevance of documents and provide more accurate search results.



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Word Embeddings (e.g., Word2Vec, GloVe)



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

**Faculty of Engineering,
Built Environment and
Information Technology**

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

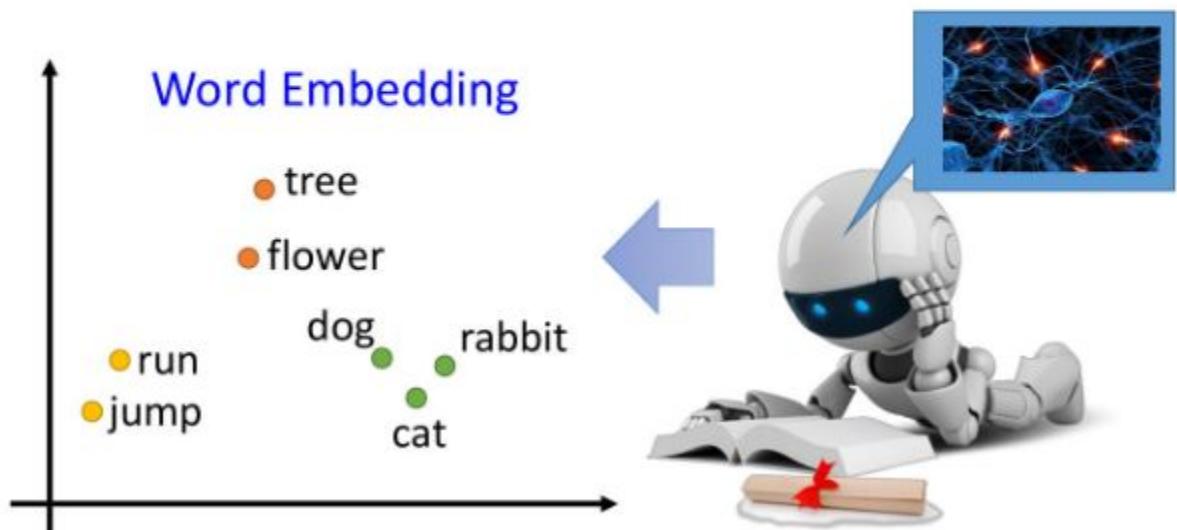
Make today matter

www.up.ac.za

Word Embeddings

Word Embedding

- Machine learns the meaning of words from reading a lot of documents without supervision



- Definition:** Word embeddings are dense vector representations of words in a continuous vector space, capturing semantic relationships between words. Word2Vec and GloVe are popular algorithms for generating word embeddings.
- Unlike Bag of Words or TF-IDF, word embeddings take into account the context in which words appear, allowing them to capture more nuanced relationships, such as "king" being closer to "queen" than to "man."
- Use Case:** Word embeddings are used in tasks like machine translation, question answering, and text similarity, where understanding the semantic meaning of words is essential.



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

These techniques provide foundational methods for processing and analyzing text in NLP, enabling machines to better understand and generate human language.



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter
www.up.ac.za

Role of Lexicons in NLP

Type	Corresponding Synsets	Freq
art	artifact, artefact	8283
act	act, human_action, human_activity	6606
hum	person, individual, someone, mortal, human, soul	6303
grb	biological_group	4933
atr	attribute	4137
psy	psychological_feature	3456
com	communication	3336
anim	animal, animate_being, beast, brute, creature, fauna	2703
plt	plant, flora, plant_life	2311
sta	state	2266
foo	food, nutrient	1541
log	region 1 (geographical location)	1282
nat	natural_object - water, - land,	1277
sub*	substance, matter	1189
evt	event	1082
prt	part, piece	992
gra	social_group - people	940
qud	definite_quantity	777
pro	process	773
chm	compound, chemical_compound - chemical_element, element	699

Type	Corresponding Synsets	Freq
time	time_period, period, period_of_time, amount_of_time - time_unit, unit_of_time - time	628
agt	causal_agent, cause, causal_agency	624
pos	possession	571
loc*	location, (any other location)	567
rel*	relation	506
frm	shape, form	420
grp*	group, grouping (any other group)	345
phm*	phenomenon	342
qui	indefinite_quantity	295
pho*	object, inanimate_object, physical_object	186
mic	microorganism	178
lme	linear_measure, long_measure	100
lfr*	life_form, organism, being, living_thing	61
cel	cell	57
meas*	measure, quantity, amount, quantum	38
ent*	entity	28
con	consequence, effect, outcome, result, upshot	21
spc	space	21
abs*	abstraction	8

- Lexicons are essential resources in **Natural Language Processing (NLP)**, offering predefined lists of words along with relevant linguistic properties, such as their meanings, sentiment scores, and part-of-speech tags. Lexicons serve as foundational tools for analyzing and understanding text in tasks like machine translation, text classification, and sentiment analysis. They provide structured information that helps algorithms interpret the meaning and emotional tone of words, which is crucial for accurately processing language in tasks like text mining and semantic analysis. Lexicon-based approaches are especially valuable for tasks that require interpretability, as the **lexicons provide explicit mappings between words and their meanings**.



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology
Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Sentiment Analysis Lexicons

SentiWordNet

- **Definition:** SentiWordNet is an extension of **WordNet**, assigning sentiment scores (positive, negative, and neutral) to each word. It is widely used in **sentiment analysis** tasks for **extracting the emotional tone of words in various contexts**.
- **Use Case:** This lexicon is commonly used for applications requiring fine-grained sentiment detection, such as product reviews or movie rating analysis.

VADER (Valence Aware Dictionary and sEntiment Reasoner)

- **Definition:** VADER is a lexicon and **rule-based** model designed to handle sentiment analysis, particularly suited for **social media** content due to its ability to handle informal language, **emoticons**, and abbreviations.
- **Use Case:** VADER is often used for **real-time sentiment analysis in platforms like Twitter or Facebook**, where informal language prevails and quick assessment of public sentiment is necessary.

AFINN-111

- **Definition:** AFINN-111 is a lexicon **containing words rated for valence on a scale from negative to positive**. It is primarily used for binary or multi-class sentiment classification tasks.
- **Use Case:** It is commonly applied in sentiment analysis tasks for customer feedback or survey responses where simplicity and quick evaluation of sentiment are essential.



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Lexicon-Based vs. Machine Learning Approaches

Lexicon-Based Approaches

- **Advantages:** These methods rely on **predefined word lists**, making them easy to implement and interpret. They are particularly **effective in smaller datasets or applications** requiring transparency and traceability, as they provide **clear mappings between words and sentiments**.
- **Limitations:** Lexicon-based methods may **lack flexibility when encountering out-of-vocabulary words** or complex contexts. **They often struggle with detecting nuanced sentiments or slang not covered by the lexicon.**

Machine Learning Approaches

- **Advantages:** ML models are **more flexible and can learn patterns from data**, making them more adaptable to complex contexts, **idiomatic expressions**, and evolving language use. They generally **perform better in large-scale datasets** and can automatically capture more nuanced sentiments.
- **Limitations:** Machine learning models can be **less interpretable**, **requiring more resources for training** and validation.

Additionally, they often **require extensive labeled data** and may suffer from **overfitting or bias if not properly managed**.



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology
Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter
www.up.ac.za

In summary, while lexicon-based methods offer simplicity and interpretability, machine learning models provide greater flexibility and accuracy, making the choice between the two dependent on the specific needs and constraints of the NLP task.



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter
www.up.ac.za

Text Preprocessing

Text preprocessing is a critical step in **Natural Language Processing (NLP)**, **transforming raw text into a format that can be easily analyzed** by algorithms. It involves various techniques to clean, structure, and prepare text data before it can be used for tasks such as text classification, sentiment analysis, or machine translation.

Tokenization:

- **Definition:** Tokenization refers to the **process of splitting text into smaller units**, known as tokens. These tokens can be **words, subwords, or even characters**, depending on the application. For example, a sentence like "Natural Language Processing is fascinating!" could be tokenized into individual words: ["Natural", "Language", "Processing", "is", "fascinating", "!"].
- **Importance:** Tokenization helps **break down large chunks of text** into manageable pieces, making it easier for NLP models to process and analyze language at a finer level, such as word-level or sentence-level analysis.



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Text Preprocessing

Stop Words Removal

- **Definition:** Stop words are common words that usually carry little meaning in text analysis (e.g., "and," "the," "is," "in"). Removing these **words reduces the noise in the data and helps focus the analysis on the more meaningful words that convey key information.**
- **Importance:** By removing stop words, algorithms can improve performance in tasks such as **text classification or sentiment analysis** because they no longer waste resources on processing common but uninformative words. However, stop word removal should be done carefully, as some contexts may require these words.



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Text Preprocessing

Handling Polysemy (Polysemy Dilemma)

- **Definition:** Polysemy refers to the phenomenon where a single word has multiple meanings. For example, the word "bank" can mean a financial institution or the side of a river. Handling polysemy is essential for accurately interpreting text because the correct meaning depends on the context in which the word appears.
- **Solutions:** Techniques like **Word Sense Disambiguation (WSD)** are employed to resolve polysemy by using the surrounding words (context) to infer the correct meaning. Machine learning models like **Word2Vec** can also help handle polysemy by learning different word embeddings for the various meanings of a word based on context.



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology
Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

These preprocessing techniques ensure that raw text is transformed into a structured format, allowing machine learning algorithms to work more efficiently and produce more accurate results.



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter
www.up.ac.za

NLP and Machine Translation

Machine translation (MT) is a key application of **Natural Language Processing (NLP)**, aiming to automatically translate text or speech from one language to another. Over the years, several methodologies have been developed for text translation, each with its own strengths and limitations.

1. Text Translation Methods:

- **Rule-based Systems:** Early machine translation systems relied on handcrafted linguistic rules and **bilingual dictionaries**. These systems translated text by applying syntactic, semantic, and grammatical rules of the source and target languages. While they provided high-quality translations in certain cases, they were limited by their dependency on predefined rules, which made them difficult to scale across diverse languages and contexts.

Statistical Machine Translation (SMT): SMT emerged as a more data-driven approach, where translation models were built using **large bilingual corpora**. Instead of relying on rules, SMT used probabilities to predict the most likely translation based on patterns observed in parallel texts. Phrases or words in one language were matched to their translations in another language, and the best translations were selected based on statistical models. However, SMT struggled with complex sentence structures and context beyond short phrases.



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

NLP and Machine Translation

Neural Machine Translation (NMT): NMT represents the state of the art in machine translation. It uses deep learning models, particularly **Recurrent Neural Networks (RNNs)** or **Transformer models**, to encode entire sentences or paragraphs into high-dimensional vector representations. These vectors capture the meaning of the input text, allowing the model to generate more accurate and contextually appropriate translations. NMT has shown remarkable improvements in fluency and coherence compared to SMT, particularly with the rise of the Transformer-based architecture like **Google's BERT** and **OpenAI's GPT** models.



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

NLP and Machine Translation

Challenges in Machine Translation:

- **Translating Polysemic Words:** One of the major challenges in machine translation is handling polysemic words—words with multiple meanings depending on the context. For instance, the English word "bark" can mean the sound a dog makes or the outer layer of a tree. Choosing the correct meaning in translation is essential to preserving the meaning of the sentence. NMT models mitigate this issue by learning contextual representations of words, but even advanced systems sometimes struggle with ambiguous terms.
- **Maintaining Meaning Across Languages:** Maintaining meaning during translation can be complex, especially for languages with different grammatical structures, idiomatic expressions, or cultural references. Some languages, for example, use more context-specific information than others. For example, translating gender-neutral sentences from English into gendered languages like French or Spanish requires additional contextual understanding to maintain meaning and avoid errors.



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

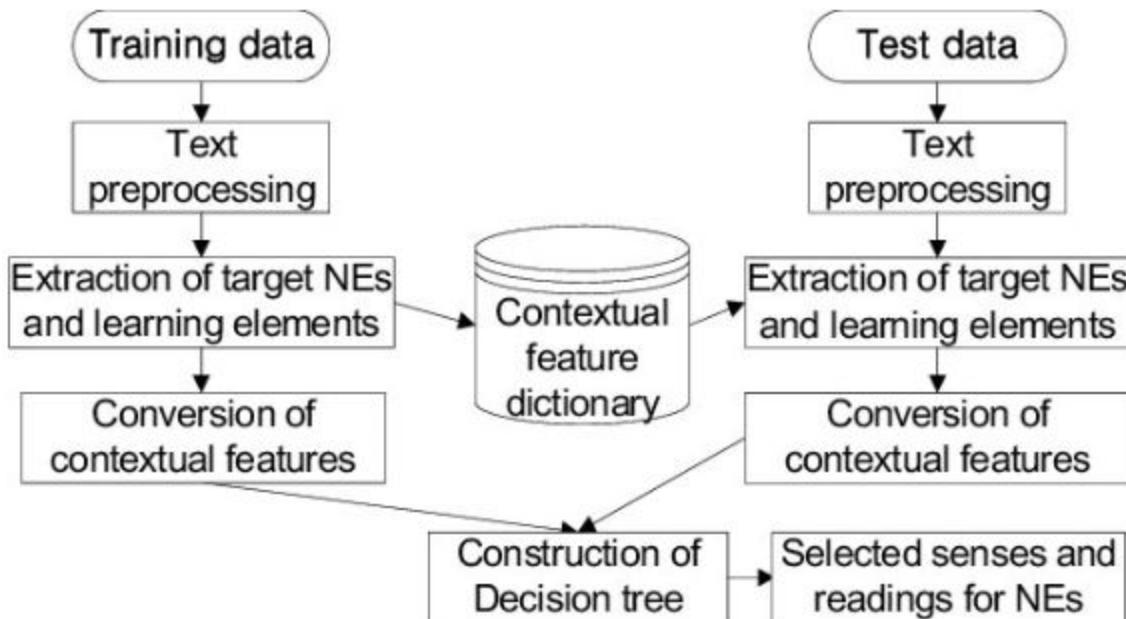
Modern machine translation has evolved from rule-based systems to powerful NMT techniques. Despite the significant advancements, challenges such as polysemy and preserving cultural meaning across languages remain ongoing areas of research in NLP



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

**Faculty of Engineering,
Built Environment and
Information Technology**
Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

In linguistics, a *corpus* is a collection of linguistic data (usually contained in a computer database) used for research, scholarship, and teaching. Also called a *text corpus*. Plural: *corpora*.



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Tools and Libraries for NLP and Lexicography

- **Natural Language Toolkit (NLTK)**: Popular Python library for building NLP applications.
- **SpaCy**: Another library used for faster NLP processing.
- **Stanford NLP**: A suite of NLP tools for text analysis.



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
BESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Computational Lexicography in Practice

1. Building and Maintaining Lexicons:

Computational lexicography involves the creation of electronic dictionaries or lexicons using **corpus data**—large collections of text that reflect real-world language usage. The process starts with collecting **extensive text corpora, which are then analyzed to identify word frequencies, meanings, usage patterns, and contexts**. The goal is to build lexicons that are accurate, comprehensive, and representative of how language is used. Techniques like **corpus linguistics** are often employed to extract lexical information such as word senses, collocations, and grammatical behavior, making the lexicons useful for tasks like **machine translation, speech recognition, and sentiment analysis**.

Once a lexicon is created, it needs to be continually maintained. Words evolve over time, new terms emerge (e.g., from technology or pop culture), and meanings shift, making regular updates necessary. **Corpus-based approaches** allow **lexicographers to detect these changes, helping to keep the lexicon relevant and up-to-date**. Additionally, the maintenance process often involves adding multilingual support, integrating synonyms and antonyms, and enriching the entries with semantic annotations, such as word senses, synonyms, and usage notes.

2. Annotating and Enriching Dictionaries with Semantic Information:

To enhance the utility of lexicons in computational tasks, lexicographers annotate dictionaries with semantic information. This includes assigning word senses, labeling parts of speech, and identifying semantic relations like synonymy, antonymy, and hypernymy. For instance, resources like **WordNet** are built around semantic networks, where words are connected through relationships that provide context and depth to their meanings. This enrichment allows for better natural language understanding (NLU) by enabling more nuanced interpretations of text, crucial for tasks like **semantic search, question answering, and sentiment analysis**.



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Computational Lexicography in Practice

3. Use of Crowdsourcing and AI to Build Large-Scale Lexicons:

Crowdsourcing and AI are becoming increasingly essential in the development of large-scale lexicons. Platforms like **Wiktionary** and **Wordnik** have utilized crowdsourcing to expand their entries by allowing users to contribute definitions, examples, and usage notes. This democratization of lexicography enables rapid growth in the lexicon, especially for **emerging words and regional dialects**. **AI and machine learning techniques are also being used to automatically extract lexical data from corpora and web sources, significantly accelerating the lexicon-building process**. AI-driven models can learn patterns in language, such as common collocations or context-dependent meanings, and help scale up the creation of lexicons for **low-resource languages** or specialized domains (e.g., medical terminology).



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

These innovations make it possible to develop more expansive and contextually rich lexicons, supporting a wide range of NLP applications and improving the quality of language-related tasks like machine translation and text classification.



