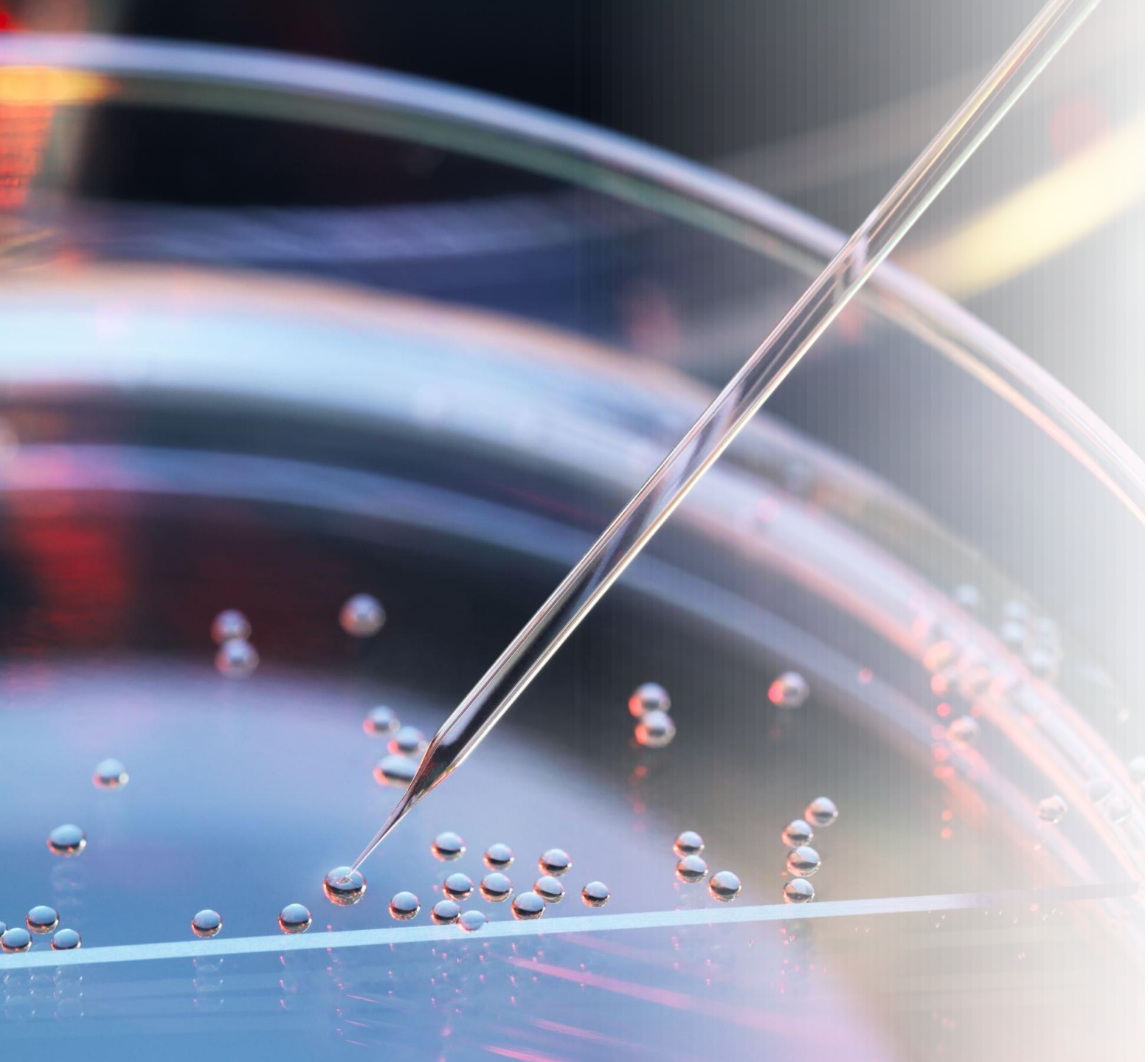


Analysis of Adverse Drug Effects Using Big Data and Cloud Computing

Leveraging data to enhance drug safety and protect patients

General description

- What are Adverse Drug Effects (ADEs)?
  - Undesirable and harmful reactions resulting from medications.
- Need for ADE Analysis:
  - Early detection
  - Traditional methods are time-consuming
- Role of Big Data and Cloud Computing:
  - Enables analysis of vast amounts of drug and patient data.
  - Scalable, fast, and cost-efficient solutions

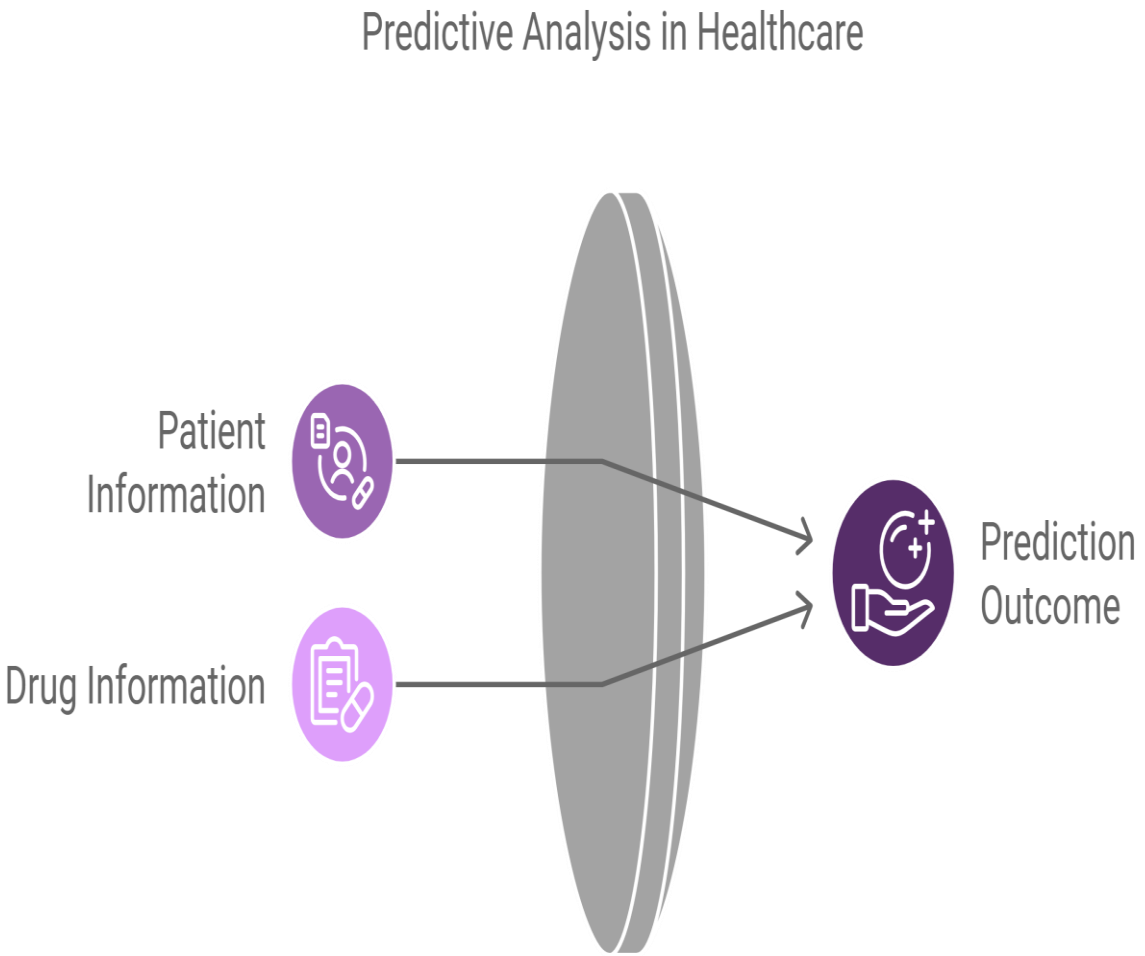
General description

- **Patient Information:**
  - Seriousness of the disease.
  - Hospitalization status (whether the patient is hospitalized or not).
  - Life-threatening condition status.
  - Patient age

- **Drug Information:**
  - Drug name
  - Number of active substances in the drug
  - Medicinal product details

