IBM Data Science

SHR1_AIS3_M1e

DEPI Graduation Project Final Report

HealthCare Predictive Analytics

Team 2 Group Members

Ahmed Magdy Ahmed (Leader)

Khaled Tarek Mohamed

Modaher Abdelmohsen Abdelmawgood

Executive Summary

This report explores the potential of health predictive analytics in improving patient outcomes and operational efficiencies within healthcare systems. By leveraging historical data, machine learning, and statistical methods, predictive analytics can forecast health events, enhance decision-making, and optimize resource allocation. The report examines key methodologies, applications, challenges, and future directions in the field.

Introduction

Health predictive analytics involves analyzing vast amounts of health data to identify patterns, predict future health outcomes, and inform clinical decisions. With the rapid advancement of technology and increasing availability of health data, predictive analytics has become a vital component in modern healthcare strategies.

Tasks Assigned by Each Group Member

- **Ahmed Magdy Ahmed (Leader)**
  - Data Collection
  - ML-Flow
- **Khaled Tarek Mohamed**
  - Data Exploration
  - Data Analysis
  - Data Visualization
- **Modather Abdelmohsen Abdelmawgood**
  - Data Prediction

Data Collection

➢ Data Source

This synthetic healthcare dataset has been created to serve as a valuable resource for data science, machine learning, and data analysis enthusiasts. It is designed to mimic real-world healthcare data, enabling users to practice, develop, and showcase their data manipulation and analysis skills in the context of the healthcare industry.

➢ Data Sample View

Name	Age	Gender	Blood Type	Medical C...
Bobby JacksOn	30	Male	B-	Cancer
LesLie TErRy	62	Male	A+	Obesity

Date of A...	Doctor	Hospital	Insurance...	Billing Am...
2024-01-31	Matthew Smith	Sons and Miller	Blue Cross	18856.281305978155
2019-08-20	Samantha Davies	Kim Inc	Medicare	33643.327286577885

Room Nu...	Admissio...	Discharge...	Medication	Test Resu...
328	Urgent	2024-02-02	Paracetamol	Normal
265	Emergency	2019-08-26	Ibuprofen	Inconclusive

➢ Data Information

Each column provides specific information about the patient, their admission, and the healthcare services provided, making this dataset suitable for various data analysis and modeling tasks in the healthcare domain. Here's a brief explanation of each column in the dataset: -

- **Name:** This column represents the name of the patient associated with the healthcare record.
- **Age:** The age of the patient at the time of admission, expressed in years.
- **Gender:** Indicates the gender of the patient, either "Male" or "Female."
- **Blood Type:** The patient's blood type, which can be one of the common blood types (e.g., "A+", "O-", etc.).
- **Medical Condition:** This column specifies the primary medical condition or diagnosis associated with the patient, such as "Diabetes," "Hypertension," "Asthma," and more.
- **Date of Admission:** The date on which the patient was admitted to the healthcare facility.
- **Doctor:** The name of the doctor responsible for the patient's care during their admission.
- **Hospital:** Identifies the healthcare facility or hospital where the patient was admitted.
- **Insurance Provider:** This column indicates the patient's insurance provider, which can be one of several options, including "Aetna," "Blue Cross," "Cigna," "UnitedHealthcare," and "Medicare."
- **Billing Amount:** The amount of money billed for the patient's healthcare services during their admission. This is expressed as a floating-point number.
- **Room Number:** The room number where the patient was accommodated during their admission.
- **Admission Type:** Specifies the type of admission, which can be "Emergency," "Elective," or "Urgent," reflecting the circumstances of the admission.
- **Discharge Date:** The date on which the patient was discharged from the healthcare facility, based on the admission date and a random number of days within a realistic range.
- **Medication:** Identifies a medication prescribed or administered to the patient during their admission. Examples include "Aspirin," "Ibuprofen," "Penicillin," "Paracetamol," and "Lipitor."
- **Test Results:** Describes the results of a medical test conducted during the patient's admission. Possible values include "Normal," "Abnormal," or "Inconclusive," indicating the outcome of the test.