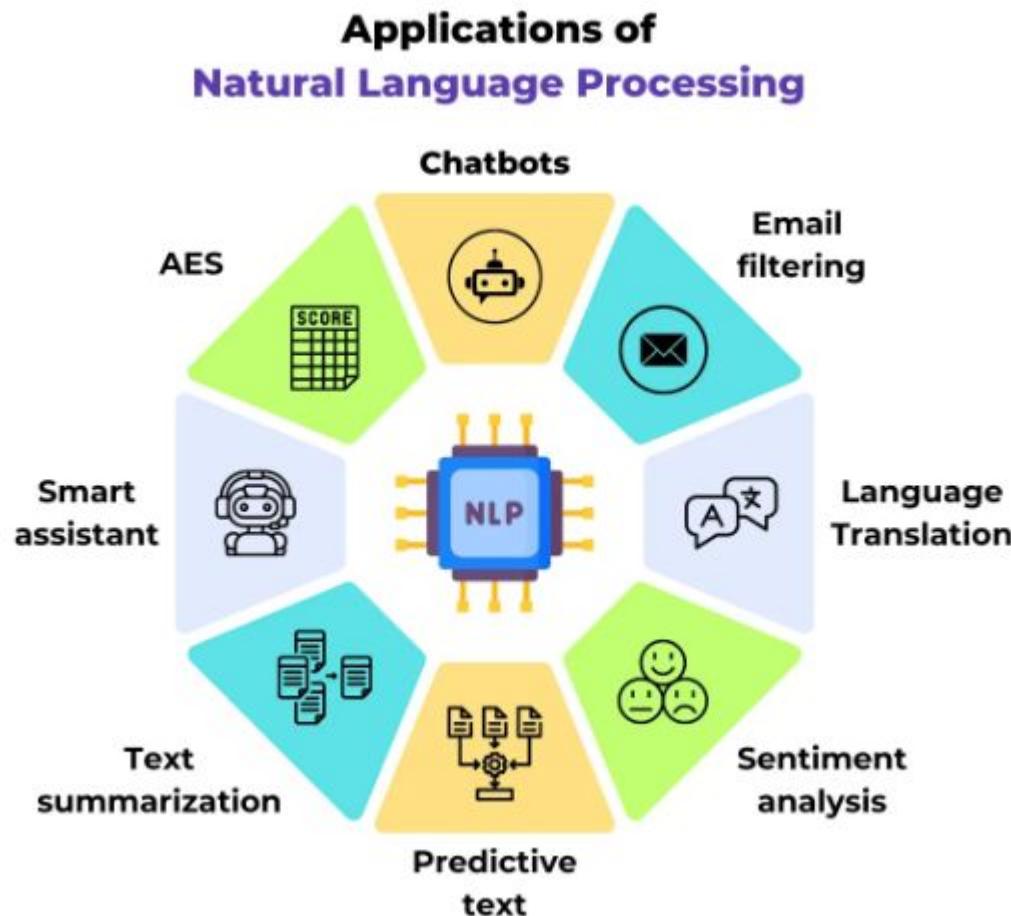
UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

**Faculty of Engineering,
Built Environment and
Information Technology**

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Applications of NLP and Computational Lexicography

- **Search Engines:** How computational lexicons improve search accuracy.
- **Chatbots and Virtual Assistants:** NLP-based interaction using lexicons for understanding.
- **Machine Translation:** Role of lexicons and syntactic parsing in translation systems.
- **Text Summarization:** Reducing large texts into concise summaries using NLP techniques.



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

**Faculty of Engineering,
Built Environment and
Information Technology**

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za

Challenges and Future Directions

- **Ambiguity in Language:** Handling polysemy, homonymy, and context sensitivity.
- **Multilingual Lexicons:** Building resources that support multiple languages effectively.
- **Explainable NLP:** Making NLP systems more interpretable and transparent, especially when using lexicons.



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Evaluation



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

**Faculty of Engineering,
Built Environment and
Information Technology**

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter
www.up.ac.za

The evaluation of computational lexicography models involves assessing their **accuracy, coverage, consistency, and usefulness for various linguistic and NLP tasks**. Below are key methods used to evaluate these models:

1. **Lexicon** Coverage:
Coverage measures how well the lexicon includes the vocabulary used in a specific corpus or domain. A high-coverage lexicon should contain most of the words and expressions found in the target corpus, including rare or domain-specific terms. Evaluators often compare the lexicon to reference corpora or test sets to **calculate the percentage of words included**.
2. **Precision** and Recall:
Like in machine learning, **precision** and **recall** are used to evaluate the quality of lexical entries. Precision measures the correctness of the entries (i.e., **how many of the words in the lexicon are accurate**), while recall measures how many of the actual words from the target dataset are included in the lexicon. **F1 score**, the harmonic mean of precision and recall, is often used for balanced evaluation.
3. **Sense Disambiguation** and Accuracy:
For lexicons that include word senses or polysemy (multiple meanings of a word), it is crucial to **evaluate how well the model assigns the correct meaning based on context**. Models are evaluated by **comparing their word sense assignments to a gold-standard dataset where human annotators have pre-identified the correct senses**. Tools like **Word Sense Disambiguation (WSD)** systems are commonly used to benchmark accuracy.
4. **Consistency** and Redundancy:
Lexicons are evaluated for internal consistency, ensuring that similar words or concepts are treated similarly throughout the lexicon. **Redundancy checks** are also carried out to avoid duplication of entries or conflicting meanings, which can degrade the quality of NLP tasks like machine translation and sentiment analysis.
5. **Application-Based** Evaluation:
A common method is to **test lexicon-based models within specific NLP tasks such as machine translation, sentiment analysis, or named entity recognition (NER)**. Evaluators **measure the performance improvements provided by the lexicon compared to other models or against baseline methods without lexicons**. For example, how well the lexicon enhances machine translation or sentiment analysis can be quantified by **BLEU scores** (for translation) or **accuracy** in sentiment classification.



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology
Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Benchmarking against Standard Lexical Resources

Computational lexicons can be evaluated by comparing them with established resources such as **WordNet**, **Oxford English Dictionary**, or **BabelNet**. Metrics such as overlap in vocabulary, semantic relations, and coverage of word senses provide insights into the model's performance.



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Computational lexicography models are evaluated both quantitatively (using statistical metrics) and qualitatively (through human evaluation and usability testing in NLP applications). This combined approach ensures that lexicons are both linguistically sound and practically useful in a wide range of AI-driven language tasks.



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za

VADER

Input	neg	neu	pos	compound
"This computer is a good deal."	0	0.58	0.42	0.44
"This computer is a very good deal."	0	0.61	0.39	0.49
"This computer is a very good deal!!!"	0	0.57	0.43	0.58
This computer is a very good deal!! :-)"	0	0.44	0.56	0.74
This computer is a VERY good deal!! :-)"	0	0.393	0.61	0.82

```
lexique = {  
    "amazing": 2.0,  
    "good": 1.5,  
    "okay": 0.0,  
    "bad": -1.5,  
    "terrible": -2.0  
}
```

```
def analyse_sentiment(text):  
    words = text.lower().split()  
  
    score = sum(lexique.get(word, 0) for word in words)  
  
    return score  
  
text = "The movie was amazing and the plot was good"  
print(analyse_sentiment(text))  
  
# Output: 3.5
```

Lexicon

Sentence	Positive score	Negative score	Binary prediction
I can play chess	+1	-1	+1
I can play chess!!!	+2	-1	+1
I like to read science fictions	+2	-1	+1
I do not like to read science fictions	+1	-1	+1
I left early because the film was boring	+1	-2	-1
I hate you	+1	-4	-1
I really love you, but dislike your cold sister	+4	-3	+1



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Code



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

**Faculty of Engineering,
Built Environment and
Information Technology**

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za

VADER Lexicon Code with Sentiment Analysis

```
import nltk
from nltk.sentiment.vader import SentimentIntensityAnalyzer

# Download VADER
nltk.download('vader_lexicon')

# Initialise sentiment analyser
sid = SentimentIntensityAnalyzer()

# Inputs texts
texts = [
    "I love this product! It's absolutely amazing.",
    "This is the worst experience I have ever had.",
    "I am not sure how I feel about this.",
    "It's okay, not the best but not the worst either.",
    "I am extremely happy with the service!",
]

# Function to analyze sentiments
def analyze_sentiment(text):
    scores = sid.polarity_scores(text)
    print(f"Text: {text}")
    print(f"Scores: {scores}")
    if scores['compound'] >= 0.05:
        print("Sentiment: Positive")
    elif scores['compound'] <= -0.05:
        print("Sentiment: Negative")
    else:
        print("Sentiment: Neutral")
    print("")

# Analyze the sentiments of the example texts
for text in texts:
    analyze_sentiment(text)
```

```
# Analyze the sentiments of the example texts
for text in texts:
    analyze_sentiment(text)
```

```
Text: I love this product! It's absolutely amazing.
Scores: {'neg': 0.0, 'neu': 0.318, 'pos': 0.682, 'compound': 0.862}
Sentiment: Positive
```

```
Text: This is the worst experience I have ever had.
Scores: {'neg': 0.369, 'neu': 0.631, 'pos': 0.0, 'compound': -0.6249}
Sentiment: Negative
```

```
Text: I am not sure how I feel about this.
Scores: {'neg': 0.246, 'neu': 0.754, 'pos': 0.0, 'compound': -0.2411}
Sentiment: Negative
```

```
Text: It's okay, not the best but not the worst either.
Scores: {'neg': 0.145, 'neu': 0.464, 'pos': 0.391, 'compound': 0.5729}
Sentiment: Positive
```

```
Text: I am extremely happy with the service!
Scores: {'neg': 0.0, 'neu': 0.539, 'pos': 0.461, 'compound': 0.6468}
Sentiment: Positive
```

```
[nltk_data] Downloading package vader_lexicon to
[nltk_data]     C:\Users\u21629545\AppData\Roaming\nltk_data...
[nltk_data]     Package vader_lexicon is already up-to-date!
```



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter
www.up.ac.za

Assignments

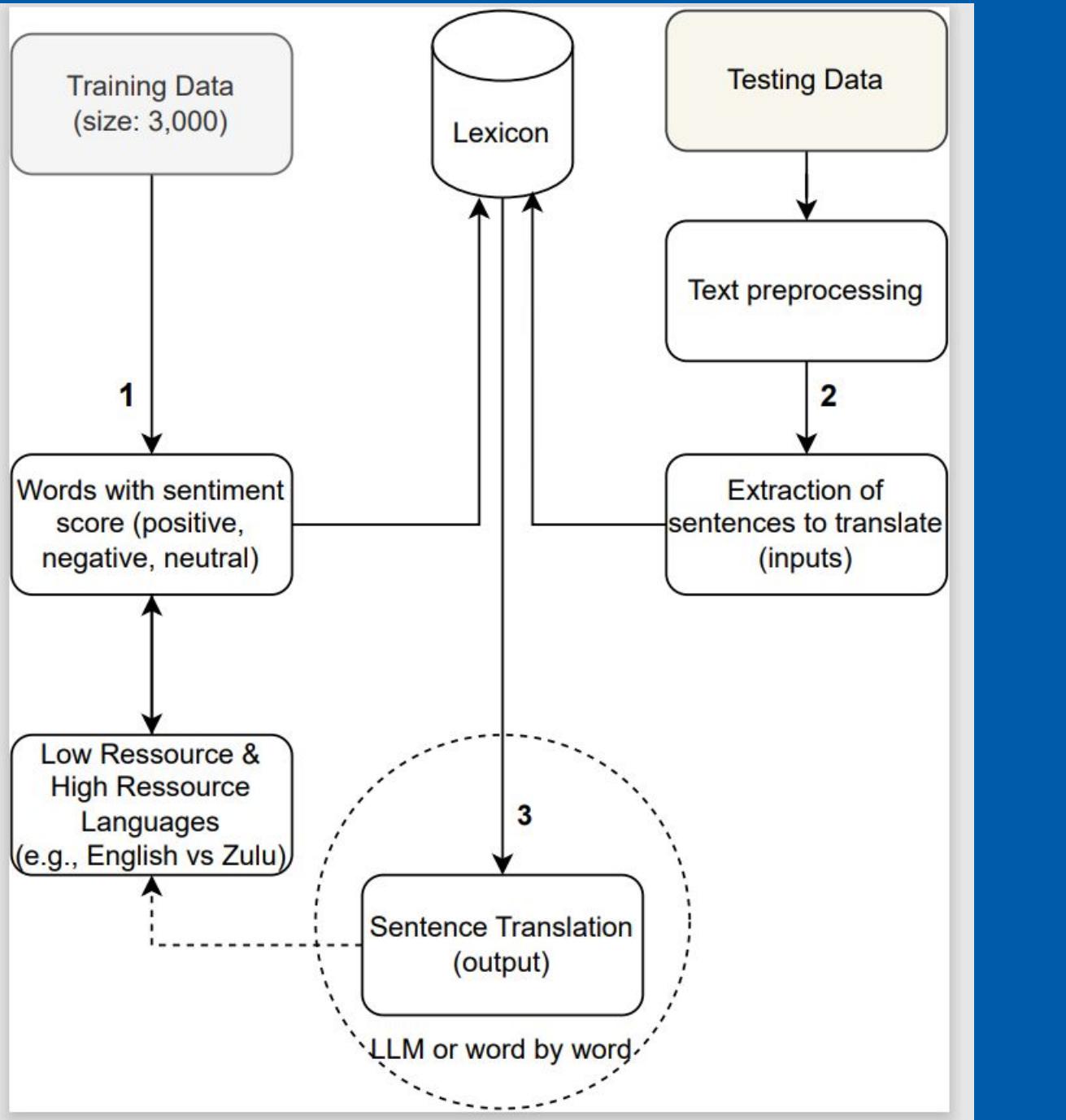


UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

**Faculty of Engineering,
Built Environment and
Information Technology**

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter
www.up.ac.za



Enjoy recess!



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

**Faculty of Engineering,
Built Environment and
Information Technology**

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za

Explainable Artificial Intelligence (XAI) & Large Language Models (LLMs)

By Dr. MIKE WA NKONGOLO
Department of Informatics, University of Pretoria



Content

1. Introduction to Explainable AI (XAI)

2. XAI Techniques

2.1. Popular XAI Techniques

2.2. Trade-offs in XAI

3. Introduction to Large Language Models (LLMs)

Content

3.1. What are LLMs?

3.2. Popular Examples of LLMs

3.3. Key Concepts in LLMs

3.3. Applications of LLMs

4. The Intersection of XAI and LLMs

4.1. Challenges with Interpretability in LLMs

4.2. Applying XAI to LLMs

4.3. Use Cases for XAI in LLMs

Content

5. Limitations and Future Directions

6. XAI Code Demonstration

7. Assignment 2 Discussion

Learning Outcomes

By the end of this lecture, students will be able to:

1. **Understand the Core Concepts of Explainable AI (XAI):**
 - Define Explainable AI (XAI) and articulate its importance in making AI systems transparent and interpretable.
 - Explain why XAI is crucial for ***trust, accountability, regulatory compliance, and model optimization*** in real-world applications.
2. **Identify and Apply Common XAI Techniques:**
 - Recognize and describe popular XAI methods such as SHAP, LIME, and Partial Dependence Plots (PDP).
 - Differentiate between ***model-specific and model-agnostic XAI techniques*** and explain the trade-offs between interpretability and model complexity.
3. **Comprehend the Architecture and Functionality of Large Language Models (LLMs):**
 - Explain the role of ***transformers in LLMs*** and describe key concepts such as self-attention and contextual understanding.
 - Understand the pretraining and fine-tuning process in LLMs and how it enables these models to perform complex language tasks.

Learning Outcomes

4. Evaluate the Applications and Capabilities of LLMs:

- Identify various applications of LLMs, including ***text generation, translation, sentiment analysis***, and question answering.
- Critically assess the performance of LLMs in different natural language processing tasks.

2. Analyze the Intersection of XAI and LLMs:

- Discuss the challenges of interpreting LLMs due to their complexity and size.
- Apply XAI techniques such as LIME and attention visualization to understand and explain the decision-making process of LLMs.

3. Understand the Ethical and Practical Implications of XAI and LLMs:

- Evaluate the limitations of both **XAI and LLMs, such as computational costs, model bias, and overfitting.**
- Discuss future research directions and the importance of integrating XAI into the development of fair and trustworthy LLMs.

Definition and Importance

- Explainable AI (XAI) refers to a set of techniques and methods that make the decisions, results, and outputs of AI or ML models interpretable and understandable by **humans**.
- The rise of complex AI/ML models (black-box), especially deep learning, has led to the need for transparency and trust.
- XAI addresses the "black-box" problem by providing insights into how models make decisions.

Definition and Importance

- XAI addresses the "black-box" problem by providing insights into how models make decisions.

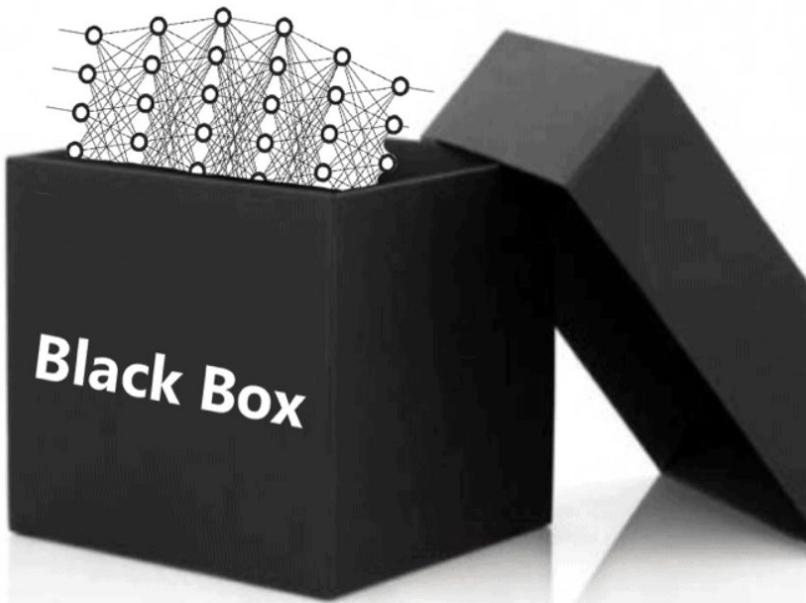
Key Characteristics of Black-Box Models

A **black-box model** refers to an AI or machine learning model whose internal workings are not easily interpretable or transparent to users.