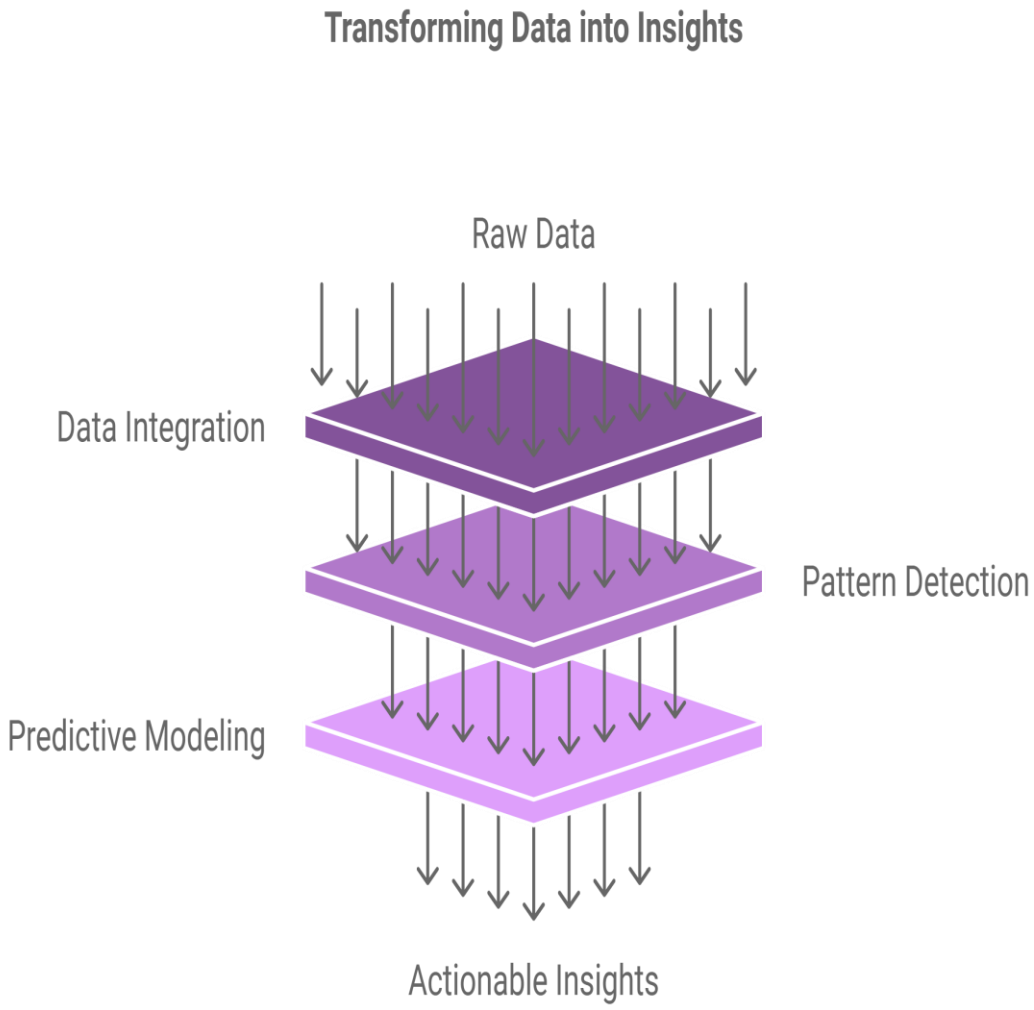
- **Prediction Outcome:**

  - Based on input data, the project will predict the reaction outcome.
  - If the outcome is serious, a recommendation can be made to consult the doctor for alternative drugs.



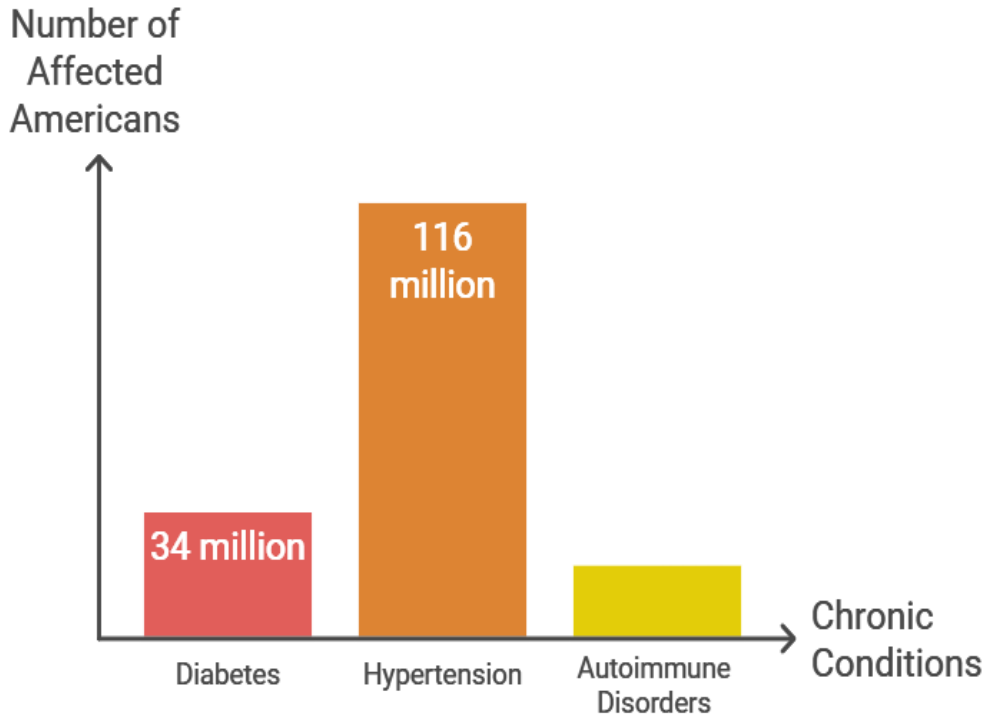
Objectives

- **Data Integration**: Integrate diverse data sources
- **Pattern Detection:** Identify correlations and trends
- **Develop Predictive Models:** Create and validate predictive models to forecast ADEs
  - Random Forest
  - Multilayer Perceptron (MLP)



Motivation

- **High Medication Usage**:
  - 70% of adults in the U.S. take one prescription medication.
  - 50% of adults take two or more medications.
- **Prevalence of Chronic Conditions**:
  - **Diabetes**: Over **34 million Americans** have diabetes, requiring ongoing medication management.
  - **Hypertension**: Nearly **116 million adults** in the U.S. suffer from high blood pressure.
  - **Autoimmune Disorders**: Approximately **20 million Americans** are affected by conditions like thyroid disease.



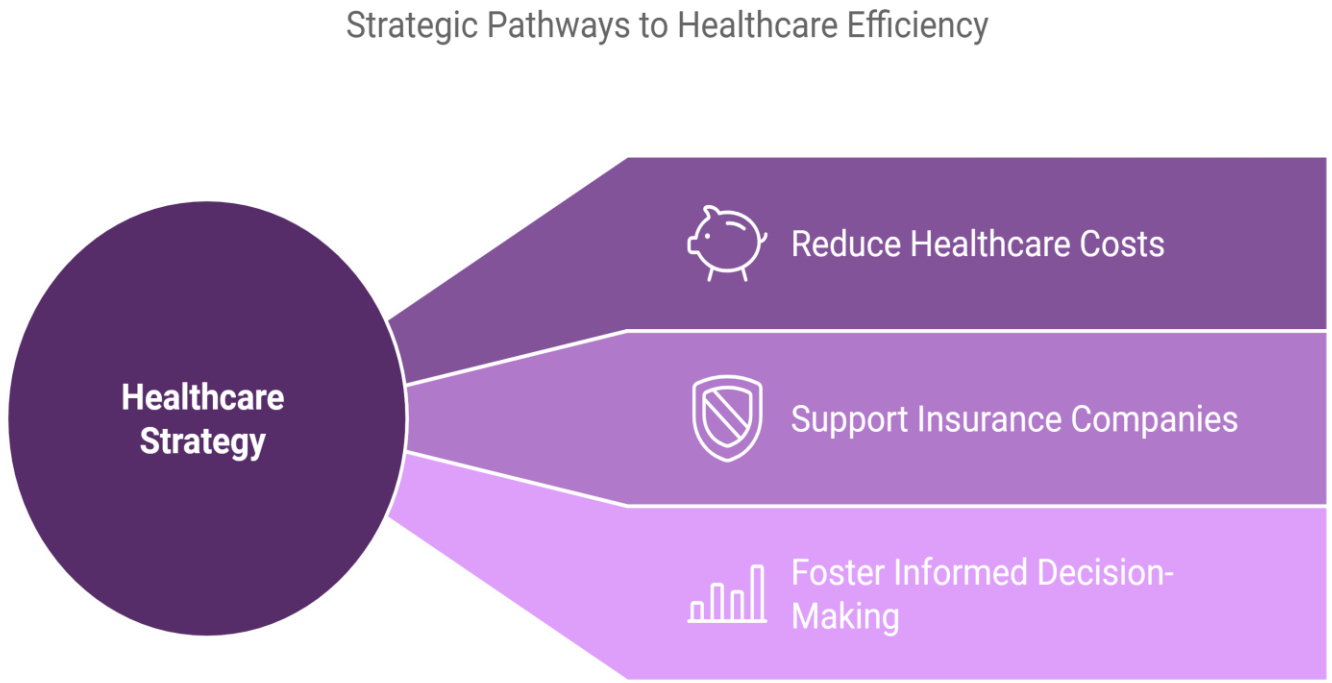
Prevalence of Chronic Conditions in the U.S.