Data Exploration

➢ Shape

```
(55500, 15)
```

➢ Description

	Age	Billing Amount	Room Number
count	55500.000000	55500.000000	55500.000000
mean	51.539459	25539.316097	301.134829
std	19.602454	14211.454431	115.243069
min	13.000000	-2008.492140	101.000000
25%	35.000000	13241.224652	202.000000
50%	52.000000	25538.069376	302.000000
75%	68.000000	37820.508436	401.000000
max	89.000000	52764.276736	500.000000

➢ Information

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 55500 entries, 0 to 55499
Data columns (total 15 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Name                55500 non-null  object
 1   Age                 55500 non-null  int64
 2   Gender              55500 non-null  object
 3   Blood Type          55500 non-null  object
 4   Medical Condition   55500 non-null  object
 5   Date of Admission   55500 non-null  object
 6   Doctor              55500 non-null  object
 7   Hospital            55500 non-null  object
 8   Insurance Provider  55500 non-null  object
 9   Billing Amount      55500 non-null  float64
 10  Room Number         55500 non-null  int64
 11  Admission Type      55500 non-null  object
 12  Discharge Date      55500 non-null  object
 13  Medication          55500 non-null  object
 14  Test Results        55500 non-null  object
dtypes: float64(1), int64(2), object(12)
memory usage: 6.4+ MB
```

Data Cleaning

- **Duplicates**
  - **Before Handling:** 534 Duplicates
  - **After Handling:** 0 Duplicates

- **Missing Values:** No Missing Values