In such models, we can observe the inputs/datasets and outputs/results, but we cannot easily understand how the model processes the inputs to produce the outputs.

These models are often complex, involving numerous layers of computations and parameters, making it challenging to explain why they make specific predictions or decisions.

Black-Box Models Architecture



Examples of Black-Box Models

- **Deep Neural Networks (DNNs):** DNNs consist of many layers, and their internal connections make it nearly impossible for a human to follow the logic behind individual decisions.
- **Ensemble Models:** Models like random forests and gradient boosting combine multiple decision trees, making the overall model's decision process hard to understand, even though individual trees may be interpretable.
- **Support Vector Machines (SVMs):** SVMs are hard to interpret when using complex kernel functions to transform data into high-dimensional spaces.

Why Black-Box Models are a Concern?



- **Trust:** Users may **hesitate to trust a system they cannot understand, especially in critical applications like medical diagnosis, finance, or legal decisions.**
- **Accountability:** In cases of errors or biases, it's difficult to determine what caused the incorrect outcome or who is responsible.
- **Ethical and Legal Issues:** Regulations, such as **GDPR, require transparency in automated decision-making systems, and black-box models often do not comply with these transparency requirements.**

Mitigating the Black-Box Nature with XAI

To address these challenges, Explainable AI (XAI) techniques are used to make black-box models more interpretable by providing explanations of how decisions are made.

Techniques like LIME, SHAP, and Attention Visualizations **allow users to gain insights into the model's decision process without fully exposing its complexity.**

Why XAI is Crucial

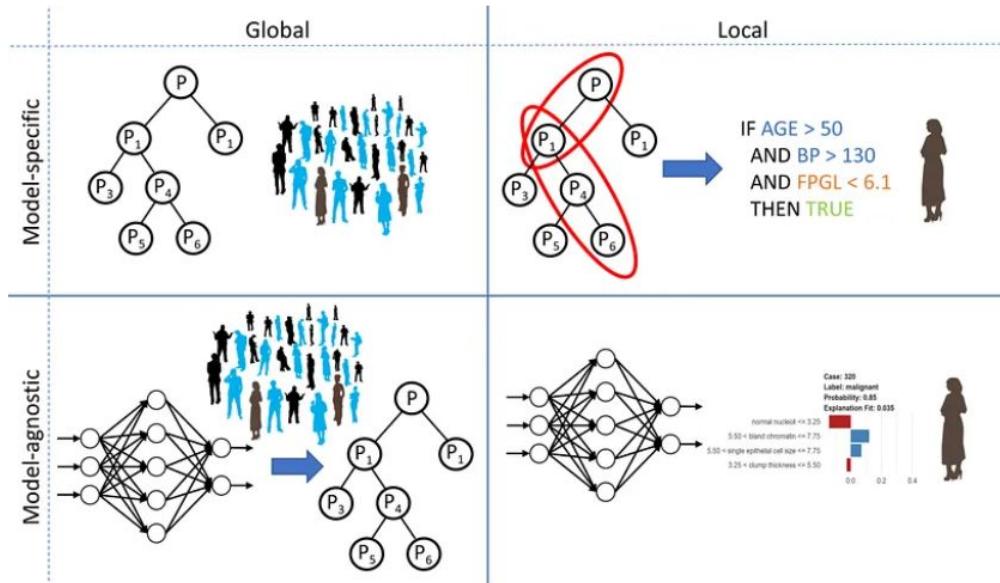
- **Trust and Accountability:** In fields like **healthcare, finance, and law**, users need to understand AI decisions to trust them and ensure ethical usage.
- **Debugging and Optimization:** Helps **data scientists and developers identify weaknesses or biases in models, improving their performance.**
- **Regulatory Compliance:** Compliance with regulations like **GDPR (General Data Protection Regulation) which require explanations for automated decisions.**

XAI Techniques

- Model-Specific vs Model-Agnostic
 - Model-Specific: Methods designed for specific types of models (e.g., decision trees, linear models) which are naturally interpretable.
 - Model-Agnostic: Techniques applicable to any machine learning model, especially black-box models like neural networks and random forests.

Said so, here we are going to focus on Post-Hoc Techniques, i.e. explanation methods that work on top of complex black-box methods. So that anyone can have fun with the ML model he prefers.

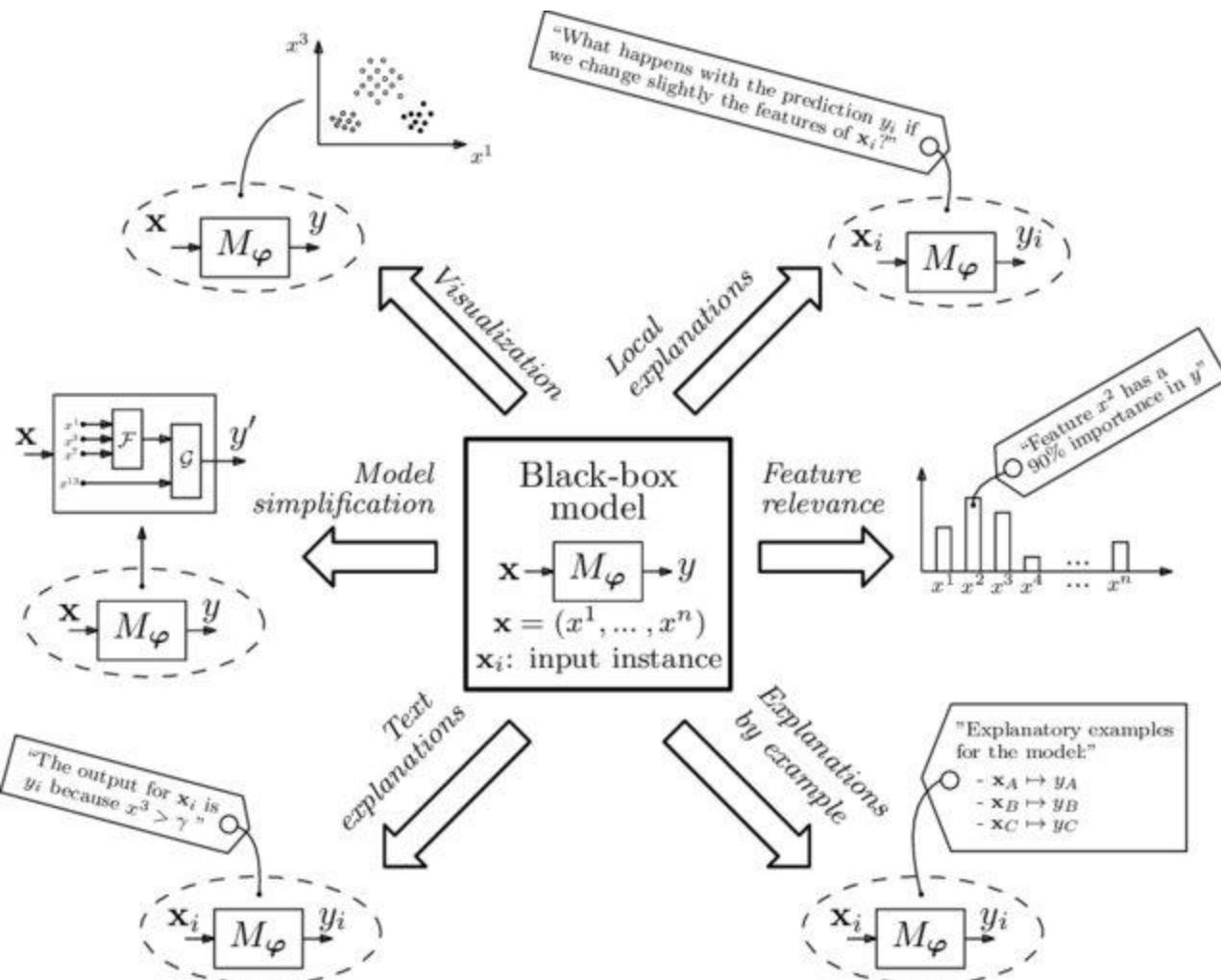
Post-Hoc Methods are partitioned following two main concepts: Model Agnostic/Model Specific and Local/Global techniques.



Model Agnostic techniques work for any kind of ML models, while Model Specific ones rely on a certain model structure. Global methods give an explanation for all the units in the dataset, whereas Local ones just for a bunch of dataset units (but you may always repeat the Local explanation on all the units of interest). Image credits to [Stiglic](#)

Pre-hoc and post-hoc Methods

Pre-hoc and **post-hoc** methods are two primary approaches in Explainable AI (XAI) used to explain the behavior of machine learning models. They differ in **when and how the explanations are generated, in relation to the model's training and prediction phases.**

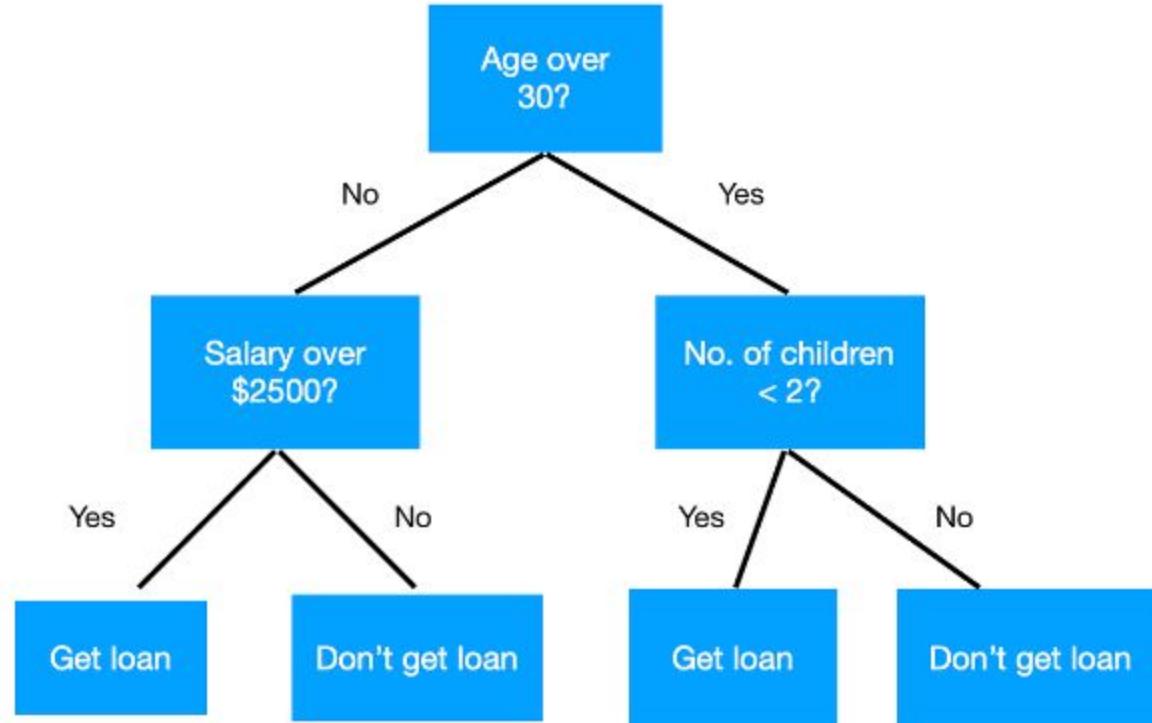


Pre-hoc Methods (Intrinsic Interpretability)

Pre-hoc methods, also known as **intrinsic** or **built-in interpretability** methods, are applied **before or during model training**. These methods focus on using inherently interpretable models, meaning that the models themselves are designed to be simple and transparent, allowing users to understand their decision-making process directly.

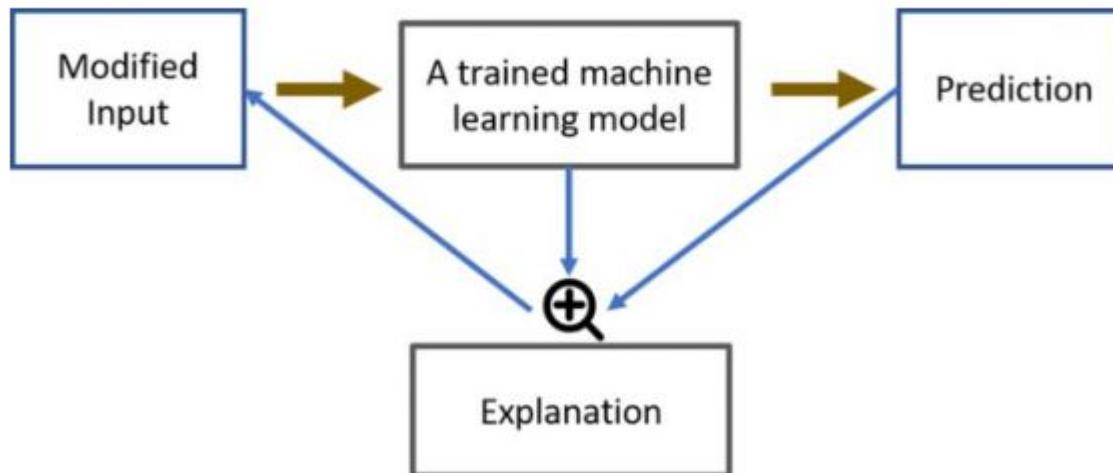
Characteristics of Pre-hoc Methods

- **Model Simplicity:** These methods rely on models that are simple enough for humans to easily interpret without the need for additional tools. Common models include:
 - **Linear regression**
 - **Logistic regression**
 - **Decision trees**
 - **Rule-based models**



Disadvantages

- Pre-hoc models often sacrifice accuracy and performance in complex tasks in favor of interpretability.



Post-hoc Methods (Post-training Interpretability)

Post-hoc methods, on the other hand, are applied **after the model has been trained**. These methods attempt to explain the decisions of complex, black-box models like deep neural networks, random forests, and ensemble methods, which are not inherently interpretable.

Post-hoc approaches provide explanations for specific predictions or model behaviors without altering the model itself.

Characteristics of Post-hoc Methods

- **Applied After Training:** These methods are used once the model has been trained, making them suitable for explaining black-box models.
- **Flexible Application:** They can be used with any type of model, making them more versatile than pre-hoc methods.
- **Local or Global Explanations:** Some post-hoc methods provide explanations for individual predictions (local), while others explain the overall model behavior (global).

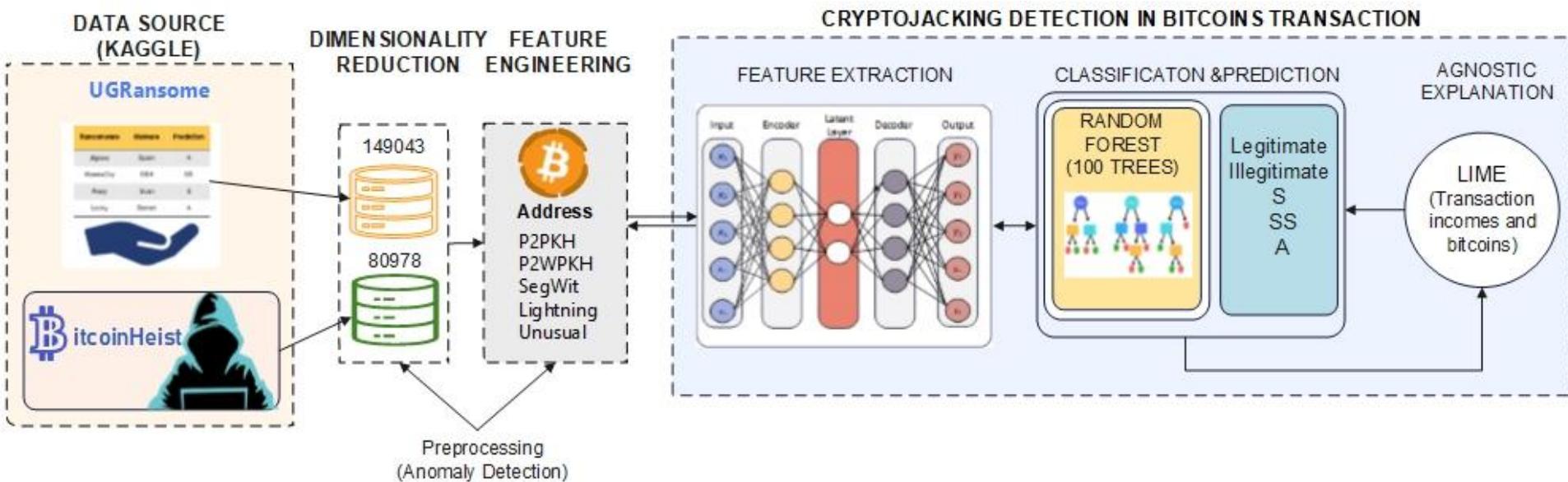
Common Post-hoc Techniques

- LIME (Local Interpretable Model-Agnostic Explanations): LIME **approximates the model locally using simpler, interpretable models (e.g., linear models, Random Forest)**, which help explain a single prediction.
- SHAP (SHapley Additive exPlanations): SHAP **assigns an importance score to each feature for a specific prediction based on cooperative game theory, offering local and global explanations**.
- Partial Dependence Plots (PDP): PDPs show how a feature affects the model's predictions on average, helping to explain the global behavior of the model.
- Feature Importance Scores: These scores rank the features based on how much they contribute to the model's predictions.