Motivation

- **Growing Concern of Adverse Drug Effects (ADEs)**
- **Limitations of Traditional Methods**
- **Impact on Healthcare and Patient Safety**
- **Personal Interest in the Topic:**
  - Aiming to leverage advanced technologies to solve real-world healthcare challenges.
  - Contributing to a critical area where technology can directly impact human well-being.

Business Goal

- **Reduce Healthcare Costs**
- **Support Insurance Companies**: Help insurance companies reduce claims related to ADEs, ultimately lowering their overall expenditure on patient care.
- **Foster Informed Decision-Making**

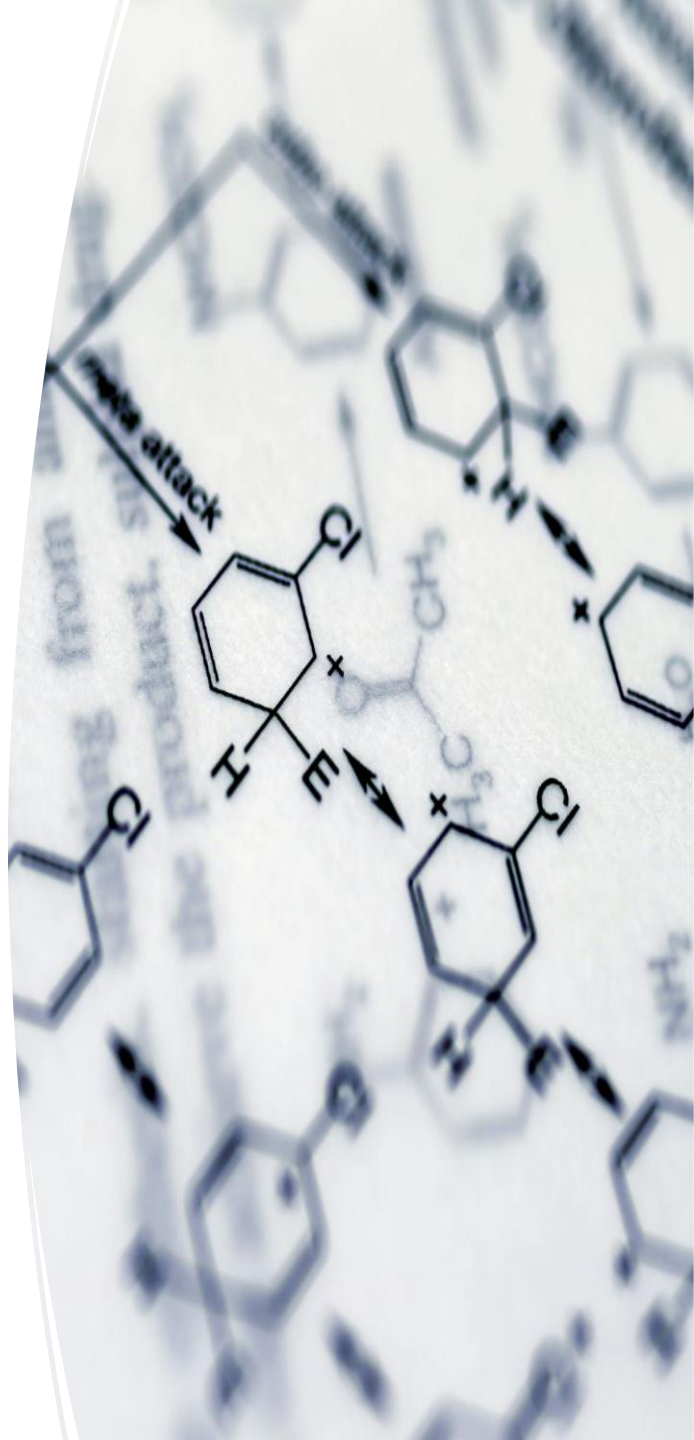


Dataset Description & Collection Process

Source of Data

FDA Adverse Event Reporting System (FAERS):

- The dataset contains **4 million rows**.
- Includes a wide range of drugs, patient demographics, and reaction descriptions.



Animal and Veterinary

Animal And Veterinary Event

Food

Food Enforcement

Food Event

Human Drug

Human Drug Event

Human Drug Label

Human NDC Directory

Human Drug Enforcement

Medical Device

Medical Device 510k

Medical Device Classification

Medical Device Enforcement

Medical Device Event

Medical Device PMA

Medical Device Recall

Human Drug

Drug Adverse Events [/drug/event]

This endpoint's data may be downloaded in zipped JSON files. Records are represented in the same format as API calls to this endpoint. update to the data in this endpoint could change old records. You need to download all the files to ensure you have a complete and up-t dataset, not just the newest files. For more information about openFDA downloads, see the API basics.

There are **1571** files, last updated on **2024-11-05**.

Hide all 1571 download files

2004

2004 Q3 (part 1 of 5) 60.66 mb
2004 Q3 (part 2 of 5) 3.14 mb
2004 Q3 (part 3 of 5) 45.27 mb
2004 Q3 (part 4 of 5) 101.22 mb
2004 Q3 (part 5 of 5) 56.76 mb
2004 Q2 (part 1 of 5) 58.13 mb
2004 Q2 (part 2 of 5) 3.34 mb
2004 Q2 (part 3 of 5) 57.29 mb
2004 Q2 (part 4 of 5) 104.52 mb
2004 Q2 (part 5 of 5) 35.04 mb
2004 Q1 (part 1 of 5) 58.31 mb
2004 Q1 (part 2 of 5) 2.92 mb

2005

2005 Q2 (part 1 of 5) 61.90 mb
2005 Q2 (part 2 of 5) 3.14 mb
2005 Q2 (part 3 of 5) 44.75 mb
2005 Q2 (part 4 of 5) 106.90 mb
2005 Q2 (part 5 of 5) 82.26 mb
2005 Q4 (part 1 of 5) 61.79 mb
2005 Q4 (part 2 of 5) 2.57 mb
2005 Q4 (part 3 of 5) 21.61 mb
2005 Q4 (part 4 of 5) 102.03 mb
2005 Q4 (part 5 of 5) 130.35 mb
2005 Q1 (part 1 of 6) 64.20 mb
2005 Q1 (part 2 of 6) 2.23 mb

2006

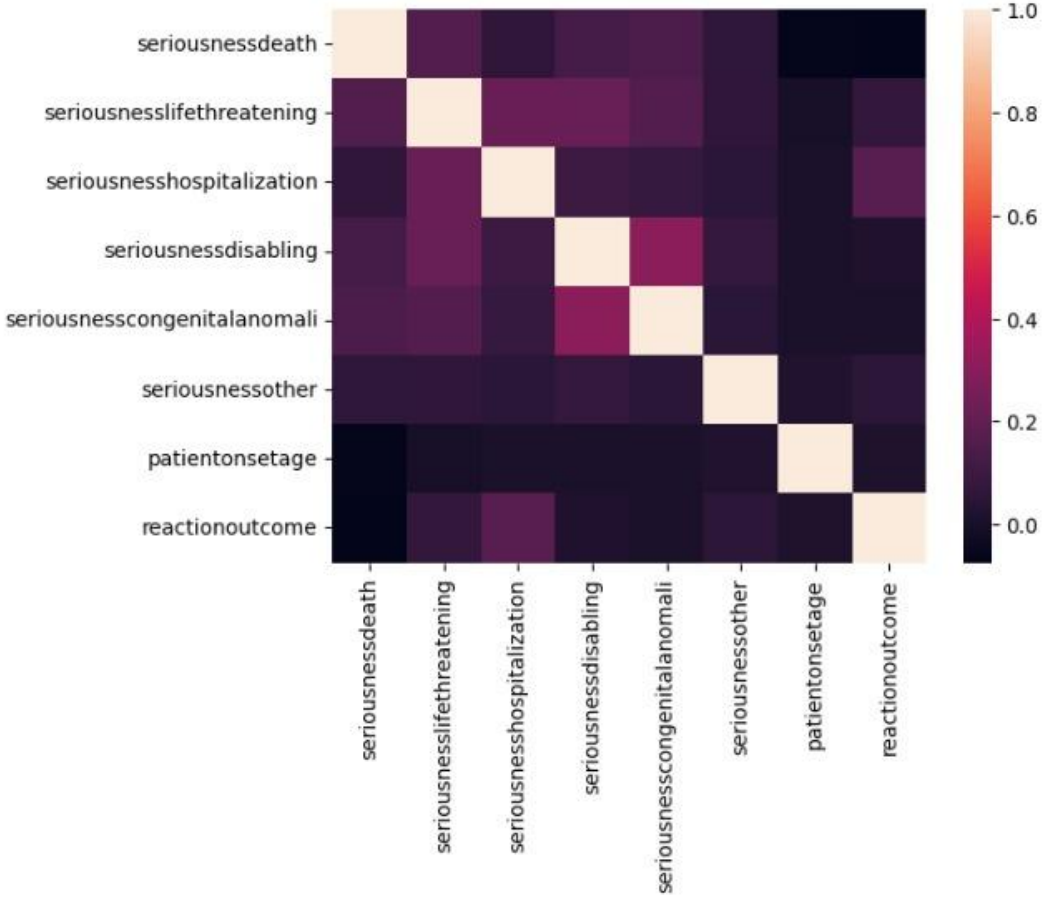
2006 Q1 (part 1 of 6) 62.02 mb
2006 Q1 (part 2 of 6) 2.12 mb
2006 Q1 (part 3 of 6) 7.81 mb
2006 Q1 (part 4 of 6) 93.33 mb
2006 Q1 (part 5 of 6) 105.32 mb
2006 Q1 (part 6 of 6) 29.37 mb
2006 Q2 (part 1 of 5) 54.71 mb
2006 Q2 (part 2 of 5) 3.01 mb
2006 Q2 (part 3 of 5) 18.22 mb
2006 Q2 (part 4 of 5) 88.88 mb
2006 Q2 (part 5 of 5) 93.62 mb
2006 Q4 (part 1 of 5) 60.47 mb