Data Analysis

Gender	count
Male	27496
Female	27470

dtype: int64

Blood Type	count
A-	6898
A+	6896
B+	6885
AB+	6882
AB-	6874
B-	6872
O+	6855
O-	6804

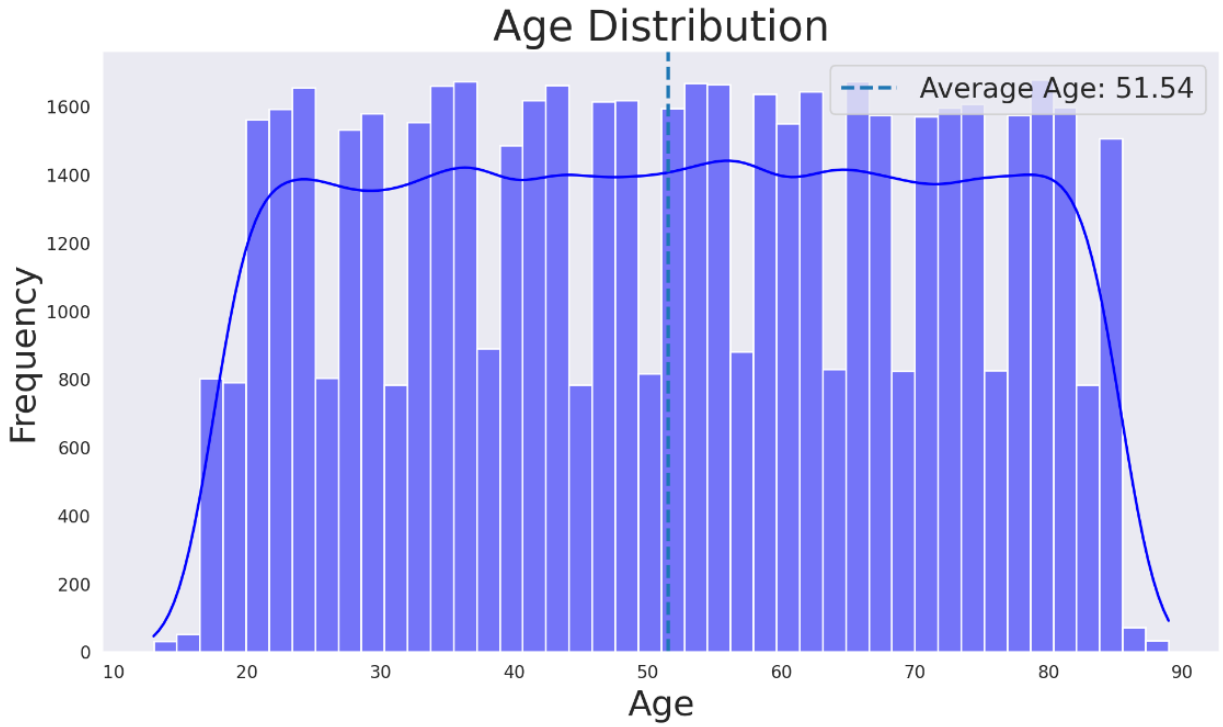
dtype: int64

Medical Condition	count
Arthritis	9218
Diabetes	9216
Hypertension	9151
Obesity	9146
Cancer	9140
Asthma	9095

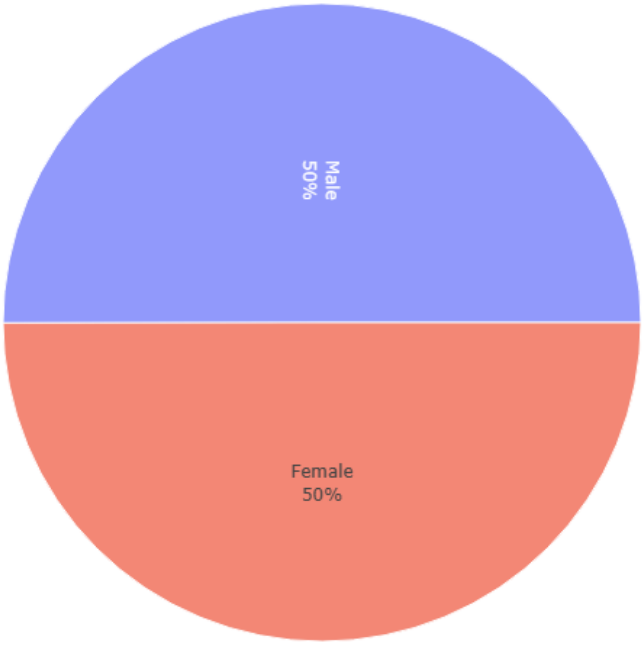
dtype: int64

	Insurance Provider	Billing Amount	Patients
0	Aetna	276498741.0	10822
1	Blue Cross	280409101.0	10952
2	UnitedHealthcare	279915371.0	11014
3	Medicare	282911027.0	11039
4	Cigna	284334099.0	11139

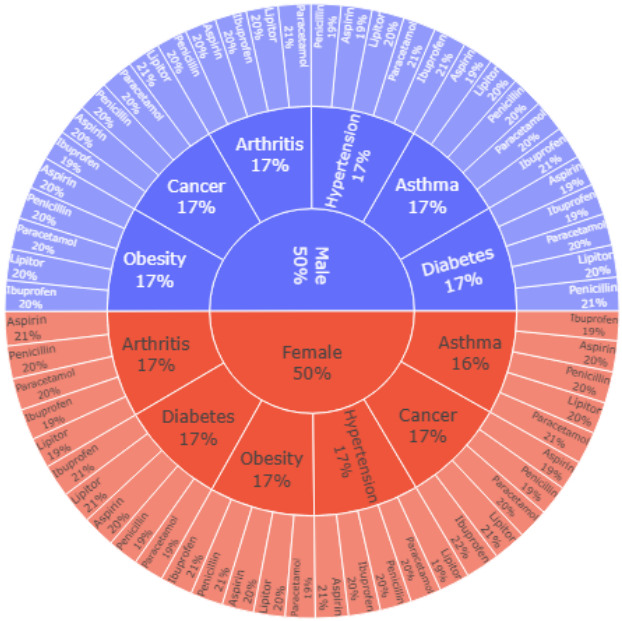
Data Visualization



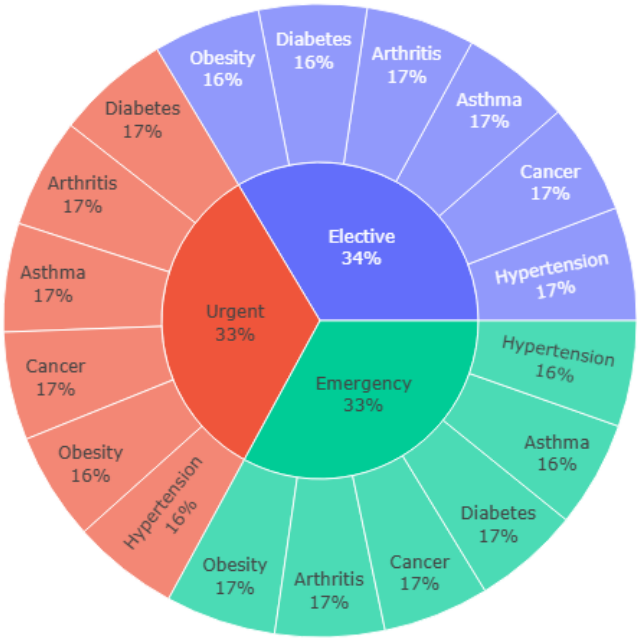
Patient's Gender



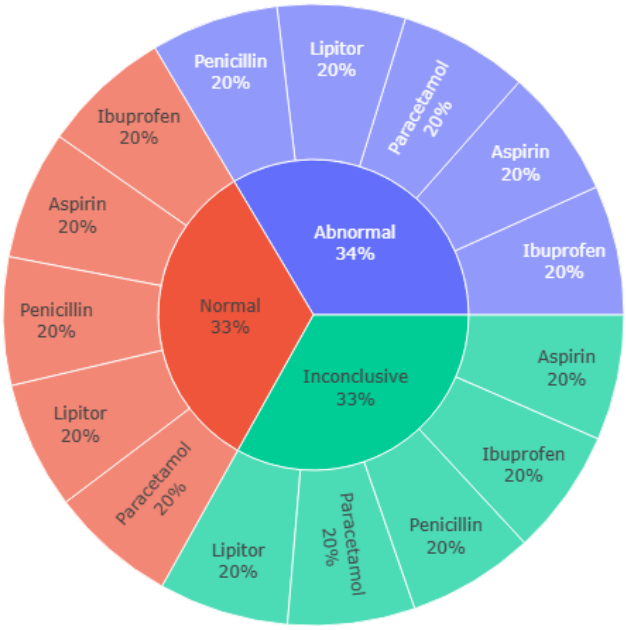
Patient's Status