Examples of Post-hoc Methods

- **LIME:** When a neural network makes a classification decision, LIME can be used to approximate the network's behavior around a particular input by fitting an interpretable model (like a linear model) locally.
- **SHAP:** After training a random forest, SHAP can be used to compute the contribution of each feature to a specific prediction, enabling more understandable explanations.

Examples of Post-hoc Methods



Advantages:

- Can explain highly accurate and complex models (e.g., deep learning models) without changing their structure.
- Flexible and can be applied to any black-box model.

Disadvantages:

- Explanations may not always be perfect or fully accurate representations of how the black-box model works.
- Computational cost can be high, especially for large models and datasets.

Pre-hoc vs. Post-hoc Comparison

Pre-hoc vs. Post-hoc Comparison

Aspect	Pre-hoc Methods	Post-hoc Methods
When Applied	During model design and training	After the model is trained
Model Type	Simple, interpretable models	Complex, black-box models
Interpretability	Inherently interpretable	Requires external methods for interpretation
Examples	Linear regression, Decision trees, Logistic regression	LIME, SHAP, PDP, Feature Importance
Trade-offs	May sacrifice performance for transparency	More powerful models but require extra effort to explain
Use Case	Suitable for simpler tasks where interpretability is critical	Suitable for complex tasks requiring high performance and later explanation
Explainability Level	Global (the whole model is interpretable)	Can be local (specific to predictions) or global

Trade-offs in XAI

- **Accuracy vs Interpretability:** Simple models (e.g., decision trees) are interpretable but may be less accurate. Complex models (e.g., deep neural networks) are accurate but harder to explain.
- **Global vs Local Explanations:** Some techniques provide global explanations (understanding the model as a whole), while others focus on local explanations (understanding specific predictions).

Code

```
# Import necessary libraries
import pandas as pd
from sklearn.preprocessing import LabelEncoder
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
import lime
import lime.lime_tabular
```

Code

```
# Import necessary libraries
import pandas as pd
from sklearn.preprocessing import LabelEncoder
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
import lime
import lime.lime_tabular
```

Code

```
# Load Sample data to create a dataframe
```

```
data = {
```

```
}
```

Code

```
# Load Sample data to create a dataframe
```

```
data = {
```

```
}
```

Code

```
# Encode the data  
label_encoder = LabelEncoder()  
df['Address'] = label_encoder.fit_transform(df['Address'])
```

Code

```
# Split the data into features and target  
  
X = df.drop('Attack', axis=1)  
  
y = df['Attack']  
  
# Split the data into training and testing sets  
  
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)  
  
# Train a BLACK-BOX MODEL (E.g., RandomForest model)  
  
rf_model = RandomForestClassifier(random_state=42)  
  
rf_model.fit(X_train, y_train)
```

Code

```
# Initialize LIME explainer  
  
explainer = lime.lime_tabular.LimeTabularExplainer(  
  
    training_data=X_train.values,  
  
    feature_names=X_train.columns,  
  
    class_names=['Legitimate', 'Illegitimate'],  
  
    mode='classification'  
  
)  
  
# Choose an instance to explain  
  
instance_to_explain = X_test.iloc[0]
```

Code

```
# Generate explanation for the chosen instance  
explanation = explainer.explain_instance(  
    data_row=instance_to_explain.values,  
    predict_fn=rf_model.predict_proba  
)
```



Figure 10. Signature(S) attack prediction using LIME



Figure 11. Seed address prediction using LIME



Introduction to Large Language Models (LLMs)

What are LLMs?

- Large Language Models are **AI models trained on vast amounts of text data to understand, generate, and manipulate human language**. They are typically based on architectures like transformers and have billions of parameters, allowing them to perform complex natural language tasks.

Popular Examples of LLMs

- **GPT (Generative Pretrained Transformer)**: A family of models by **OpenAI**, including GPT-3 and GPT-4, known for tasks like text generation, translation, and summarization.
- **BERT (Bidirectional Encoder Representations from Transformers)**: A **transformer-based model that performs well on tasks like text classification and question answering**.
- **T5 (Text-to-Text Transfer Transformer)**: A unified model that frames all NLP tasks as text-to-text tasks, improving flexibility across various natural language tasks.

Key Concepts in LLMs

- **Transformers Architecture**
 - Introduced in the paper "Attention is All You Need" (2017), transformers rely on attention mechanisms, which allow models to weigh the importance of different words in a sequence, improving performance on long-range dependencies.
- **Pretraining and Fine-tuning**
 - **Pretraining:** LLMs are trained on large corpora of text data using unsupervised tasks (e.g., next-word prediction) to learn language representations.
 - **Fine-tuning:** Once pretrained, LLMs are adapted to specific tasks (e.g., sentiment analysis, summarization) using labeled datasets.

Self-Attention Mechanism

- A core component of transformers, self-attention allows the model to focus on different parts of the input sentence, capturing context more effectively.

Contextual Understanding

- Unlike earlier models (e.g., RNNs, LSTMs), LLMs capture bidirectional context, meaning they can understand the full context of a word based on its surrounding words, making them highly effective for complex NLP tasks.

Applications of LLMs

- **Text Generation:** LLMs can generate coherent and contextually appropriate text based on prompts, useful for chatbots, content creation, and automated storytelling.
- **Translation and Summarization:** Translate text from one language to another or summarize lengthy documents into concise versions.
- **Sentiment Analysis:** LLMs can analyze the sentiment behind text (e.g., positive, neutral, negative), useful in social media analysis, customer feedback, etc.
- **Question Answering:** Models like GPT-4 can answer factual questions, leveraging their vast knowledge base learned during pretraining.

The Intersection of XAI and LLMs

- **Challenges with Interpretability in LLMs**
 - Due to their massive size and complexity, LLMs are often **considered black-box models**, making it difficult to understand how they make specific decisions.
- **Applying XAI to LLMs**
 - Techniques like **LIME** and **SHAP** can be used to **explain why a certain word or phrase was important in an LLM's prediction.**
 - **Attention Visualization:** In transformer models, the **self-attention weights can be visualized to show which words the model focused on when making predictions.**

Use Cases for XAI in LLMs

- **Bias Detection:** XAI methods can help identify and mitigate biases in LLMs (e.g., gender or racial biases) by analyzing how different inputs affect the model's predictions.
- **Interpretation of Generated Text:** For sensitive applications (e.g., legal document generation), XAI can help ensure that LLMs generate appropriate and fair text.

Limitations and Future Directions

- **XAI Limitations**
 - Most XAI methods provide **approximations of model behavior, which may not fully explain the internal workings of complex models like LLMs.**
 - Some XAI techniques can be **computationally expensive or difficult to apply to models with billions of parameters.**
- **LLM Challenges**
 - **Data Bias:** LLMs often **reflect the biases present in the training data.**
 - **Overfitting to Language Patterns:** LLMs might **generate plausible but factually incorrect text because they optimize for fluency over factual accuracy.**

Future of XAI in LLMs

- The development of more robust and scalable XAI methods that can be applied to larger, more complex models.
- Integration of ethical and fairness considerations into both LLM training and XAI explanations to create more trustworthy AI systems.

Explainable AI (XAI) and Large Language Models (LLMs) represent two critical developments in the AI landscape. While LLMs offer impressive capabilities in understanding and generating human language, XAI plays a vital role in ensuring that these systems are transparent, accountable, and trustworthy. The intersection of XAI with LLMs is an active area of research, aiming to make advanced AI systems interpretable without sacrificing their performance. Understanding the trade-offs and challenges involved will be key to developing future AI systems that are both powerful and understandable.

Explainable Artificial Intelligence (XAI) & Large Language Models (LLMs)

By Dr. MIKE WA NKONGOLO
Department of Informatics, University of Pretoria



Content

1. Introduction to Explainable AI (XAI)

2. XAI Techniques

2.1. Popular XAI Techniques

2.2. Trade-offs in XAI

3. Introduction to Large Language Models (LLMs)

Content

3.1. What are LLMs?

3.2. Popular Examples of LLMs

3.3. Key Concepts in LLMs

3.3. Applications of LLMs

4. The Intersection of XAI and LLMs

4.1. Challenges with Interpretability in LLMs

4.2. Applying XAI to LLMs

4.3. Use Cases for XAI in LLMs

Content

5. Limitations and Future Directions

6. XAI Code Demonstration

7. Assignment 2 Discussion

Learning Outcomes

By the end of this lecture, students will be able to:

1. **Understand the Core Concepts of Explainable AI (XAI):**
 - Define Explainable AI (XAI) and articulate its importance in making AI systems transparent and interpretable.
 - Explain why XAI is crucial for ***trust, accountability, regulatory compliance, and model optimization*** in real-world applications.
2. **Identify and Apply Common XAI Techniques:**
 - Recognize and describe popular XAI methods such as SHAP, LIME, and Partial Dependence Plots (PDP).
 - Differentiate between ***model-specific and model-agnostic XAI techniques*** and explain the trade-offs between interpretability and model complexity.
3. **Comprehend the Architecture and Functionality of Large Language Models (LLMs):**
 - Explain the role of ***transformers in LLMs*** and describe key concepts such as self-attention and contextual understanding.
 - Understand the pretraining and fine-tuning process in LLMs and how it enables these models to perform complex language tasks.

Learning Outcomes

4. Evaluate the Applications and Capabilities of LLMs:

- Identify various applications of LLMs, including ***text generation, translation, sentiment analysis***, and question answering.
- Critically assess the performance of LLMs in different natural language processing tasks.

2. Analyze the Intersection of XAI and LLMs:

- Discuss the challenges of interpreting LLMs due to their complexity and size.
- Apply XAI techniques such as LIME and attention visualization to understand and explain the decision-making process of LLMs.

3. Understand the Ethical and Practical Implications of XAI and LLMs:

- Evaluate the limitations of both **XAI and LLMs, such as computational costs, model bias, and overfitting.**
- Discuss future research directions and the importance of integrating XAI into the development of fair and trustworthy LLMs.

Definition and Importance

- Explainable AI (XAI) refers to a set of techniques and methods that make the decisions, results, and outputs of AI or ML models interpretable and understandable by **humans**.
- The rise of complex AI/ML models (black-box), especially deep learning, has led to the need for transparency and trust.
- XAI addresses the "black-box" problem by providing insights into how models make decisions.

Definition and Importance

- XAI addresses the "black-box" problem by providing insights into how models make decisions.

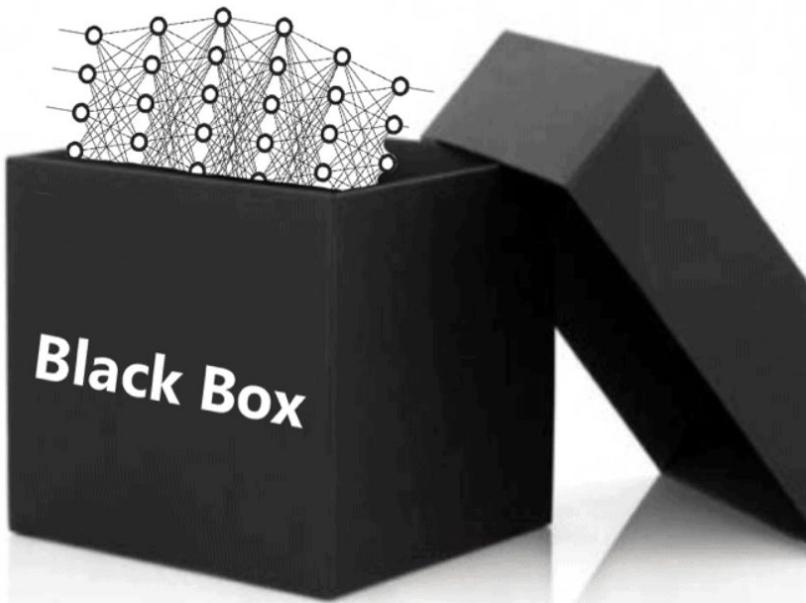
Key Characteristics of Black-Box Models

A **black-box model** refers to an AI or machine learning model whose internal workings are not easily interpretable or transparent to users.

In such models, we can observe the inputs/datasets and outputs/results, but we cannot easily understand how the model processes the inputs to produce the outputs.

These models are often complex, involving numerous layers of computations and parameters, making it challenging to explain why they make specific predictions or decisions.

Black-Box Models Architecture



Examples of Black-Box Models

- **Deep Neural Networks (DNNs):** DNNs consist of many layers, and their internal connections make it nearly impossible for a human to follow the logic behind individual decisions.
- **Ensemble Models:** Models like random forests and gradient boosting combine multiple decision trees, making the overall model's decision process hard to understand, even though individual trees may be interpretable.
- **Support Vector Machines (SVMs):** SVMs are hard to interpret when using complex kernel functions to transform data into high-dimensional spaces.

Why Black-Box Models are a Concern?



- **Trust:** Users may **hesitate to trust a system they cannot understand, especially in critical applications like medical diagnosis, finance, or legal decisions.**
- **Accountability:** In cases of errors or biases, it's difficult to determine what caused the incorrect outcome or who is responsible.
- **Ethical and Legal Issues:** Regulations, such as **GDPR, require transparency in automated decision-making systems, and black-box models often do not comply with these transparency requirements.**

Mitigating the Black-Box Nature with XAI

To address these challenges, Explainable AI (XAI) techniques are used to make black-box models more interpretable by providing explanations of how decisions are made.

Techniques like LIME, SHAP, and Attention Visualizations **allow users to gain insights into the model's decision process without fully exposing its complexity.**

Why XAI is Crucial

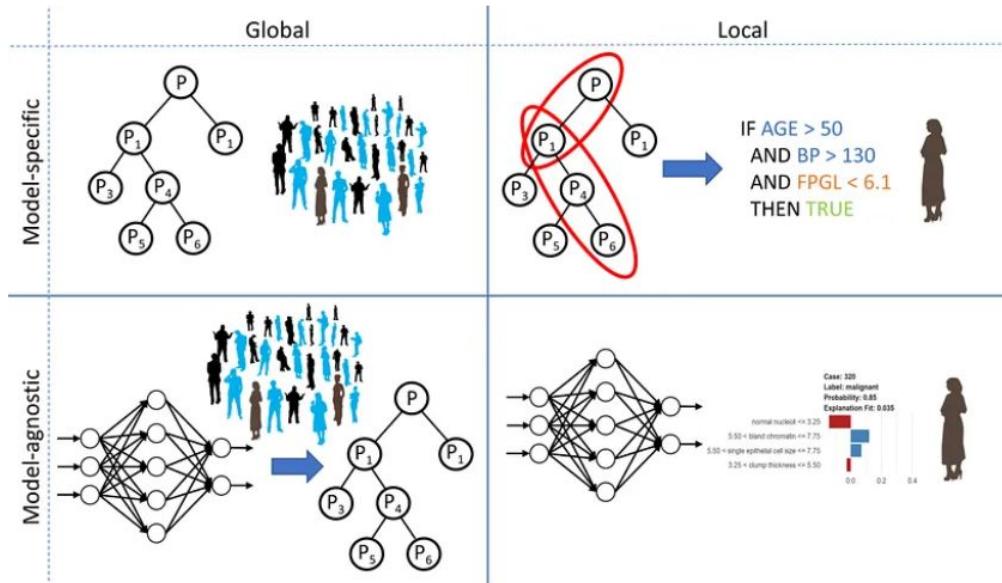
- **Trust and Accountability:** In fields like **healthcare, finance, and law**, users need to understand AI decisions to trust them and ensure ethical usage.
- **Debugging and Optimization:** Helps **data scientists and developers identify weaknesses or biases in models, improving their performance.**
- **Regulatory Compliance:** Compliance with regulations like **GDPR (General Data Protection Regulation) which require explanations for automated decisions.**

XAI Techniques

- Model-Specific vs Model-Agnostic
 - Model-Specific: Methods designed for specific types of models (e.g., decision trees, linear models) which are naturally interpretable.
 - Model-Agnostic: Techniques applicable to any machine learning model, especially black-box models like neural networks and random forests.

Said so, here we are going to focus on Post-Hoc Techniques, i.e. explanation methods that work on top of complex black-box methods. So that anyone can have fun with the ML model he prefers.