Key Columns and Their Definitions

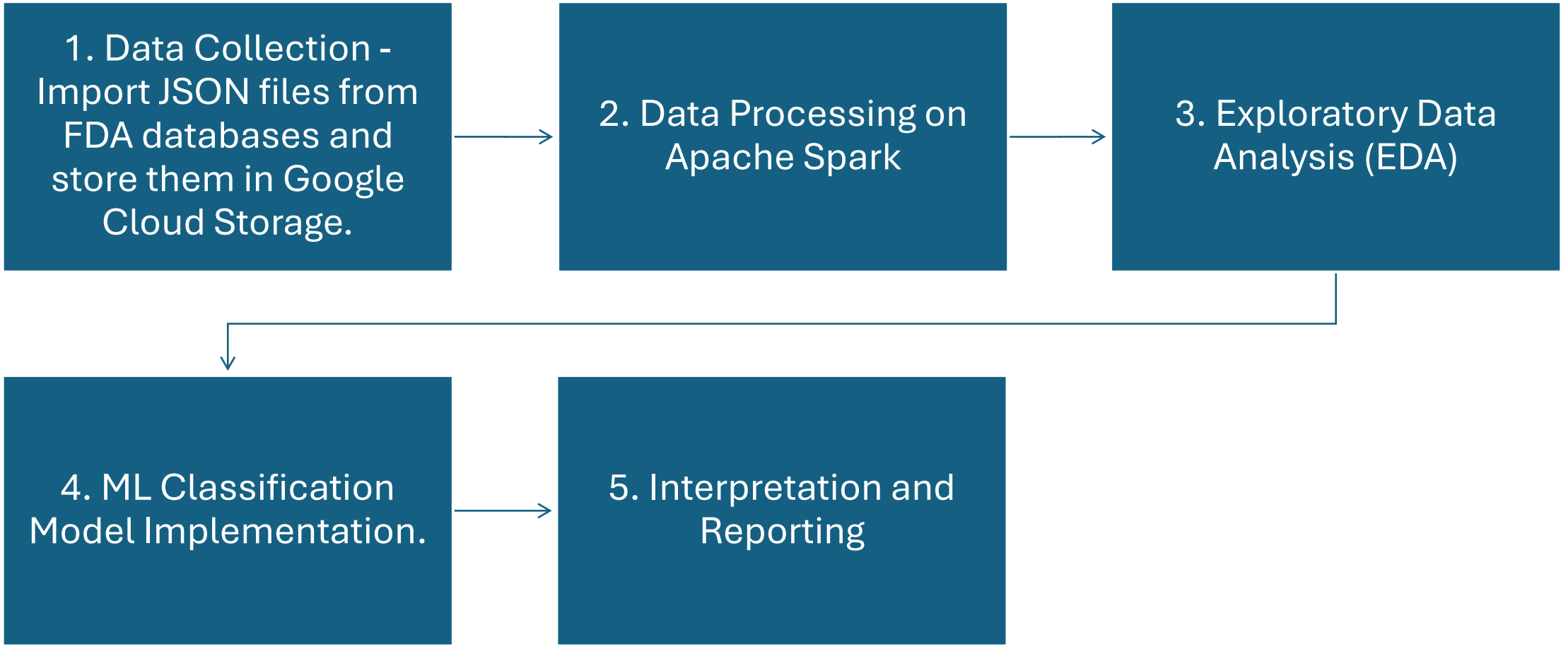
1. Safetyreportid - Unique identifier for each report.
2. Reactionmeddrapt - Type of adverse reaction (target column).
3. Reactionoutcome - Type of adverse reaction (target column)
4. Drugdosagetext - Information on the prescribed drug dosage.
5. Seriousnessdeath - Indicates if the reaction is classified as serious.

Data Pre-Processing

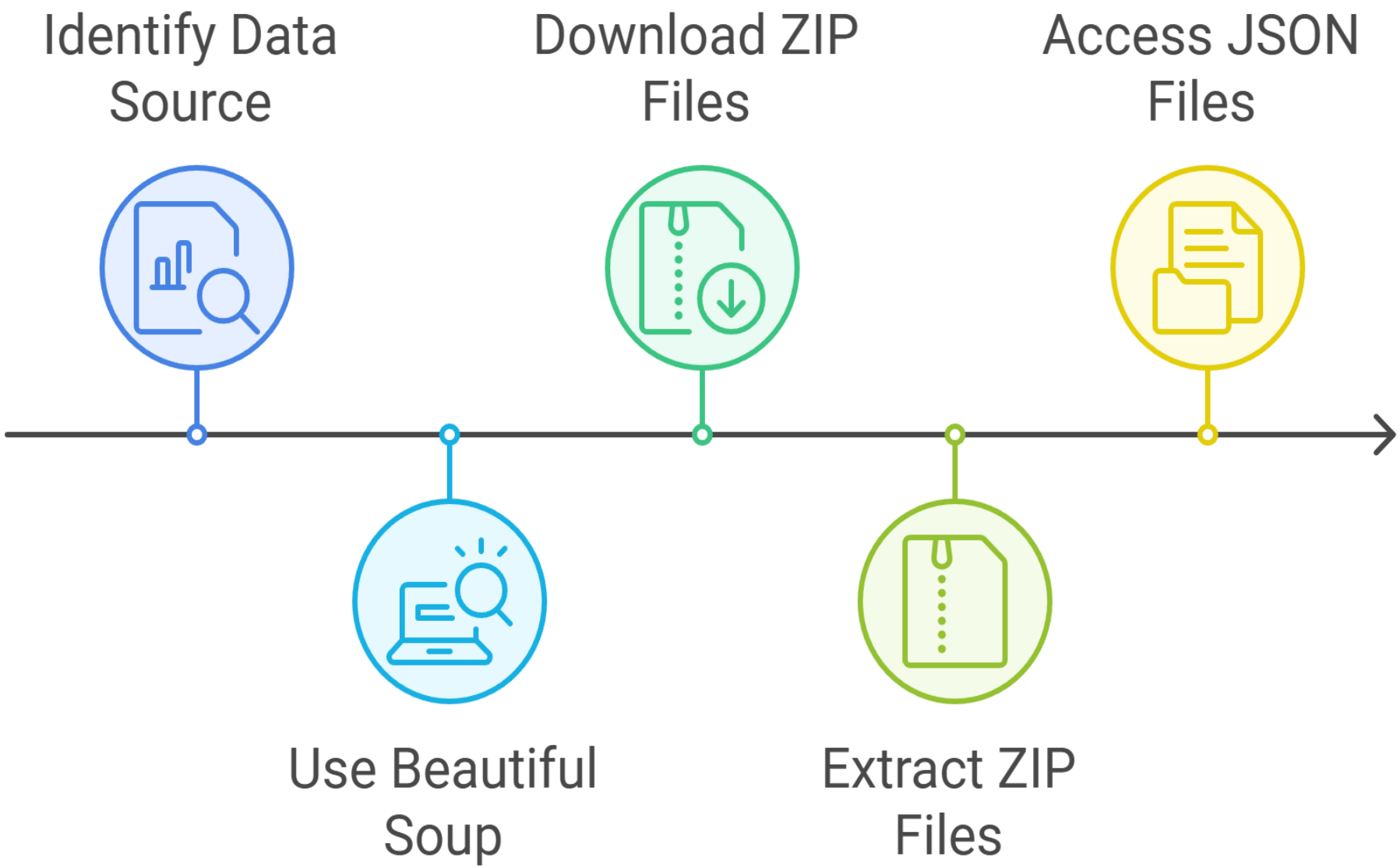
- **Drop Columns with More Than 30% Null Values:** improve dataset quality and analysis accuracy
- **Drug Treatment Duration:** Calculate the duration between 'drugstartdate' and 'drugenddate' to quantify the treatment period for ADE forecasting.
- **Feature Reduction:** Initially, 43 features were present. After removing the columns with more than 30% null values and feature engineering, we are left with 12 columns.