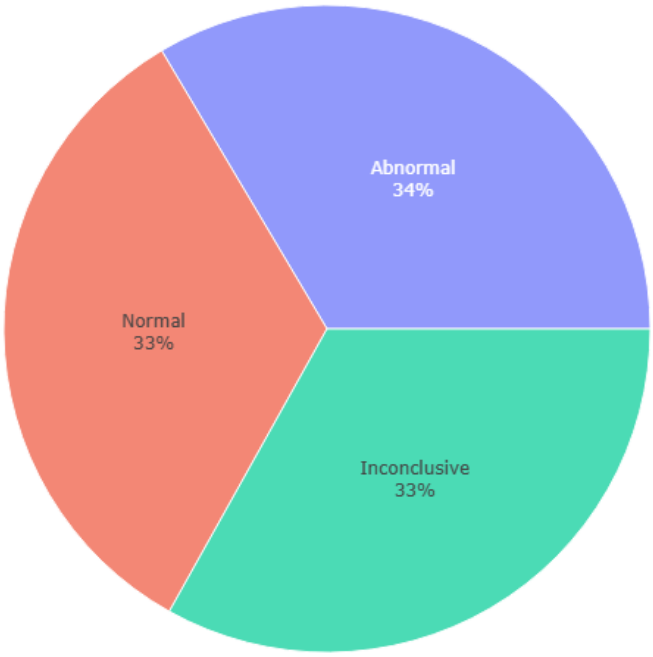
Admission Type by Medical Condition



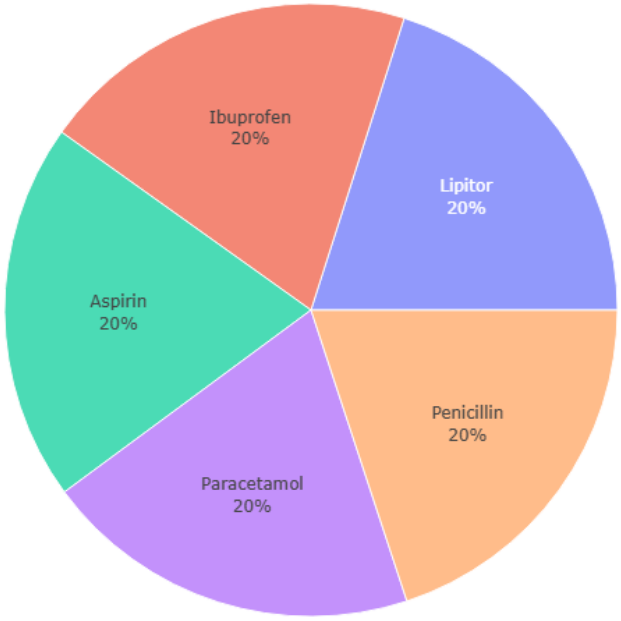
Test Results by Medication



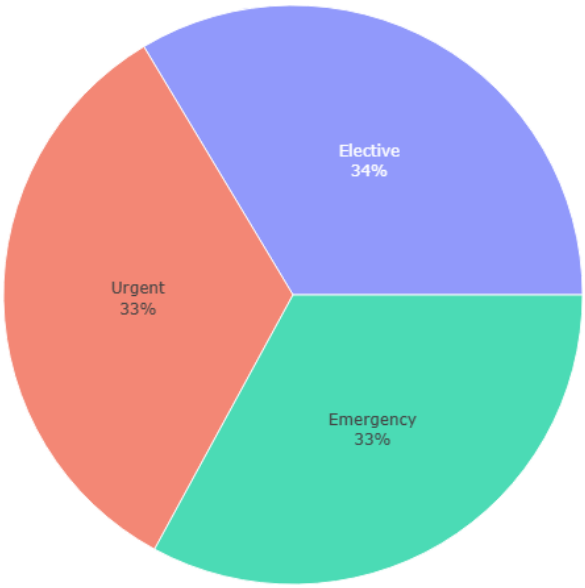
Patient's Test Results



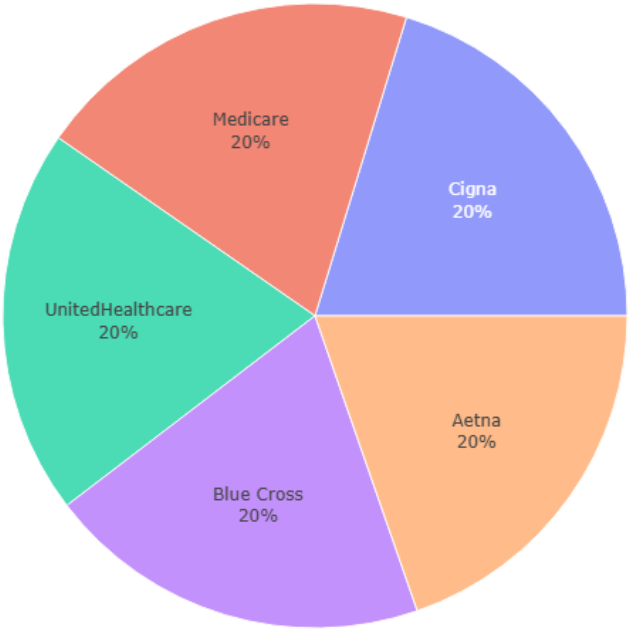
Patient's Medication



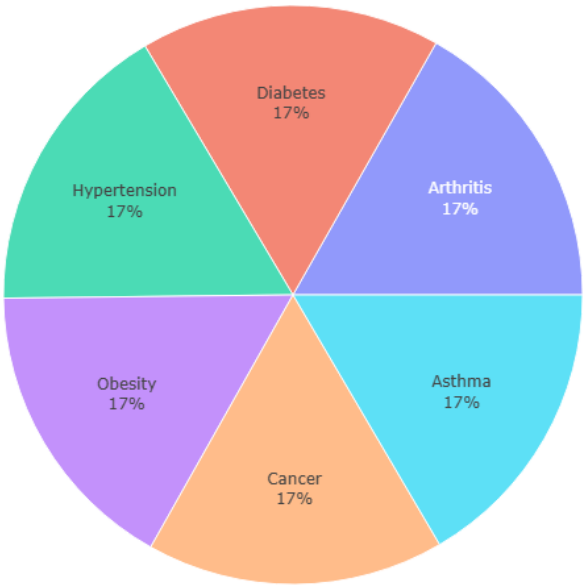
Patient's Admission Type



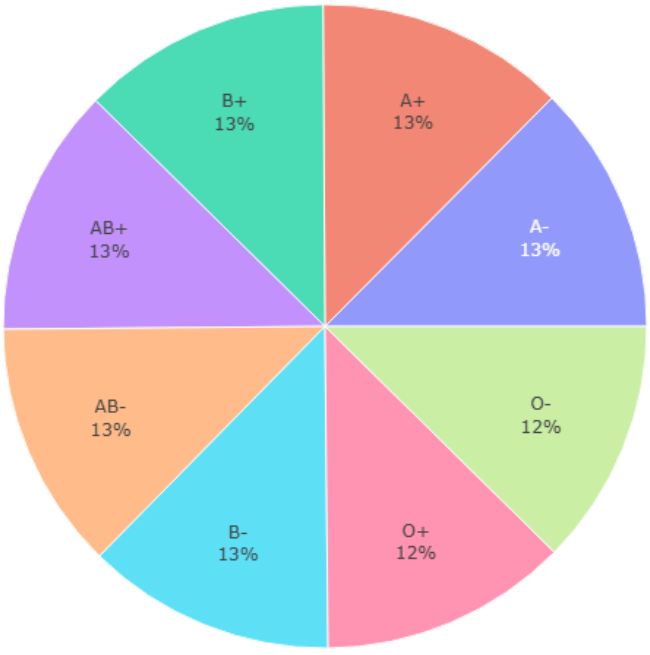
Patient's Insurance Provider



Patient's Medical Condition



Patient's Blood Type



Data Prediction

- **Prediction Target:** Test Results
- **Target Classes:** 3 (Normal – Abnormal – Inconclusive)
- **Prediction Features:** All Columns Except
- **Algorithm:** Decision Tree Classifier (DTC)
- **Methodology:** Data Fitting
- **Prediction Results: -**
  - **Precision Score:** 0.44931533227834247