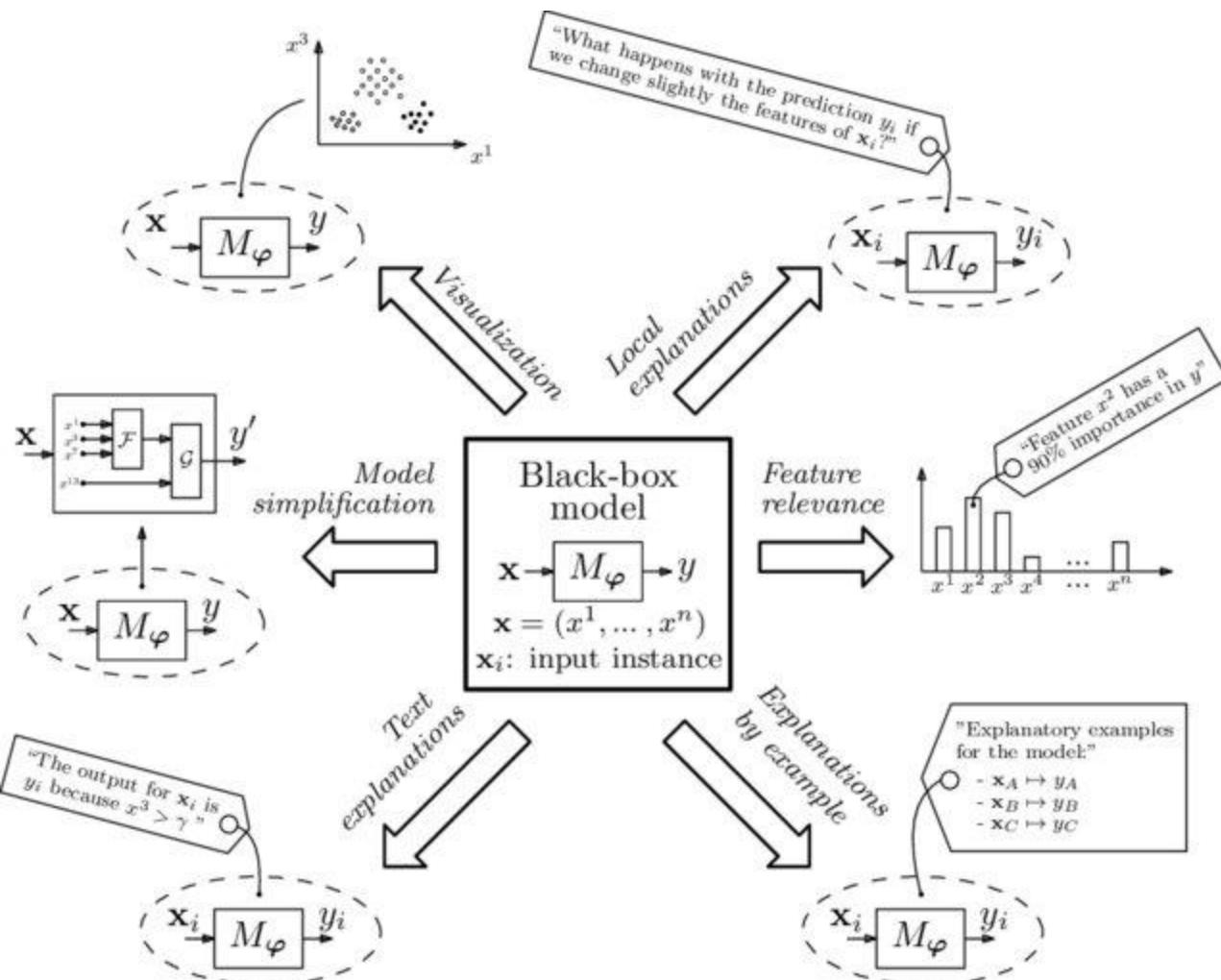
Post-Hoc Methods are partitioned following two main concepts: Model Agnostic/Model Specific and Local/Global techniques.



Model Agnostic techniques work for any kind of ML models, while Model Specific ones rely on a certain model structure. Global methods give an explanation for all the units in the dataset, whereas Local ones just for a bunch of dataset units (but you may always repeat the Local explanation on all the units of interest). Image credits to [Stiglic](#)

Pre-hoc and post-hoc Methods

Pre-hoc and **post-hoc** methods are two primary approaches in Explainable AI (XAI) used to explain the behavior of machine learning models. They differ in **when and how the explanations are generated, in relation to the model's training and prediction phases.**

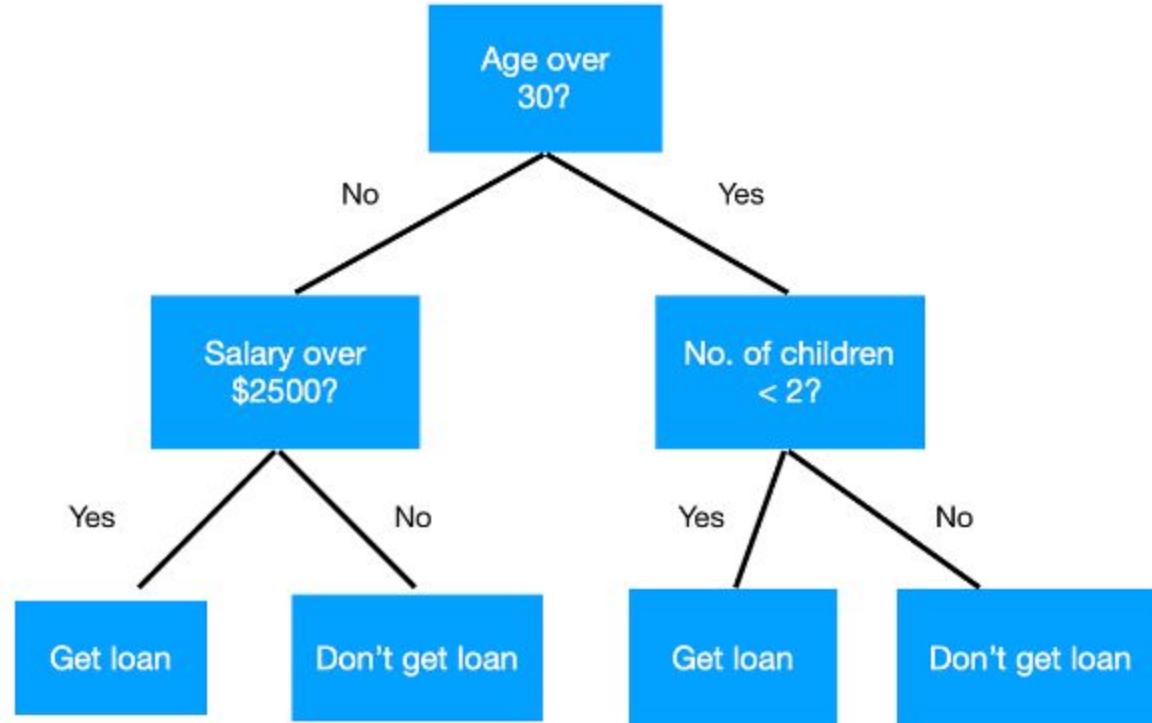


Pre-hoc Methods (Intrinsic Interpretability)

Pre-hoc methods, also known as **intrinsic** or **built-in interpretability** methods, are applied **before or during model training**. These methods focus on using inherently interpretable models, meaning that the models themselves are designed to be simple and transparent, allowing users to understand their decision-making process directly.

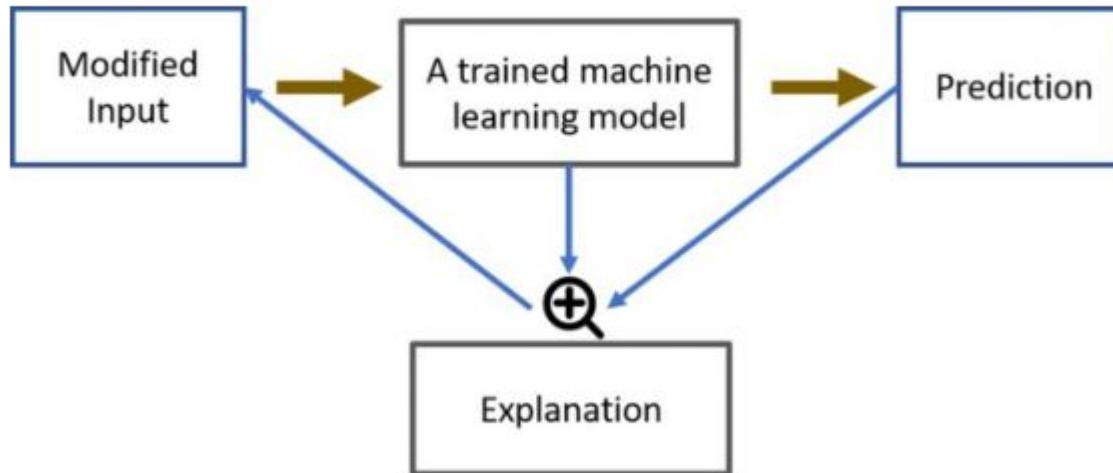
Characteristics of Pre-hoc Methods

- **Model Simplicity:** These methods rely on models that are simple enough for humans to easily interpret without the need for additional tools. Common models include:
 - **Linear regression**
 - **Logistic regression**
 - **Decision trees**
 - **Rule-based models**



Disadvantages

- Pre-hoc models often sacrifice accuracy and performance in complex tasks in favor of interpretability.



Post-hoc Methods (Post-training Interpretability)

Post-hoc methods, on the other hand, are applied **after the model has been trained**. These methods attempt to explain the decisions of complex, black-box models like deep neural networks, random forests, and ensemble methods, which are not inherently interpretable.

Post-hoc approaches provide explanations for specific predictions or model behaviors without altering the model itself.

Characteristics of Post-hoc Methods

- **Applied After Training:** These methods are used once the model has been trained, making them suitable for explaining black-box models.
- **Flexible Application:** They can be used with any type of model, making them more versatile than pre-hoc methods.
- **Local or Global Explanations:** Some post-hoc methods provide explanations for individual predictions (local), while others explain the overall model behavior (global).

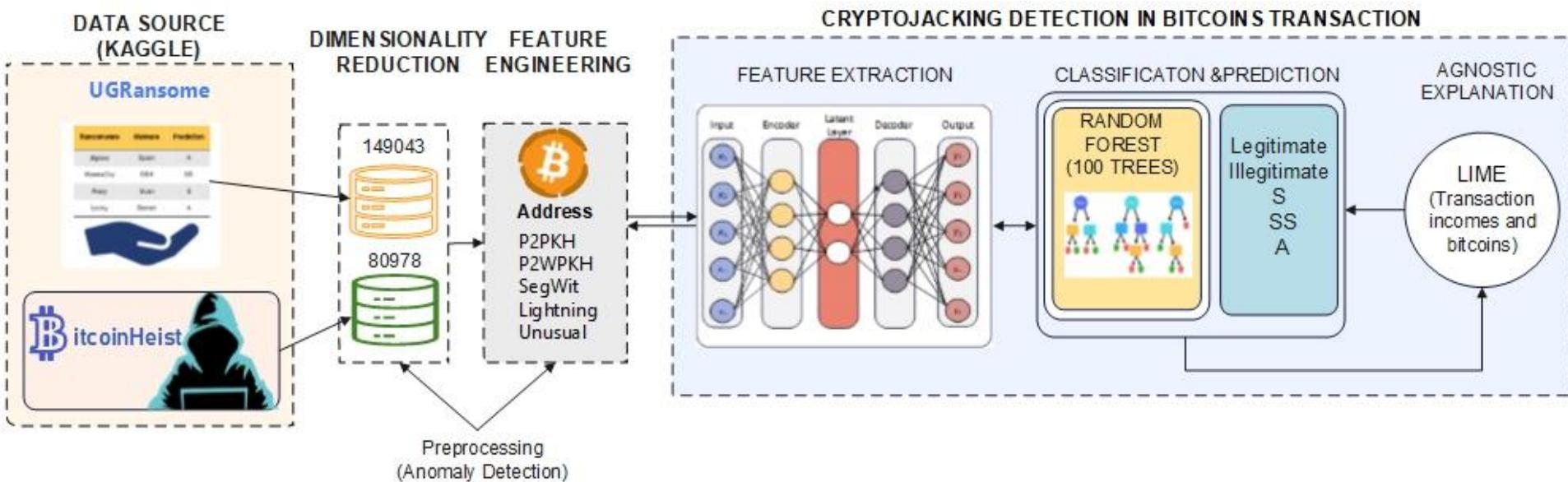
Common Post-hoc Techniques

- LIME (Local Interpretable Model-Agnostic Explanations): LIME **approximates the model locally using simpler, interpretable models (e.g., linear models, Random Forest)**, which help explain a single prediction.
- SHAP (SHapley Additive exPlanations): SHAP **assigns an importance score to each feature for a specific prediction based on cooperative game theory, offering local and global explanations**.
- Partial Dependence Plots (PDP): PDPs show how a feature affects the model's predictions on average, helping to explain the global behavior of the model.
- Feature Importance Scores: These scores rank the features based on how much they contribute to the model's predictions.

Examples of Post-hoc Methods

- **LIME:** When a neural network makes a classification decision, LIME can be used to approximate the network's behavior around a particular input by fitting an interpretable model (like a linear model) locally.
- **SHAP:** After training a random forest, SHAP can be used to compute the contribution of each feature to a specific prediction, enabling more understandable explanations.

Examples of Post-hoc Methods



Advantages:

- Can explain highly accurate and complex models (e.g., deep learning models) without changing their structure.
- Flexible and can be applied to any black-box model.

Disadvantages:

- Explanations may not always be perfect or fully accurate representations of how the black-box model works.
- Computational cost can be high, especially for large models and datasets.

Pre-hoc vs. Post-hoc Comparison

Pre-hoc vs. Post-hoc Comparison

Aspect	Pre-hoc Methods	Post-hoc Methods
When Applied	During model design and training	After the model is trained
Model Type	Simple, interpretable models	Complex, black-box models
Interpretability	Inherently interpretable	Requires external methods for interpretation
Examples	Linear regression, Decision trees, Logistic regression	LIME, SHAP, PDP, Feature Importance
Trade-offs	May sacrifice performance for transparency	More powerful models but require extra effort to explain
Use Case	Suitable for simpler tasks where interpretability is critical	Suitable for complex tasks requiring high performance and later explanation
Explainability Level	Global (the whole model is interpretable)	Can be local (specific to predictions) or global

Trade-offs in XAI

- **Accuracy vs Interpretability:** Simple models (e.g., decision trees) are interpretable but may be less accurate. Complex models (e.g., deep neural networks) are accurate but harder to explain.
- **Global vs Local Explanations:** Some techniques provide global explanations (understanding the model as a whole), while others focus on local explanations (understanding specific predictions).

Code

```
# Import necessary libraries
import pandas as pd
from sklearn.preprocessing import LabelEncoder
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
import lime
import lime.lime_tabular
```

Code

```
# Import necessary libraries
import pandas as pd
from sklearn.preprocessing import LabelEncoder
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
import lime
import lime.lime_tabular
```

Code

```
# Load Sample data to create a dataframe
```

```
data = {
```

```
}
```

Code

```
# Load Sample data to create a dataframe
```

```
data = {
```

```
}
```

Code

```
# Encode the data  
label_encoder = LabelEncoder()  
df['Address'] = label_encoder.fit_transform(df['Address'])
```

Code

```
# Split the data into features and target  
  
X = df.drop('Attack', axis=1)  
  
y = df['Attack']  
  
# Split the data into training and testing sets  
  
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)  
  
# Train a BLACK-BOX MODEL (E.g., RandomForest model)  
  
rf_model = RandomForestClassifier(random_state=42)  
  
rf_model.fit(X_train, y_train)
```

Code

```
# Initialize LIME explainer  
  
explainer = lime.lime_tabular.LimeTabularExplainer(  
  
    training_data=X_train.values,  
  
    feature_names=X_train.columns,  
  
    class_names=['Legitimate', 'Illegitimate'],  
  
    mode='classification'  
  
)  
  
# Choose an instance to explain  
  
instance_to_explain = X_test.iloc[0]
```

Code

```
# Generate explanation for the chosen instance  
explanation = explainer.explain_instance(  
    data_row=instance_to_explain.values,  
    predict_fn=rf_model.predict_proba  
)
```



Figure 10. Signature(S) attack prediction using LIME

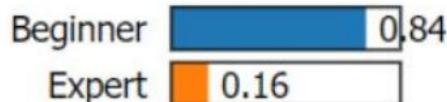


Figure 11. Seed address prediction using LIME



Figure 10: Features influencing the prediction of the ML models

Prediction probabilities



Beginner



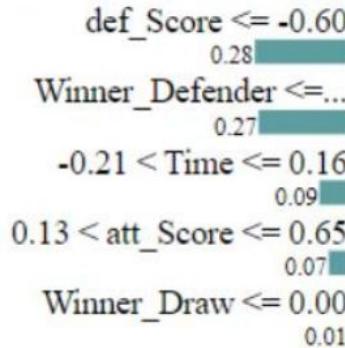
Feature Value

Feature	Value
score	51.00
time	58.00
winner	0.00

Prediction probabilities



NOT Beginner



Beginner

Feature Value

Feature	Value
def_Score	-0.60
Winner_Defender	0.00
Time	-0.15
att_Score	0.65
Winner_Draw	0.00

Introduction to Large Language Models (LLMs)

What are LLMs?

- Large Language Models are **AI models trained on vast amounts of text data to understand, generate, and manipulate human language**. They are typically based on architectures like transformers and have billions of parameters, allowing them to perform complex natural language tasks.