Data Pipeline and Workflow



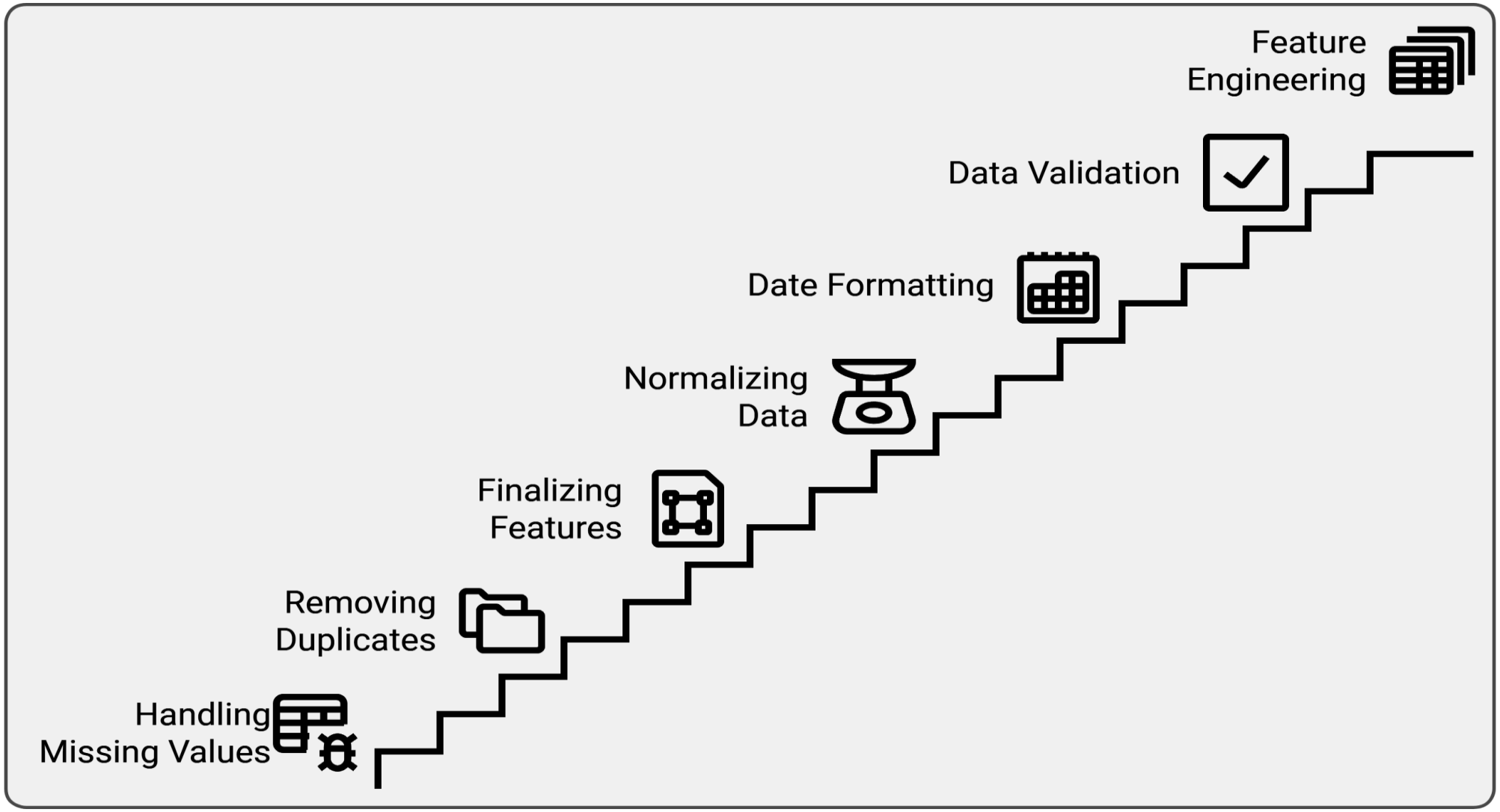
Data Collection



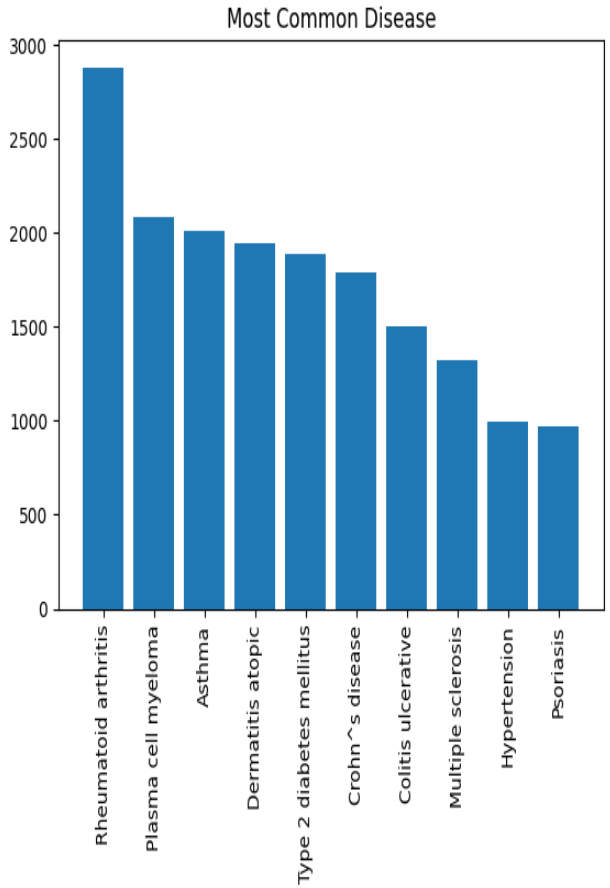
Data Schema

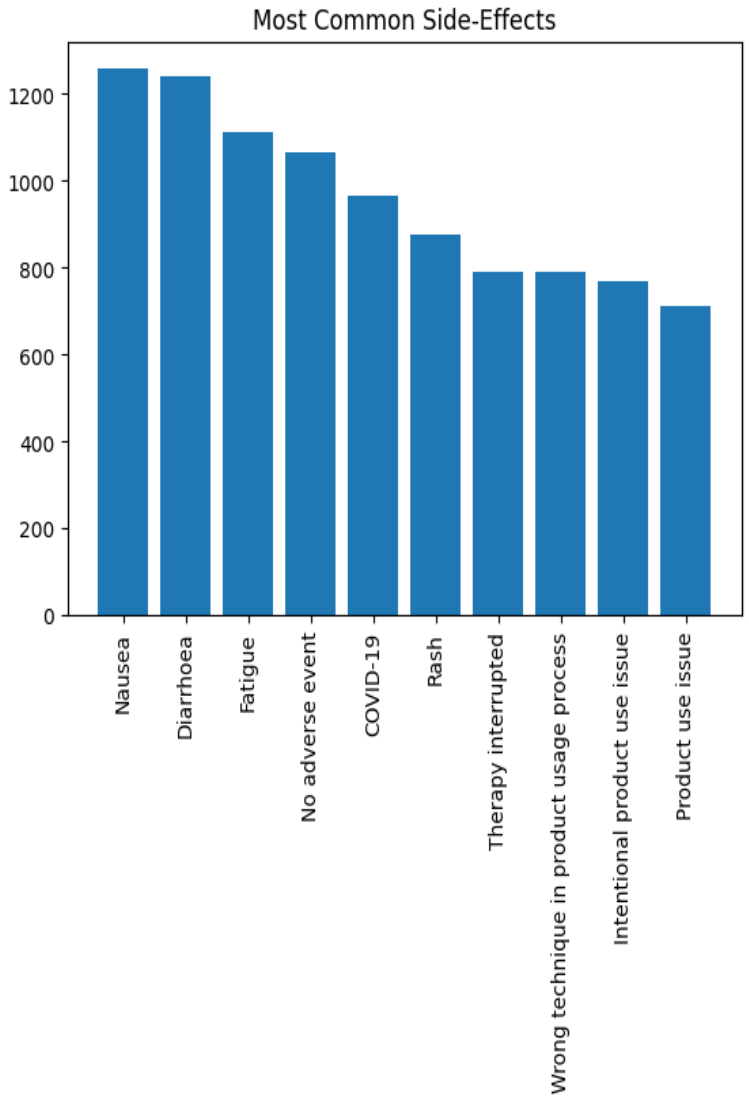
```
root
 |-- meta: struct (nullable = true)
 |    |-- disclaimer: string (nullable = true)
 |    |-- last_updated: string (nullable = true)
 |    |-- license: string (nullable = true)
 |    |-- results: struct (nullable = true)
 |    |    |-- limit: long (nullable = true)
 |    |    |-- skip: long (nullable = true)
 |    |    |-- total: long (nullable = true)
 |    |-- terms: string (nullable = true)
 |-- results: array (nullable = true)
 |    |-- element: struct (containsNull = true)
 |    |    |-- authoritynumb: string (nullable = true)
 |    |    |-- companynumb: string (nullable = true)
 |    |    |-- duplicate: string (nullable = true)
 |    |    |-- fulfillexpeditecriteria: string (nullable = true)
 |    |    |-- occurcountry: string (nullable = true)
 |    |    |-- patient: struct (nullable = true)
 |    |    |    |-- drug: array (nullable = true)
 |    |    |    |    |-- element: struct (containsNull = true)
 |    |    |    |    |    |-- actiondrug: string (nullable = true)
 |    |    |    |    |    |-- activesubstance: struct (nullable = true)
 |    |    |    |    |    |    |-- activesubstancename: string (nullable = true)
 |    |    |    |    |    |-- drugadditional: string (nullable = true)
 |    |    |    |    |    |-- drugadministrationroute: string (nullable = true)
 |    |    |    |    |    |-- drugauthorizationnumb: string (nullable = true)
 |    |    |    |    |    |-- drugbatchnumb: string (nullable = true)
 |    |    |    |    |    |-- drugcharacterization: string (nullable = true)
 |    |    |    |    |    |-- drugcumulativedosagenumb: string (nullable = true)
 |    |    |    |    |    |-- drugcumulativedosageunit: string (nullable = true)
 |    |    |    |    |    |-- drugdosageform: string (nullable = true)
 |    |    |    |    |    |-- drugdosagetext: string (nullable = true)
 |    |    |    |    |    |-- drugenddate: string (nullable = true)
 |    |    |    |    |    |-- drugenddateformat: string (nullable = true)
 |    |    |    |    |    |-- drugindication: string (nullable = true)
 |    |    |    |    |    |-- drugintervaldosagedefinition: string (nullable = true)
 |    |    |    |    |    |-- drugintervaldosageunitnumb: string (nullable = true)
 |    |    |    |    |    |-- drugrecurreadministration: string (nullable = true)
 |    |    |    |    |    |-- drugseparatedosagenumb: string (nullable = true)
 |    |    |    |    |    |-- drugstartdate: string (nullable = true)
 |    |    |    |    |    |-- drugstartdateformat: string (nullable = true)
 |    |    |    |    |    |-- drugstructuredosagenumb: string (nullable = true)
 |    |    |    |    |    |-- drugstructuredosageunit: string (nullable = true)
 |    |    |    |    |    |-- drugtreatmentduration: string (nullable = true)
 |    |    |    |    |    |-- drugtreatmentdurationunit: string (nullable = true)
 |    |    |    |    |    |-- medicinalproduct: string (nullable = true)
 |    |    |    |    |    |-- openfda: struct (nullable = true)
 |    |    |    |    |    |    |-- application_number: array (nullable = true)
 |    |    |    |    |    |    |    |-- element: string (containsNull = true)
```