  - **Recall Score:** 0.449
  - **F1 Score:** 0.449129234867101
  - **Accuracy: 44.90%**

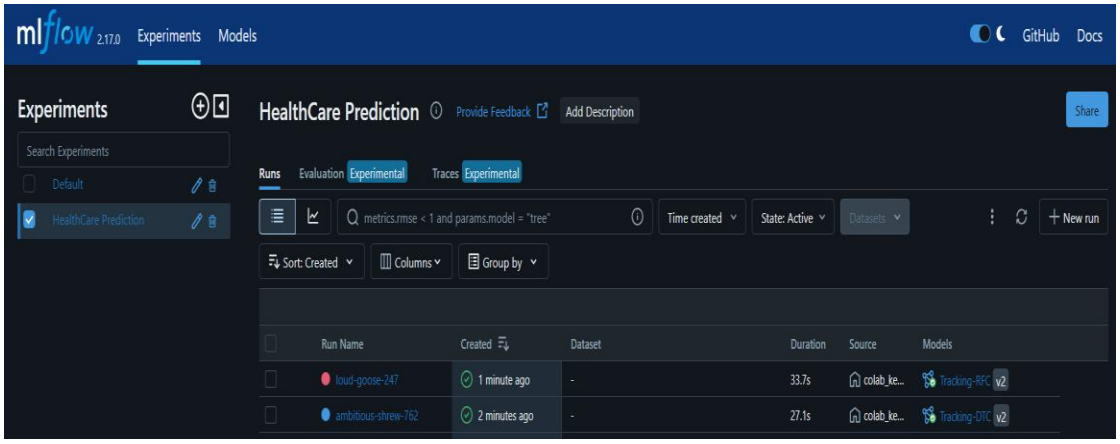
Data Optimization

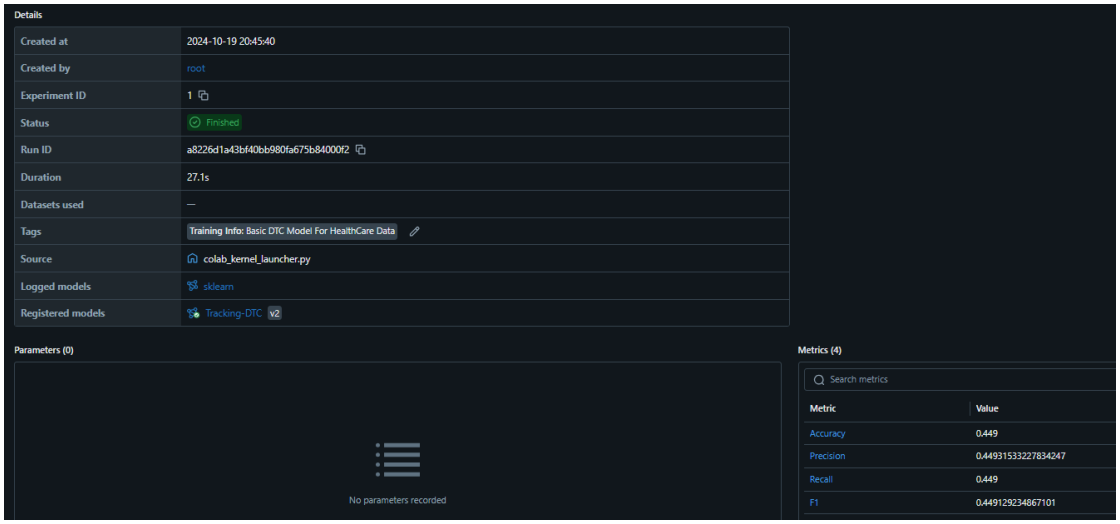
- **Algorithm: Random Forest Classifier (RFC)**
- **Methodology: Data Fitting**
- **Prediction Results After Optimization: -**
  - **Precision Score:** 0.48705099273936026
  - **Recall Score:** 0.4865
  - **F1 Score:** 0.4865200362489088
  - **Accuracy: 48.65%**

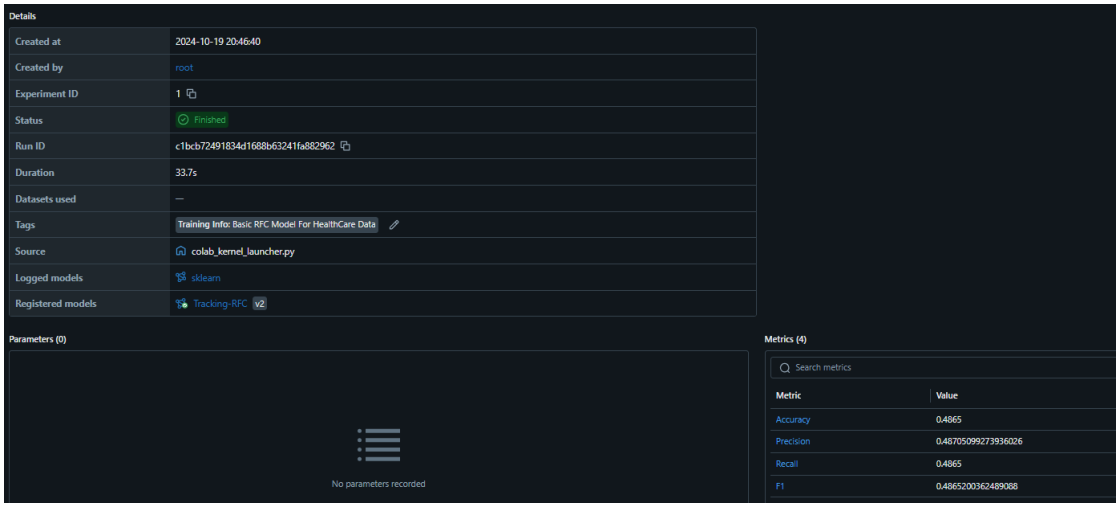
Machine Learning Flow (ML – Flow)

- **Experiment Name: HealthCare Prediction**
- **Server Tracking:**
  - **DTC Tracking**
  - **RFC Tracking**
- **Logged Metrics**
  - **Accuracy**
  - **Precision Score**
  - **Recall Score**
  - **F1 Score**