Popular Examples of LLMs

- **GPT (Generative Pretrained Transformer)**: A family of models by **OpenAI**, including GPT-3 and GPT-4, known for tasks like text generation, translation, and summarization.
- **BERT (Bidirectional Encoder Representations from Transformers)**: A **transformer-based model that performs well on tasks like text classification and question answering**.
- **T5 (Text-to-Text Transfer Transformer)**: A unified model that frames all NLP tasks as text-to-text tasks, improving flexibility across various natural language tasks.

Key Concepts in LLMs

- **Transformers Architecture**
 - Introduced in the paper "Attention is All You Need" (2017), transformers rely on attention mechanisms, which allow models to weigh the importance of different words in a sequence, improving performance on long-range dependencies.
- **Pretraining and Fine-tuning**
 - **Pretraining:** LLMs are trained on large corpora of text data using unsupervised tasks (e.g., next-word prediction) to learn language representations.
 - **Fine-tuning:** Once pretrained, LLMs are adapted to specific tasks (e.g., sentiment analysis, summarization) using labeled datasets.

Self-Attention Mechanism

- A core component of transformers, self-attention allows the model to focus on different parts of the input sentence, capturing context more effectively.

Contextual Understanding

- Unlike earlier models (e.g., RNNs, LSTMs), LLMs capture bidirectional context, meaning they can understand the full context of a word based on its surrounding words, making them highly effective for complex NLP tasks.

Applications of LLMs

- **Text Generation:** LLMs can generate coherent and contextually appropriate text based on prompts, useful for chatbots, content creation, and automated storytelling.
- **Translation and Summarization:** Translate text from one language to another or summarize lengthy documents into concise versions.
- **Sentiment Analysis:** LLMs can analyze the sentiment behind text (e.g., positive, neutral, negative), useful in social media analysis, customer feedback, etc.
- **Question Answering:** Models like GPT-4 can answer factual questions, leveraging their vast knowledge base learned during pretraining.

The Intersection of XAI and LLMs

- **Challenges with Interpretability in LLMs**
 - Due to their massive size and complexity, LLMs are often **considered black-box models**, making it difficult to understand how they make specific decisions.
- **Applying XAI to LLMs**
 - Techniques like **LIME** and **SHAP** can be used to **explain why a certain word or phrase was important in an LLM's prediction.**
 - **Attention Visualization:** In transformer models, the **self-attention weights can be visualized to show which words the model focused on when making predictions.**

Use Cases for XAI in LLMs

- **Bias Detection:** XAI methods can help identify and mitigate biases in LLMs (e.g., gender or racial biases) by analyzing how different inputs affect the model's predictions.
- **Interpretation of Generated Text:** For sensitive applications (e.g., legal document generation), XAI can help ensure that LLMs generate appropriate and fair text.

Limitations and Future Directions

- **XAI Limitations**
 - Most XAI methods provide **approximations of model behavior, which may not fully explain the internal workings of complex models like LLMs.**
 - Some XAI techniques can be **computationally expensive or difficult to apply to models with billions of parameters.**
- **LLM Challenges**
 - **Data Bias:** LLMs often **reflect the biases present in the training data.**
 - **Overfitting to Language Patterns:** LLMs might **generate plausible but factually incorrect text because they optimize for fluency over factual accuracy.**

Future of XAI in LLMs

- The development of more robust and scalable XAI methods that can be applied to larger, more complex models.
- Integration of ethical and fairness considerations into both LLM training and XAI explanations to create more trustworthy AI systems.

Explainable AI (XAI) and Large Language Models (LLMs) represent two critical developments in the AI landscape. While LLMs offer impressive capabilities in understanding and generating human language, XAI plays a vital role in ensuring that these systems are transparent, accountable, and trustworthy. The intersection of XAI with LLMs is an active area of research, aiming to make advanced AI systems interpretable without sacrificing their performance. Understanding the trade-offs and challenges involved will be key to developing future AI systems that are both powerful and understandable.

Assignment 4

See ClickUP.

Application of XAI on RTIA data.

Thank you !





ENGINEERING 4.0
UNIVERSITY OF PRETORIA



INF 491/791: Applied Data Science



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter
www.up.ac.za

L 01: Introduction to Data Science



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za

Outline

- ❖ Welcome
- ❖ Lecturing Team
- ❖ About INF 491/791
- ❖ What is Data Science?
- ❖ Software Requirements
- ❖ Assignments
- ❖ Additions
- ❖ Conclusion



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Welcome

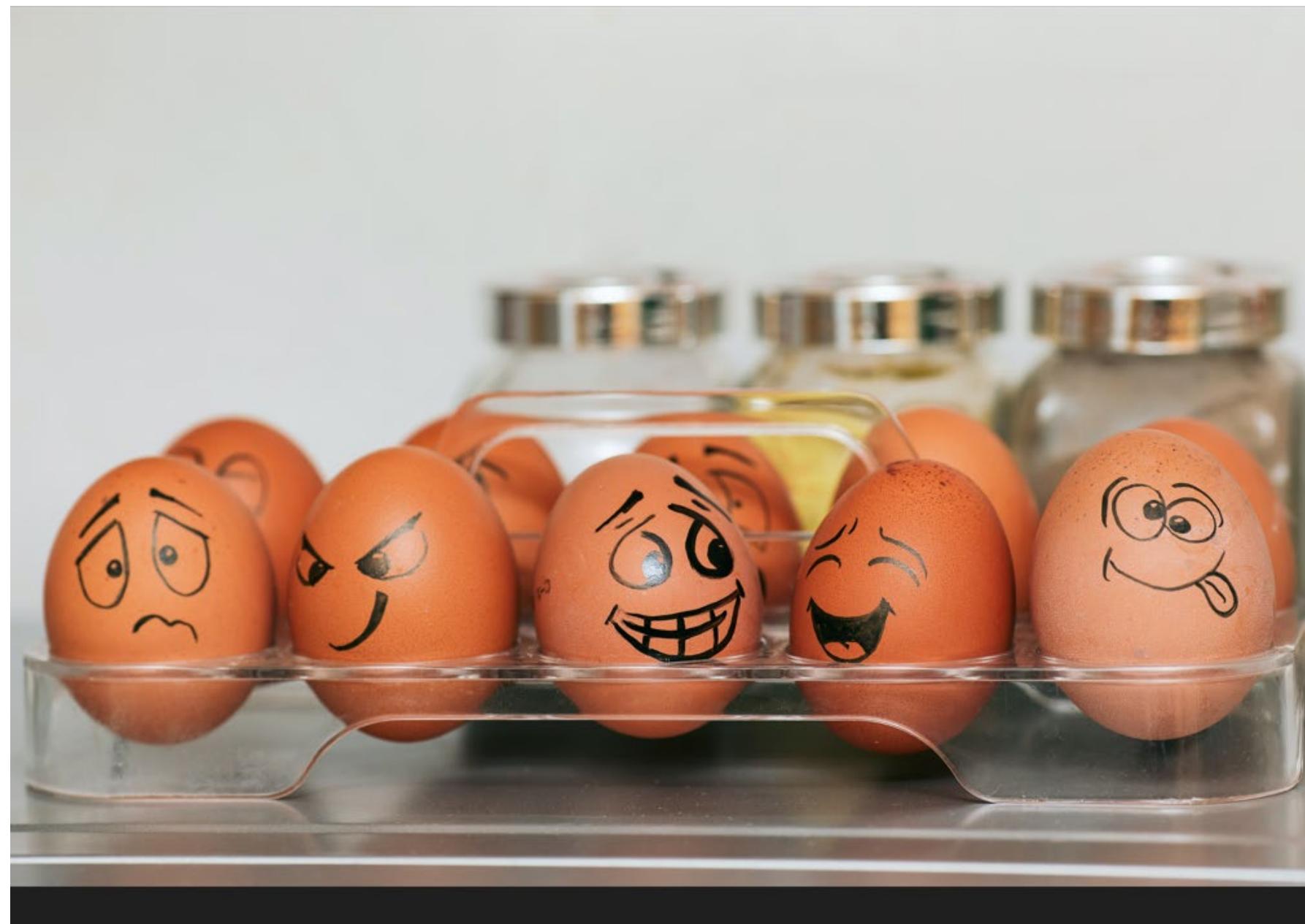


UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

**Faculty of Engineering,
Built Environment and
Information Technology**

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter
www.up.ac.za



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering, Built Environment and Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za

Lecturing Team



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

**Faculty of Engineering,
Built Environment and
Information Technology**

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter
www.up.ac.za

About INF 491/791



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za

Contact Information



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

**Faculty of Engineering,
Built Environment and
Information Technology**

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter
www.up.ac.za

Lectures

The lectures are scheduled in the following manner:

DAY	TIME	VENUE
Friday	13:30 – 15:00	IT Building, Room 2-27



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

**Faculty of Engineering,
Built Environment and
Information Technology**

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Purpose of the module

- One of the occupations that is most likely to prosper in this century is **data scientist**. It is computational, digital, and programming-oriented. Therefore, it should be clear that there is a growing need for data scientists in the labour market. **But the supply has been highly constrained**. The knowledge required to become a data scientist is challenging to obtain.
- This module aims to equip the student with some of the necessary data science competencies. These competencies contribute to applying data science practices and developing, e.g., *algorithms*, *models*, and *solutions*. Furthermore, the knowledge and skills obtained through this module will better prepare the student for potential job opportunities in the industry.



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology
Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Module Outcomes

- **MO1:** Have a good understanding of data science.
- **MO2:** Understand data science concepts such as statistical analysis, machine learning, data mining, and neural networks.
- **MO3:** Use data science tools and Python to do basic data science tasks.
- **MO4:** Implement the basics of data science concepts such as statistical analysis, machine learning, data mining, and neural networks.
- **MO5:** Understand the importance of ethical considerations in data science.
- **MO6:** Implement and demonstrate data science solutions.



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology
Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Module Structure

- Introduction to Data Science
- Data Collection and Preparation
- Data Exploration and Visualization
- Applied Statistical Analysis
- Machine Learning
- Data Mining and Big Data
- Neural Networks and Deep Learning
- Ethical Considerations in Data Science



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Mark Allocation

Pass requirements:

- A sub-minimum of 40% as a semester mark is needed to access the exam.
- According to faculty regulations, you must obtain a final mark of 50% or more to pass this module (see below).

Module Final Mark Percentages [as calculated at the end of the exam]	
Semester mark	50%
Exam mark	50%
Final module average [need to obtain a minimum of 40% in the exam and a final mark of 50% or more to pass this module]	100%



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Schedule (with Flow)

WEEK OF	DATE	SCHEDULE TYPE	DETAILS	SCOPE	WEIGHT
23/07 to 29/07	FRI:28	Lecture (L)	L 01 Introduction to Data Science	N/A	N/A
06/08 to 12/08	FRI:11	ClickUP Quiz (Q)	Q 01 – DUE: 11:59 AM	Data Collection and Preparation	1%
06/08 to 12/08	FRI:11	L	L 02 Data Collection and Preparation	N/A	N/A
20/08 to 26/08	FRI:25	ClickUP Q.	Q 02 – DUE: 11:59 AM	Data Exploration and Visualization	1%
20/08 to 26/08	FRI:25	L	L 03 Data Exploration and Visualization	N/A	N/A
03/09 to 09/09	FRI:08	ClickUP Q.	Q 03 – DUE: 11:59 AM	Applied Statistical Analysis	1%
03/09 to 09/09	FRI:08	L	L 04 Applied Statistical Analysis	N/A	N/A
10/09 to 16/09	MON:11	Assignment (A)	A 01 – DUE: 17:00 PM	L 02, 03, 04	25%
24/09 to 30/09	FRI:29	ClickUP Q.	Q 04 – DUE: 11:59 AM	Machine Learning	2%
24/09 to 30/09	FRI:29	L	L 05 Machine Learning	N/A	N/A
01/10 to 07/10	FRI:06	Sem Test	(PRELIM) TIME: 08:30 TO 11:30	L 02, 03, 04, 05	40%
08/10 to 14/10	FRI:13	ClickUP Q.	Q 05 – DUE: 11:59 AM	Data Mining and Big Data	2%
08/10 to 14/10	FRI:13	L	L 06 Data Mining and Big Data	N/A	N/A
22/10 to 28/10	FRI:27	ClickUP Q.	Q 06 – DUE: 11:59 AM	Neural Networks and Deep Learning	2%
22/10 to 28/10	FRI:27	L	L 07 Neural Networks and Deep Learning	N/A	N/A
29/10 to 04/11	MON:30	A	A 02 – DUE: 17:00 PM	L 02, 03, 04, 05, 06, 07	25%
29/11 to 24/11	FRI:03	ClickUP Q.	Q 07 – DUE: 11:59 AM	Ethical Considerations in Data Science	1%

WEEK OF	DATE	SCHEDULE TYPE	DETAILS	SCOPE	WEIGHT
29/11 to 24/11	FRI:03	L	L 08 Ethical Considerations in Data Science	N/A	N/A
19/11 to 25/11	FRI:24	Exam	(PRELIM) TIME: 07:30 TO 10:30 <i>Note: INF 491/791 exam date and time will be finalized/confirmed 2 weeks before it is to be written. The lecturers do not decide on this date and time.</i>	Everything	50%



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Schedule (with Flow) continued...

- Be aware of upcoming assignments, tests, and examination (see Study Guide and announcements of availability, etc.)



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

**Faculty of Engineering,
Built Environment and
Information Technology**
Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Assessment Criteria

- Computer-based assessments on ClickUP are used for quizzes. For the semester test(s) and examination(s), memorandums will be used to assess students' answers, while rubrics and memos will be used for assignments. All of these assessments are aligned with the learning outcomes of this module.
- At present, the venue (method) for writing the semester test(s) and examination(s) is an Informatorium Lab using ClickUP for submissions.



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology
Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Assessment Policy

- Please refer to the **Departmental Brochure** for general rules related to assessment. Review the departmental brochure for all departmental rules and requirements as listed and associated with University regulations and requirements. Test and assignment information and dates are available on ClickUP. Please make a note of these dates. **Please submit all assessments on time. No late submissions will be accepted. All assessments contribute towards your final mark.**



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology
Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Plagiarism

- Plagiarism is a **serious** form of academic misconduct. It involves appropriating someone else's work and passing it off as one's work afterward. Thus, you commit plagiarism when you present someone else's written or creative work (words, images, ideas, opinions, discoveries, artwork, music, recordings, computer-generated work, etc.) as your own.
Only hand in your original work.



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology
Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Support Services

Disability Unit	Academic support for students with learning disabilities: <ul style="list-style-type: none">▪ Assistive technological services.▪ Facilitation of test and examination accommodations.▪ Test and exam concession applications.▪ Accessible study venues and a computer lab.▪ Referrals for recommended textbooks in electronic format.	012 420 2064 email: du@up.ac.za www.up.ac.za/disability-unit	
Student Counselling Unit	Provides counselling and therapeutic support to students.	012 420 2333	
Student Health Services	Promotes and assists students with health and wellness.	012 420 5233 012 420 3423	
The Careers Office	Provides support for UP students and graduates as they prepare for their careers.	012 420 2315 careerservices@up.ac.za	
Department of Security Services	24-hour Operational Management Centre. 24-hour Operational Manager Crisis Line.	012 420-2310 012 420-2760 083 654 0476 0800 006 428	
Department of Student Affairs	Enquiries concerning studies, accommodation, food, funds, social activities and personal problems.	012 420 2371/4001 Roosmryn Building, Hatfield campus	
Centre for Sexualities, AIDS and Gender	Identifies and provides training of student peer counsellors.	012 420 4391	
Fees and funding	http://www.up.ac.za/enquiry www.up.ac.za/fees-and-funding	012 420 3111	

FLY@UP: The Finish Line is Yours	<ul style="list-style-type: none">▪ Think carefully before dropping modules (after the closing date for amendments or cancellation of modules).▪ Make responsible choices with your time and work consistently.▪ Aim for a good semester mark. Don't rely on the examination to pass.	email: fly@up.ac.za www.up.ac.za/fly@up	
-------------------------------------	---	---	---

IT Helpdesk	For IT-related student queries.	012 420 3051 studenthelp@up.ac.za	
-------------	---------------------------------	--	---



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za

What is Data Science?



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

**Faculty of Engineering,
Built Environment and
Information Technology**

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter
www.up.ac.za

Data Science

- Data, Data & More Data
 - +- 2.5 Exabytes (2.5 billion gigabytes) Per Day
 - 1 exabyte = 1,000 petabytes (PB)
 - 1 exabyte = 1,000,000 terabytes (TB)
 - 1 exabyte = 1,000,000,000 gigabytes (GB)
 - 1 exabyte = 1,000,000,000,000 megabytes (MB)
 - 1 exabyte = 1,000,000,000,000,000 kilobytes (KB)



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Data Science (continued...)



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za

Data Science (continued...)



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Data Science (continued...)

- Data Science Definitions:
 - Set of fundamental principles that guide the extraction of knowledge from data.
 - Field of study that combines domain expertise, programming skills, and knowledge of maths and statistics to extract meaningful insights from data



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology
Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Data Science (continued...)