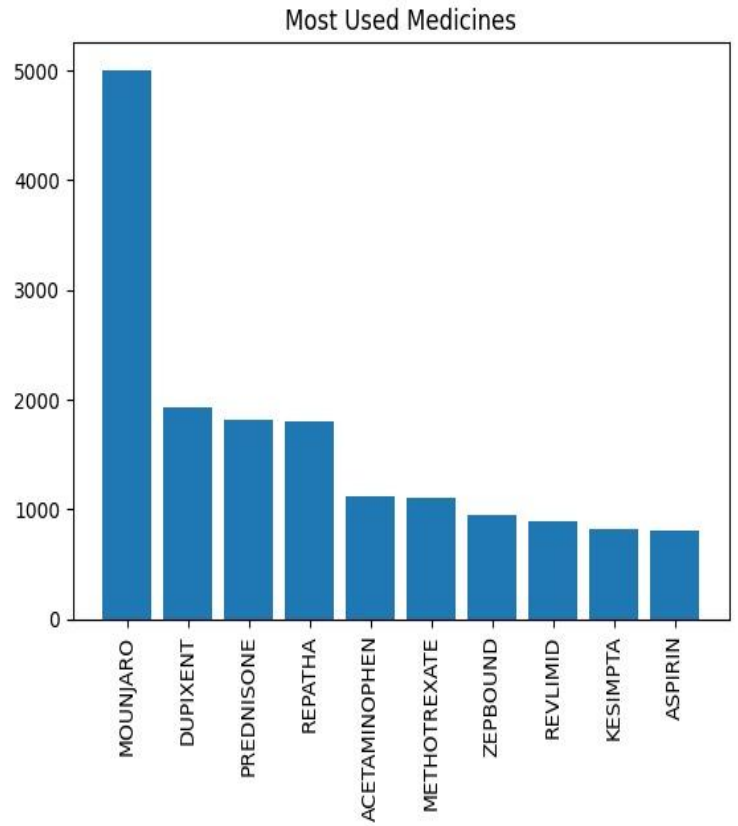
Data Cleaning & Preprocessing



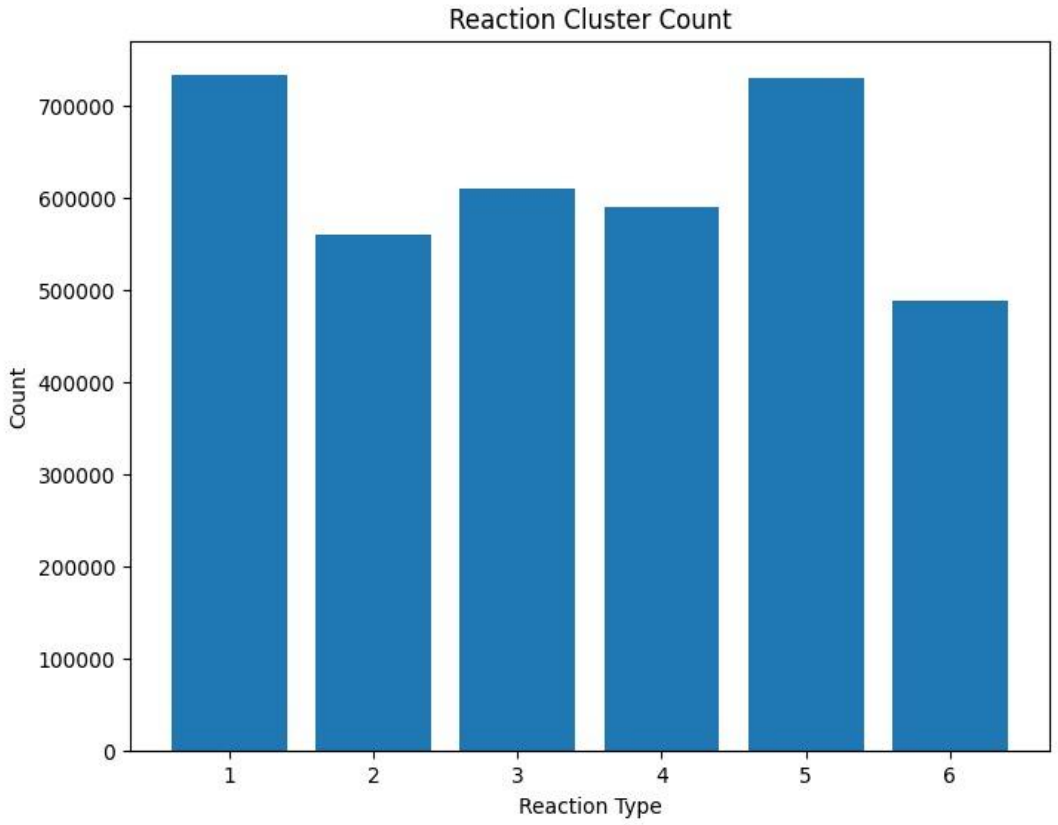
Data Visuals







Clustering



1. 'Abdominal abscess', 'Abdominal adhesions', 'Abdominal bruit', 'Abdominal compartment syndrome', 'Abdominal discomfort’

2. 'Adenocarcinoma', 'Adenocarcinoma of colon', 'Adenoma benign', 'Adenotonsillectomy', 'Adenoviral encephalitis', 'Adenovirus infection', 'Adhesion’.

3. 'Adenocarcinoma of colon', 'Adenocarcinoma pancreas', 'Adenoma benign', 'Adenomyosis', 'Adenovirus infection', 'Adhesion', 'Adjusted calcium increased', 'Adjustment disorder', 'Adjustment disorder with depressed mood'

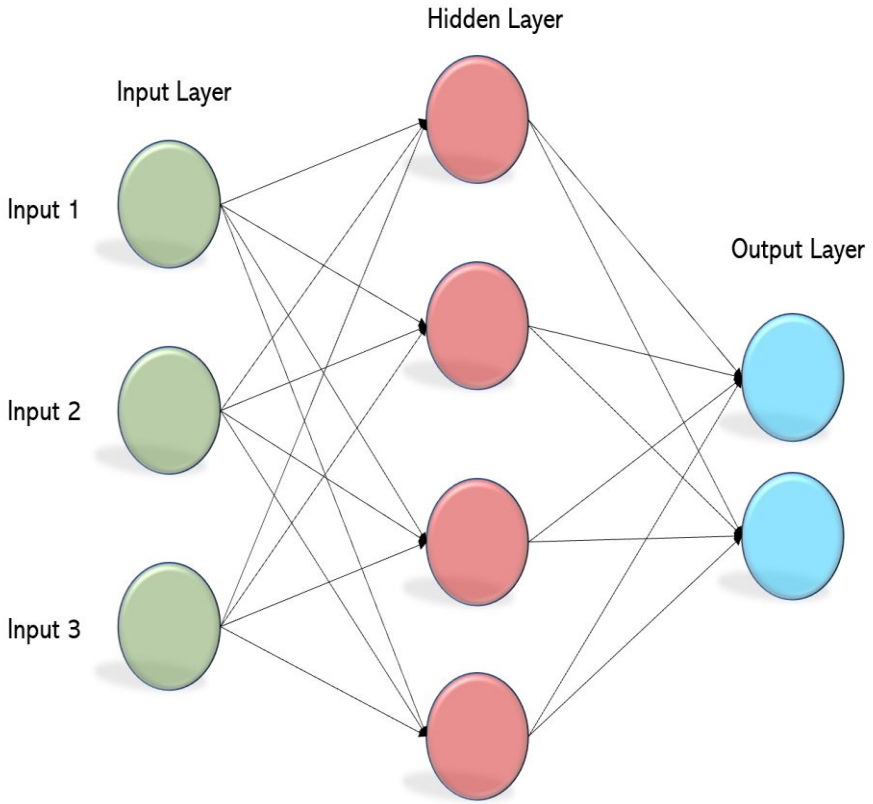
4. 'Vomiting', 'Vulval cancer metastatic', 'Vulvovaginal dryness', 'Walled-off pancreatic necrosis', 'Weight decreased', 'Weight increased’,

5. Wound dehiscence', 'Wound infection', 'Wound infection staphylococcal', 'Wound sepsis', 'Wrong patient received product', 'Wrong product administered

6. 'Xanthelasma', 'Xeroderma', 'Xerophthalmia', 'Xerosis', 'Yawning', 'Yellow skin', 'Yersinia infection

Develop Predictive Models

MLP



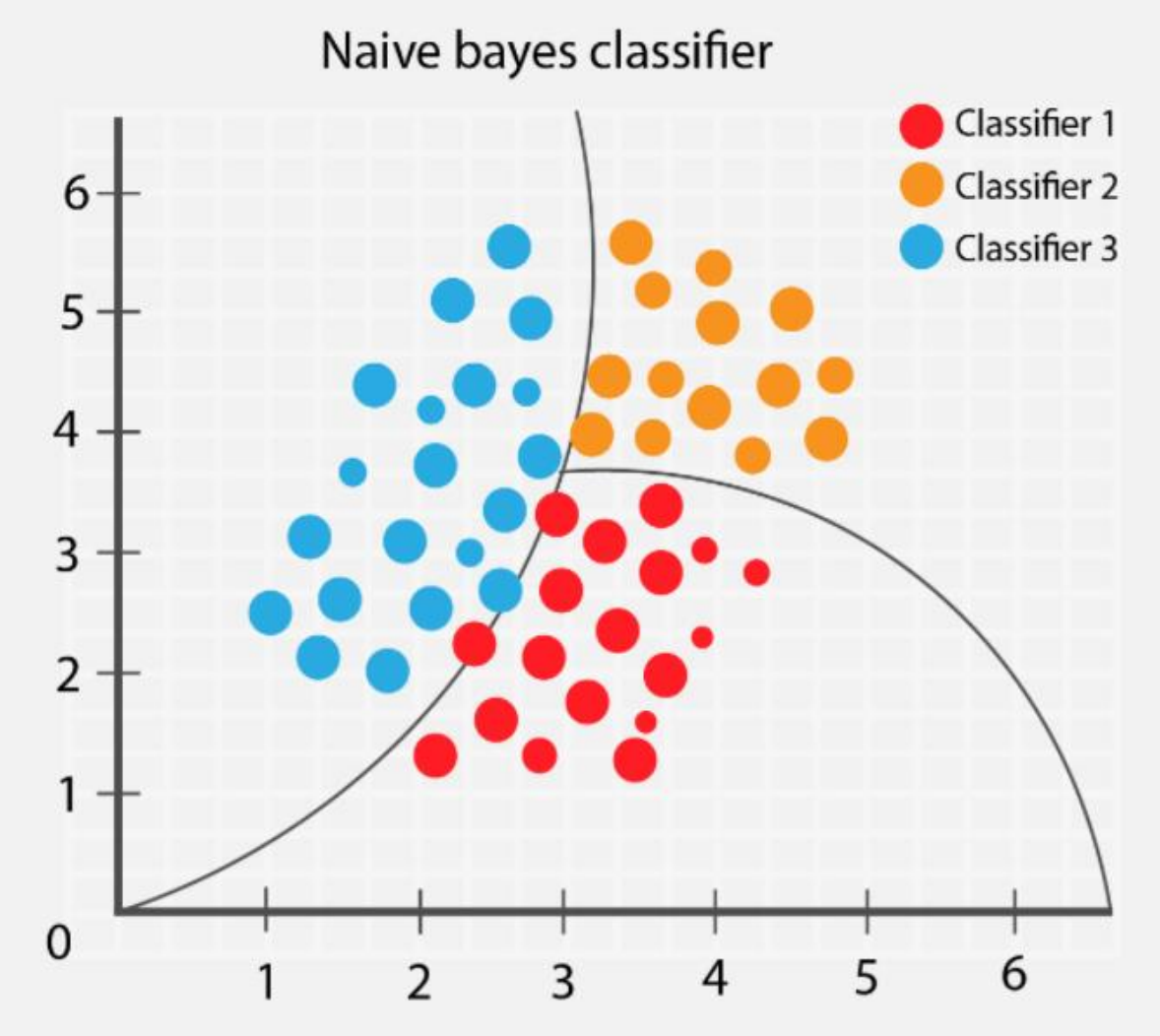
[64]:
```
predictions.select("features", "prediction","label").show(5)
```

```
+--------------------+----------+-----+
|            features|prediction|label|
+--------------------+----------+-----+
|[0.0,0.0,7.0,0.0,...|       0.0|  0.0|
|[0.0,0.0,7.0,0.0,...|       0.0|  0.0|
|[0.0,0.0,7.0,0.0,...|       0.0|  0.0|
|[0.0,0.0,2.0,0.0,...|       0.0|  0.0|
|[0.0,0.0,2.0,0.0,...|       0.0|  0.0|
+--------------------+----------+-----+
only showing top 5 rows
```

[65]:
```
# Select (prediction, true label) and compute test error
evaluator = MulticlassClassificationEvaluator(
    labelCol="label", predictionCol="prediction", metricName="accuracy")
accuracy = evaluator.evaluate(predictions)
print("Test Error = %g" % (1.0 - accuracy))
```

```
Test Error = 0.112965
```

Develop Predictive Models



[56]:
```
# Select example rows to display.
predictions.select("predictedLabel", "reactionoutcome", "features").show(5)
```

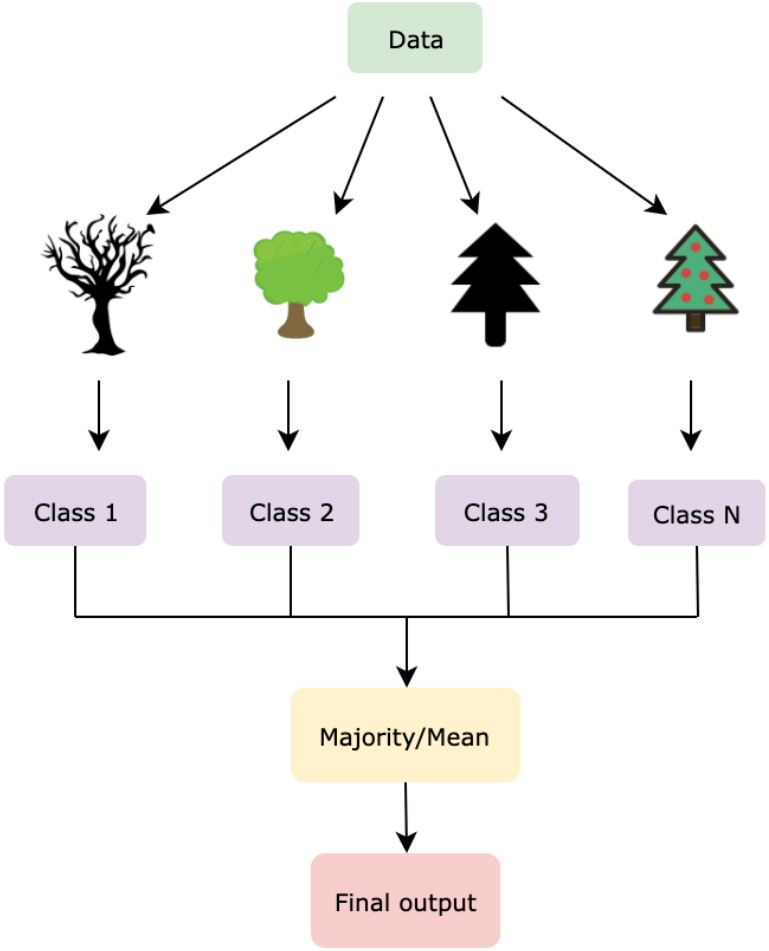
```
+--------------+--------------+--------------------+
|predictedLabel|reactionoutcome|            features|
+--------------+--------------+--------------------+
|             2|             5|[0.0,0.0,7.0,0.0,...|
|             5|             5|[0.0,0.0,2.0,0.0,...|
|             5|             5|[0.0,0.0,2.0,0.0,...|
|             5|             5|[0.0,0.0,2.0,0.0,...|
|             5|             5|[0.0,0.0,2.0,0.0,...|
+--------------+--------------+--------------------+
only showing top 5 rows
```