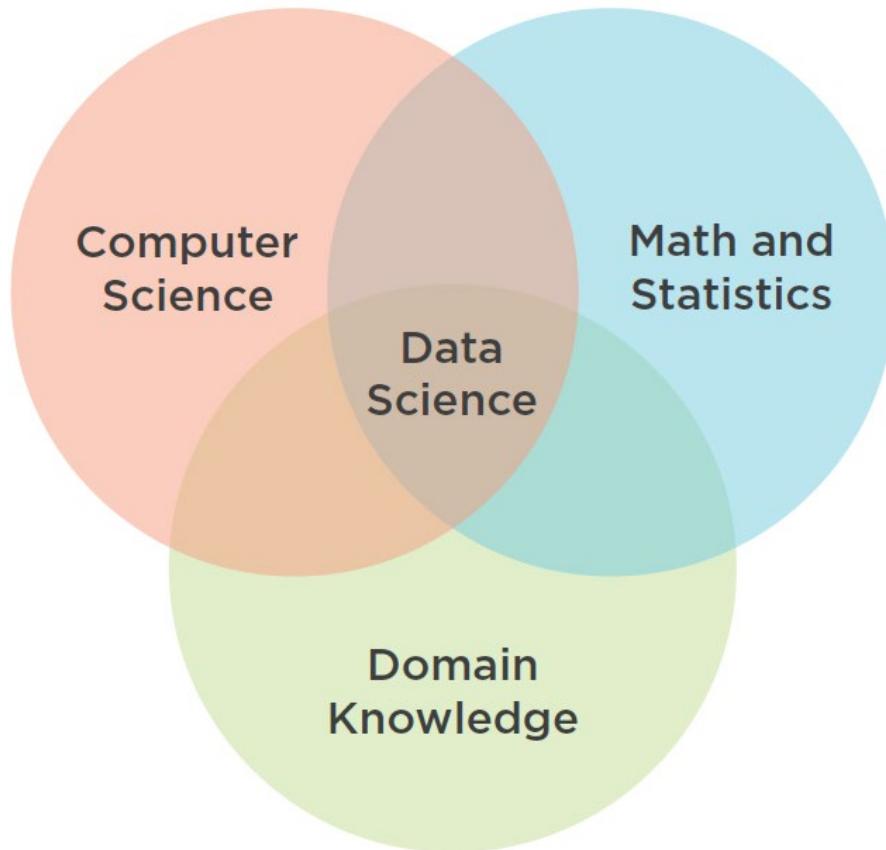
UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Data Science (continued...)

- What is Data Science:



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

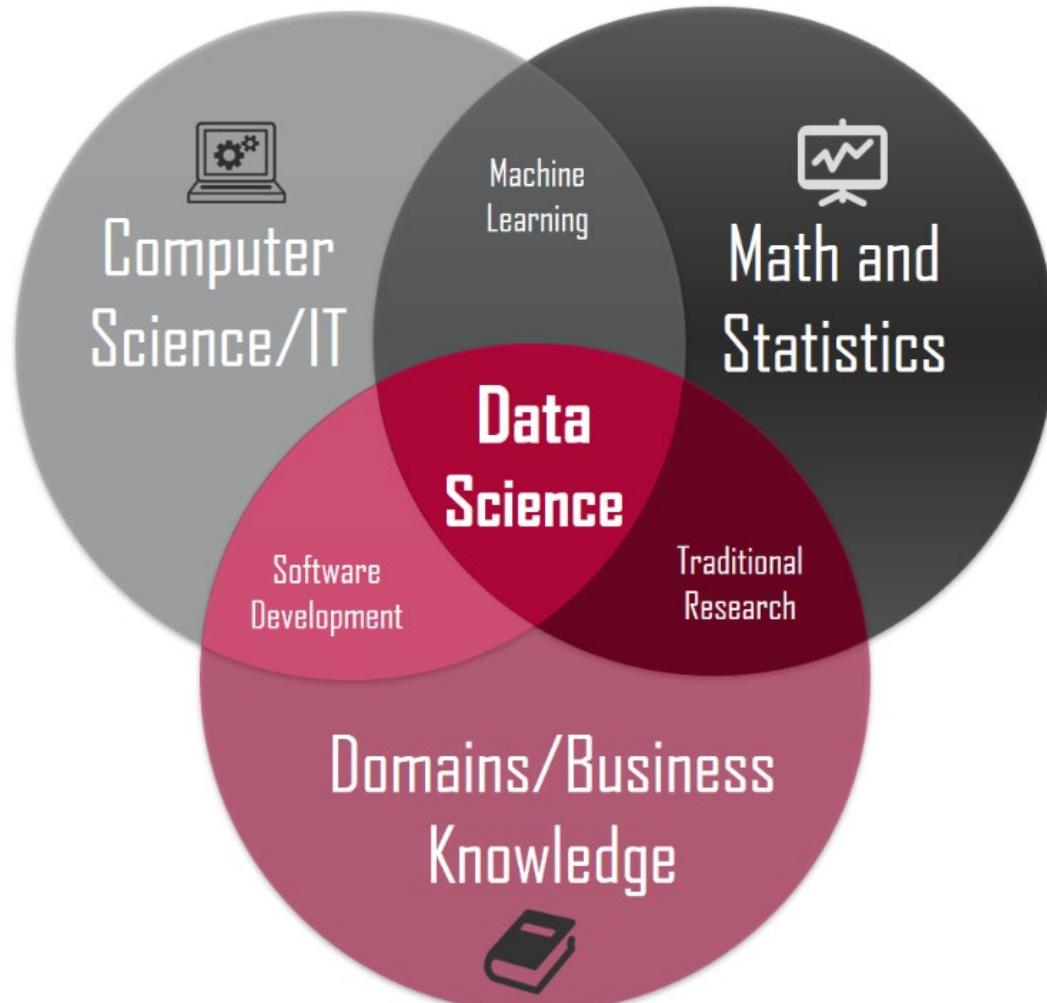
Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za

Data Science (continued...)

- What is Data Science:



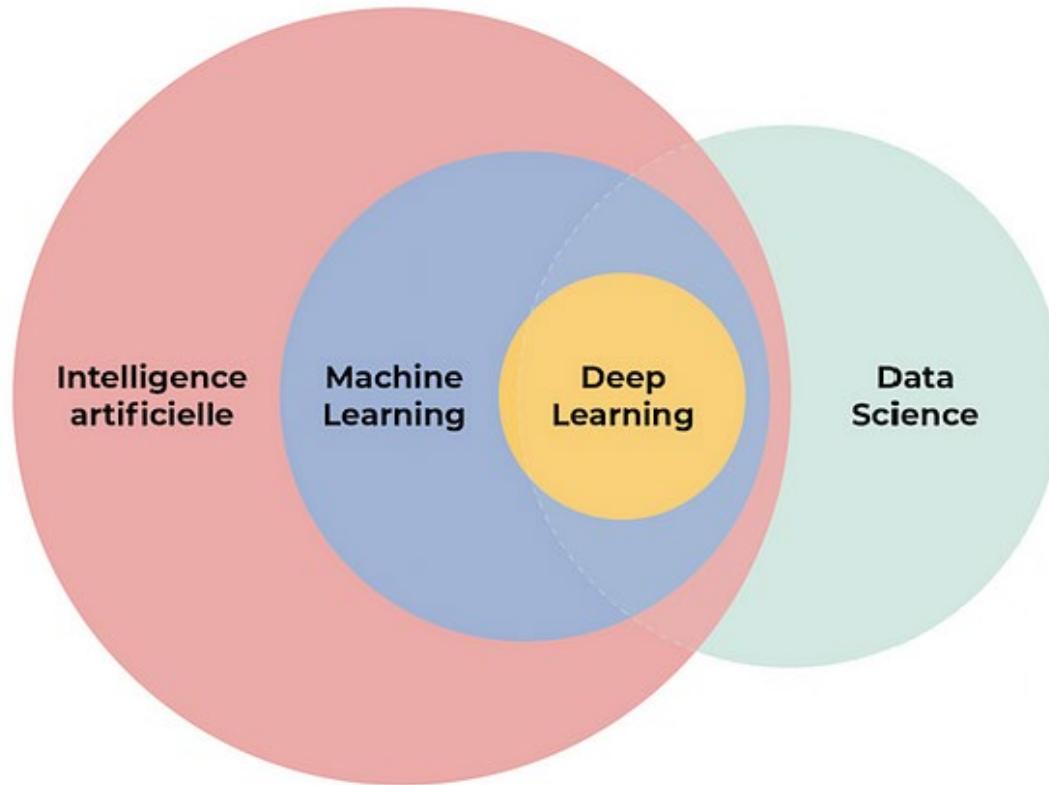
UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Data Science (continued...)

- Where Data Science is situated:



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

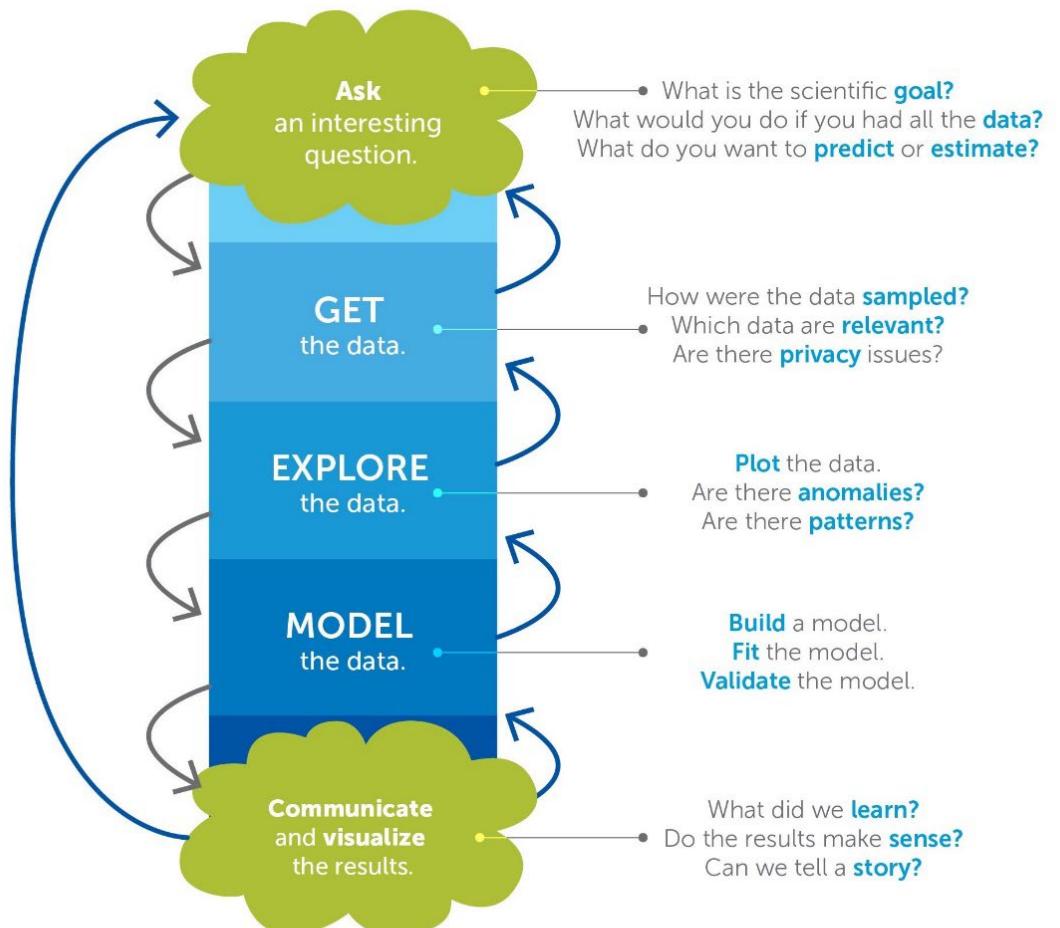
Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za

Data Science (continued...)

The Data Science Process



Derived from the work of Joe Blitzstein and Hanspeter Pfister,
originally created for the Harvard data science course <http://cs109.org/>.



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

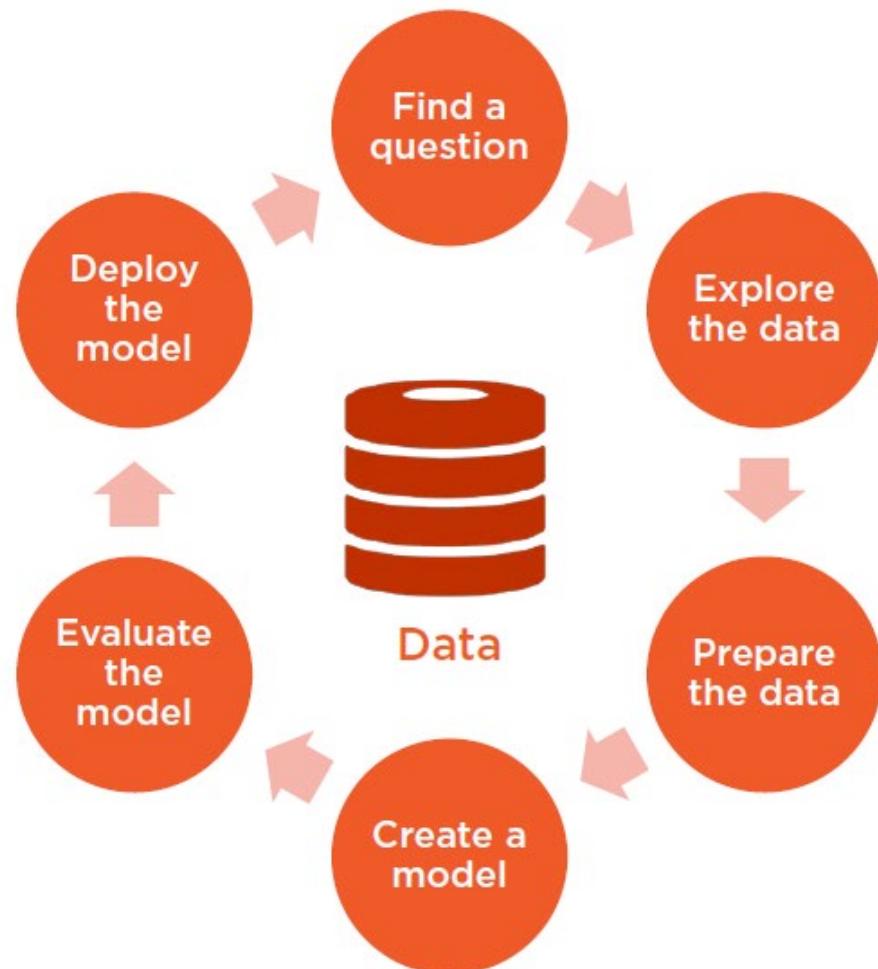
Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za

Data Science (continued...)

- The Data Science Process (another example):



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Data Science (continued...)

- What is a Data Scientist:
 - Is a Scientist
 - and
 - Is a Developer
 - and
 - Is an Analyst
- i.e. They perform data science
- Is Data Scientist still the sexiest job of the 21st century



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Data Science (continued...)

- Data Scientist Skills:
 - Programming
 - Working with data
 - Descriptive statistics
 - Data visualization
 - Statistical modeling
 - Handling Big Data
 - Machine learning
 - Deployment of solutions



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

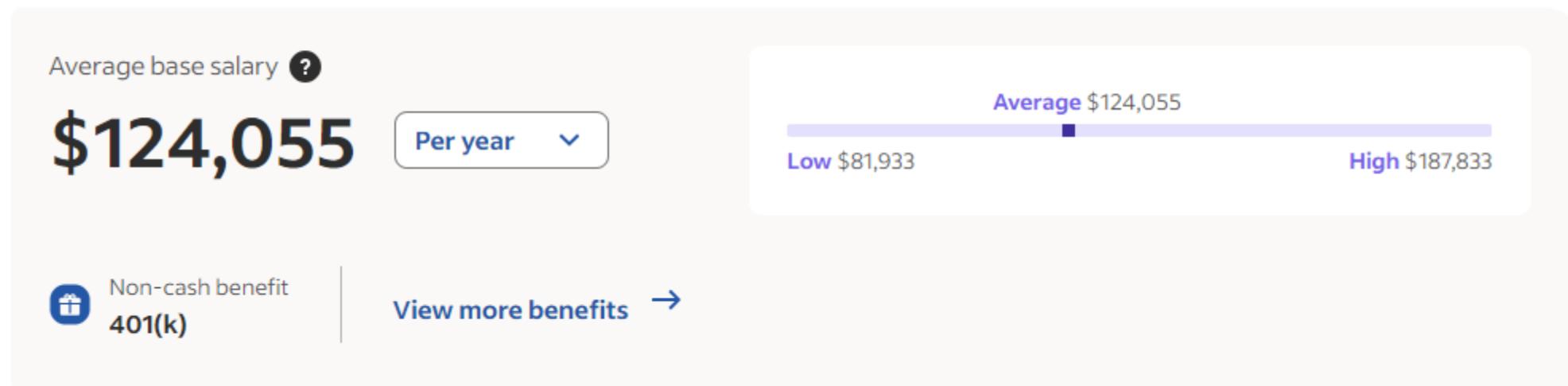
Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Data Science (continued...)

- Data Scientist Pay estimates:

Data scientist salary in United States

How much does a Data Scientist make in the United States?