[57]:
```
# Select (prediction, true label) and compute test error
evaluator = MulticlassClassificationEvaluator(
    labelCol="label", predictionCol="prediction", metricName="accuracy")
accuracy = evaluator.evaluate(predictions)
print("Test Error = %g" % (1.0 - accuracy))
```

```
Test Error = 0.2
```

Develop Predictive Models

Random Forest



[52]:
```
# Select example rows to display.
predictions.select("prediction", "reactionoutcome", "features").show(5)
```

```
+----------+--------------+--------------------+
|prediction|reactionoutcome|            features|
+----------+--------------+--------------------+
|       5.0|             5|[0.0,0.0,7.0,1.0,...|
|       5.0|             5|[0.0,0.0,2.0,1.0,...|
|       5.0|             5|[0.0,0.0,2.0,1.0,...|
|       5.0|             5|[0.0,0.0,2.0,1.0,...|
|       5.0|             5|[0.0,0.0,2.0,1.0,...|
+----------+--------------+--------------------+
only showing top 5 rows
```

[53]:
```
# Select (prediction, true label) and compute test error
evaluator = MulticlassClassificationEvaluator(
    labelCol="reactionoutcome", predictionCol="prediction", metricName="accuracy")
accuracy = evaluator.evaluate(predictions)
print("Test Error = %g" % (1.0 - accuracy))
```

```
Test Error = 0.071775
```

Model Performance and Metrics

Naive Bayes Classifier
Test Error **: 0.20**
Accuracy **: 80%**

Observations: Performs adequately on basic patterns but struggles due to feature independence assumptions.

Random Forest
Test Error: **0.07177**
Accuracy: **92.82%**
Observations: Best-performing model with strong accuracy. Captures non-linear relationships effectively and minimizes classification errors.

Multilayer Perceptron (MLP)
Test Error: **0.112965**
Accuracy: **88.7%**
Observations: Performs better than Naive Bayes but slightly lower than Random Forest. Demonstrates strong predictive power but requires further tuning.

Scaling up - Data Utilization Impact on Processing Time

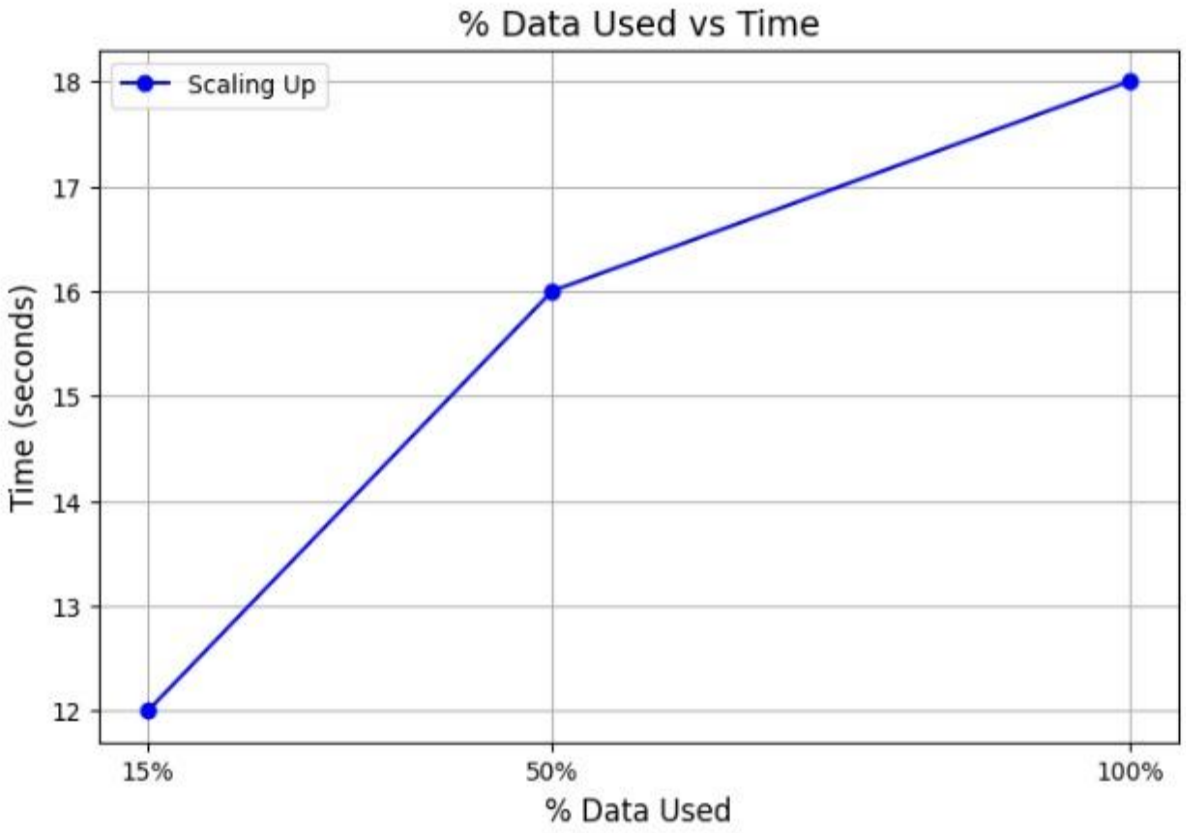
This graph illustrates the relationship between the percentage of data utilized and the time taken for processing.

Observation: The percentage of data used increases from 15% to 100%, the processing time also increases, demonstrating a direct correlation between data volume and computational time.

Time Taken:

- At 15% data usage: ~12 seconds
- At 50% data usage: ~14 seconds
- At 100% data usage: ~18 seconds

Conclusion: Efficient data management strategies are essential for handling large healthcare datasets, ensuring timely insights while minimizing computational delays.



Scaling up - Number of Workers vs Time

This graph illustrates the relationship between the number of worker nodes and the time taken for processing.

Observation: Increasing the number of workers from 3 to 7 leads to a decrease in processing time, demonstrating the benefits of parallel processing in data analysis workflows.

Time Taken:

- At 3 workers : ~37 seconds
- At 5 workers: ~25 seconds
- At 7 workers: ~20 seconds

Conclusion: Strategic allocation of computational resources enhances performance, making it critical for large datasets like drug adverse affects dataset.

Conclusion and Key Takeaways

Key Model Insights:

- Random Forest outperformed other models with an accuracy of **92.82%**, making it the most reliable for predicting adverse drug reactions.
- MLP and Naive Bayes showed moderate performance with accuracies of **88.7%** and **80%**, respectively.

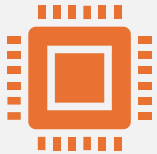
Visual Insights:

- **Most Common Diseases**: Rheumatoid arthritis and plasma cell myeloma are the leading conditions.
- **Common Side Effects**: Nausea and diarrhea are the most reported adverse effects.
- **Most Used Medicines**: Mounjaro has the highest usage, followed by Dupixent and Prednisone.

Clustering Insights:

- Reactions are grouped into distinct clusters, with cluster 6 having the **largest number of occurrences**.
- Specific issues like **abdominal discomfort, weight changes, and infections** are prevalent in clusters.

Future Enhancements and Scalability



1. Real-Time ADR Analysis: Integrate real-time data feeds for dynamic monitoring of drug safety profiles.

2. Advanced Machine Learning Models and Hyper parameter Tuning: Explore classification models to predict high-risk cases.

3. Broader Dataset Integration: Incorporate additional data sources like lifestyle or co-morbidity information for a holistic view of ADRs.

Thank you