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

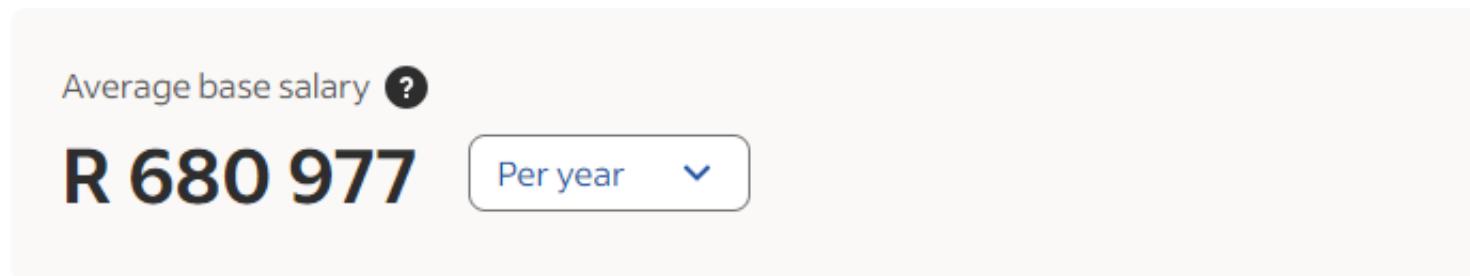
www.up.ac.za

Data Science (continued...)

- Data Scientist Pay estimates:

Data scientist salary in South Africa

How much does a Data Scientist make in South Africa?



The average salary for a data scientist is R 680 977 per year in South Africa. 110 salaries reported, updated at 25 July 2023



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

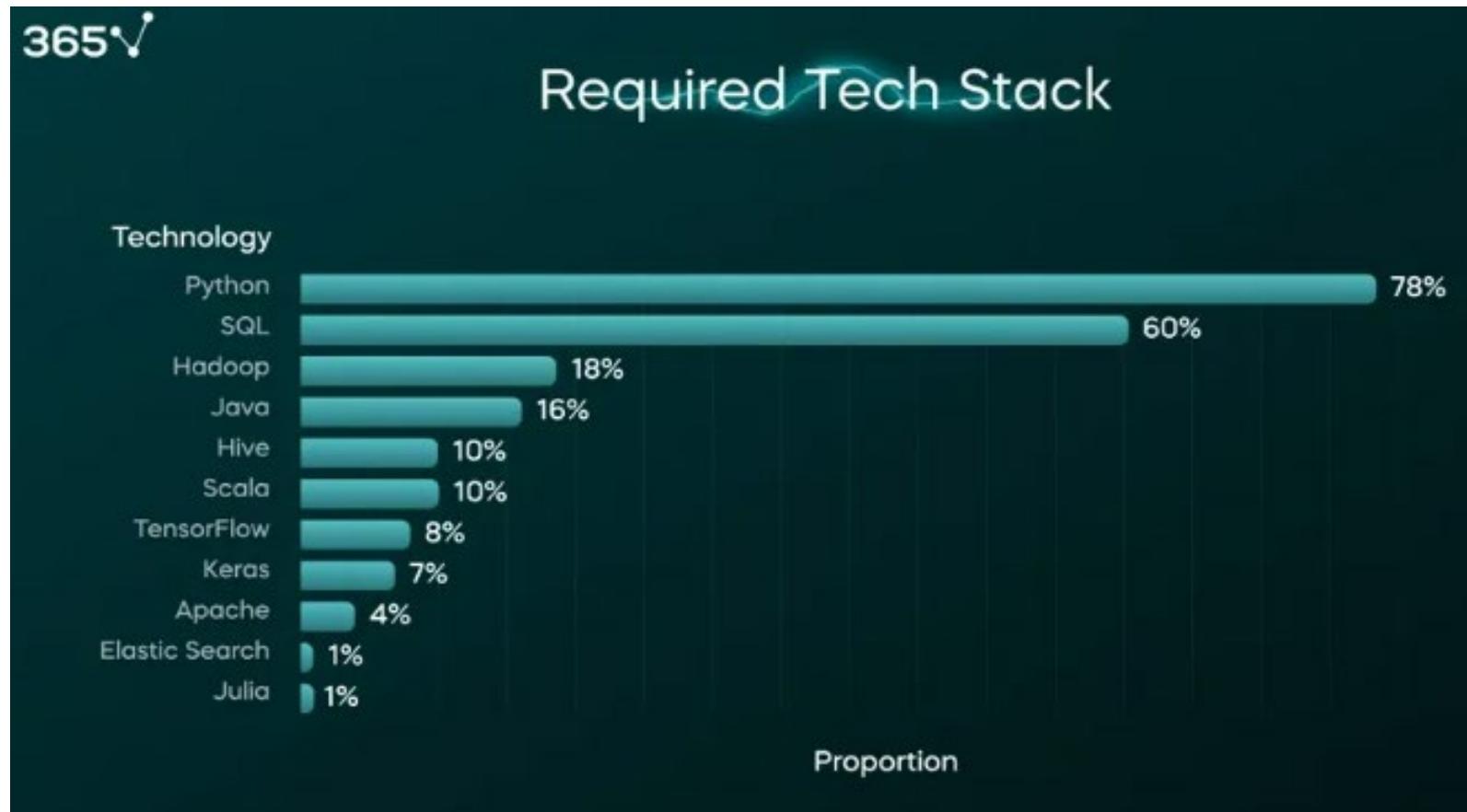
Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za

Data Science (continued...)

- Data Science Tools:



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Data Science (continued...)

■ Data Science Tools:

Worldwide, Mar 2023 compared to a year ago:

[2]

Rank	Change	Language	Share	Trend
1		Python	27.91 %	-0.6 %
2		Java	16.58 %	-1.6 %
3		JavaScript	9.67 %	+0.6 %
4		C/C++	6.93 %	-0.5 %
5		C#	6.88 %	-0.5 %
6		PHP	5.19 %	-0.6 %
7		R	4.23 %	-0.2 %
8	↑	TypeScript	2.81 %	+0.6 %
9	↑	Swift	2.28 %	+0.2 %
10	↓↓	Objective-C	2.26 %	+0.0 %



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za

Software Requirements



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

**Faculty of Engineering,
Built Environment and
Information Technology**

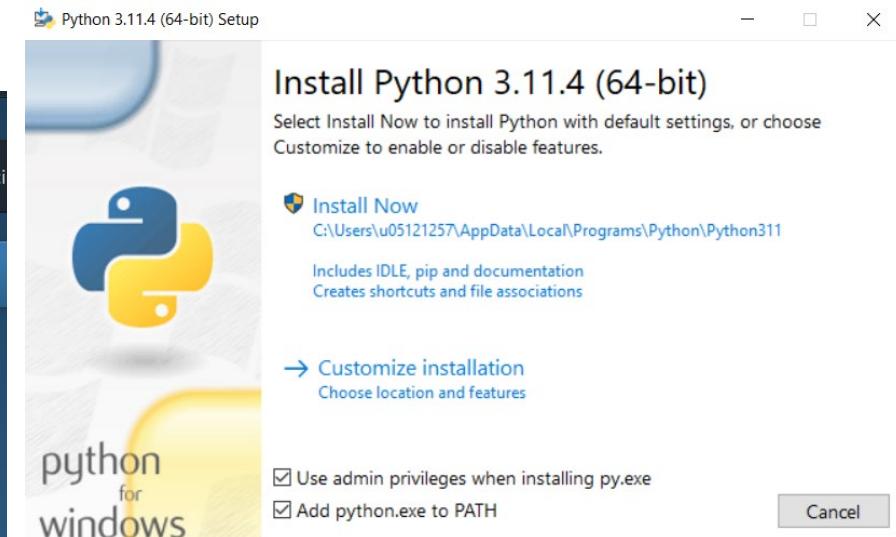
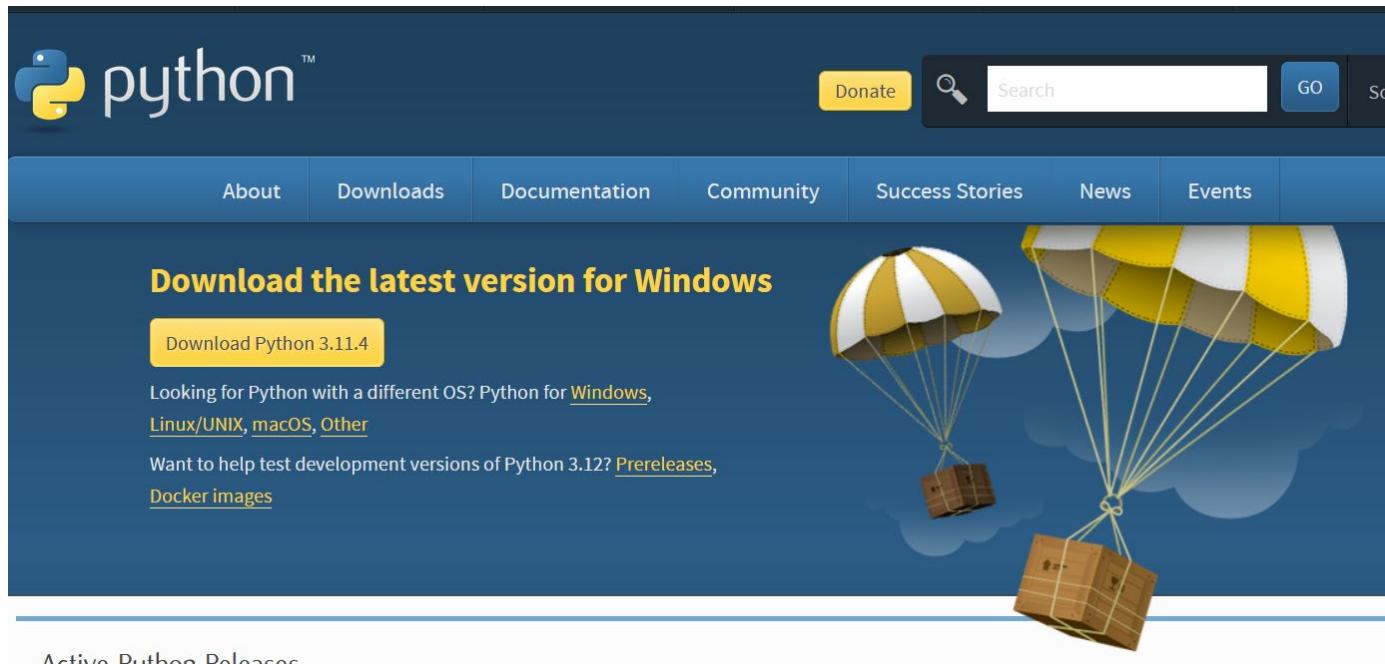
Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za

Software Requirements

- Installation 1: Python
 - <https://www.python.org/downloads/>



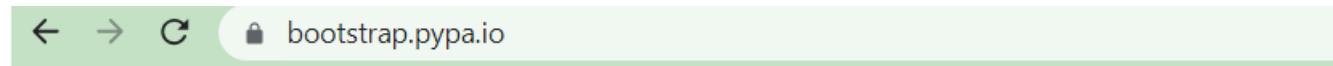
UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Software Requirements (continued ...)

- Installation 2: Pip (If not installed with Python)
 - <https://bootstrap.pypa.io/get-pip.py>



Index of /

..		
pip/	22-Jul-2023 09:45	-
virtualenv/	24-Jul-2023 15:15	-
bootstrap-buildout.py	21-Feb-2019 18:06	7458
ez_setup.py	21-Feb-2019 18:06	12537
get-pip.py	22-Jul-2023 09:45	2605506
virtualenv.pyz	24-Jul-2023 15:15	4739859

```
python get-pip.py
```



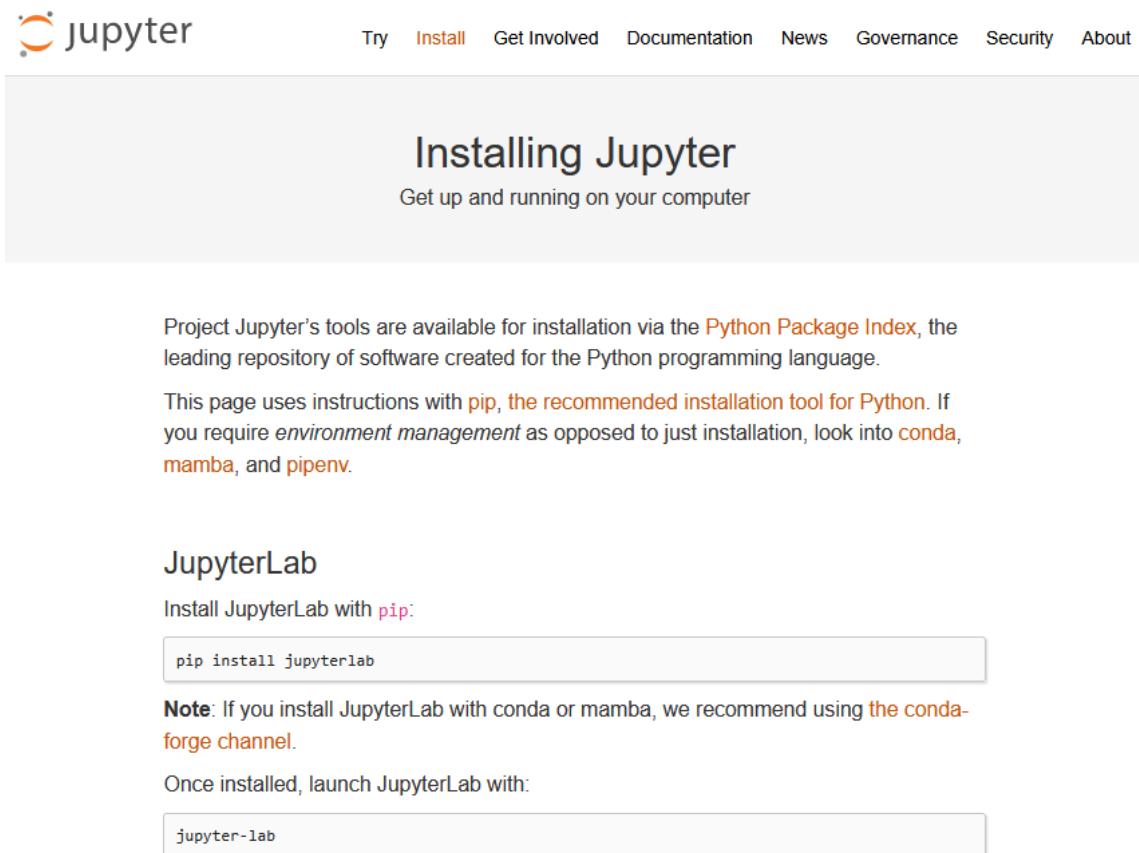
UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Software Requirements (continued ...)

- Installation 3: JupyterLab
 - <http://jupyter.org/install>



The screenshot shows the Jupyter website's "Installing Jupyter" page. The header includes links for Try, Install, Get Involved, Documentation, News, Governance, Security, and About. The main content area has a title "Installing Jupyter" and a subtitle "Get up and running on your computer". It explains that Project Jupyter's tools are available via the Python Package Index. It provides instructions for installing JupyterLab using pip, noting that conda, mamba, and pipenv are alternative environment management tools. Below this, there is a section for JupyterLab with instructions to install it using pip and launch it with jupyter-lab.

Project Jupyter's tools are available for installation via the [Python Package Index](#), the leading repository of software created for the Python programming language.

This page uses instructions with [pip](#), the recommended installation tool for Python. If you require *environment management* as opposed to just installation, look into [conda](#), [mamba](#), and [pipenv](#).

JupyterLab

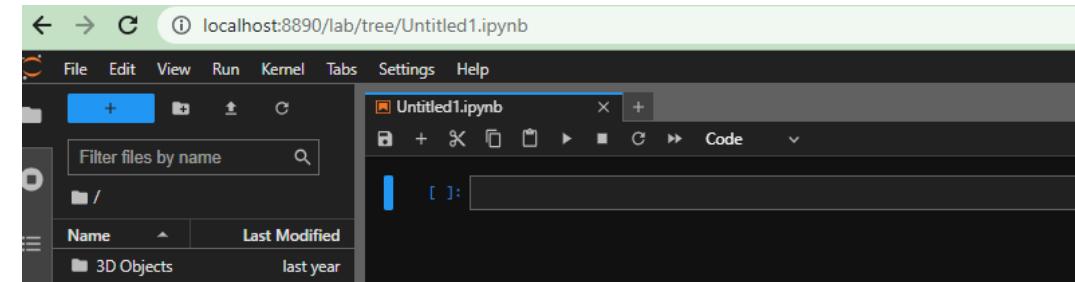
Install JupyterLab with [pip](#):

```
pip install jupyterlab
```

Note: If you install JupyterLab with conda or mamba, we recommend using [the conda-forge channel](#).

Once installed, launch JupyterLab with:

```
jupyter-lab
```



**Faculty of Engineering,
Built Environment and
Information Technology**
Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Software Requirements (continued ...)

- Note:
 - Other packages/libraries will be installed within the environments/ide's/software
 - We will inform you if you need to install any additional environment/ide/software during the semester



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology
Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Assignments



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

**Faculty of Engineering,
Built Environment and
Information Technology**

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za

2 Assignments

- Each assignment has about a month to complete
- Will use a rubric
- Information on Assignment 1 within the first half of August



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter

www.up.ac.za

Additions



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter
www.up.ac.za

- Textbook: “*Data Science from Scratch: First Principles with Python*” by Joel Grus
- Consultations from August



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering,
Built Environment and
Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Conclusion



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

**Faculty of Engineering,
Built Environment and
Information Technology**

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Make today matter
www.up.ac.za