Flow Server View







Risks & Issues

❖ Prediction

In order to process with prediction, an issue in the data had to be handled which are string data types. The algorithm used for data prediction does not accept strings and complex data types and that would lead to a compile error. For that, a technical solution was applied to change all strings into numerical data types using label encoding.

○ Data Before Encoding

Name	Age	Gender	Blood Type	Medical Condition	Date of Admission	Doctor	Hospital	Insurance Provider	Billing Amount	Room Number	Admission Type	Discharge Date	Medication	Test Results
Bobby JacksOn	30	Male	B-	Cancer	2024-01-31	Matthew Smith	Sons and Miller	Blue Cross	18856.281306	328	Urgent	2024-02-02	Paracetamol	Normal
LesLie TErRy	62	Male	A+	Obesity	2019-08-20	Samantha Davies	Kim Inc	Medicare	33643.327287	265	Emergency	2019-08-26	Ibuprofen	Inconclusive
DaNnY sMitH	76	Female	A-	Obesity	2022-09-22	Tiffany Mitchell	Cook PLC	Aetna	27955.096079	205	Emergency	2022-10-07	Aspirin	Normal
andrEw waTtS	28	Female	O+	Diabetes	2020-11-18	Kevin Wells	Hernandez Rogers and Vang,	Medicare	37909.782410	450	Elective	2020-12-18	Ibuprofen	Abnormal
adrIENNE bEll	43	Female	AB+	Cancer	2022-09-19	Kathleen Hanna	White-White	Aetna	14238.317814	458	Urgent	2022-10-09	Penicillin	Abnormal

○ Data After encoding

Name	Age	Gender	Blood Type	Medical Condition	Date of Admission	Doctor	Hospital	Insurance Provider	Billing Amount	Room Number	Admission Type	Discharge Date	Medication	Test Results
3068	30	1	5	2	1729	26612	29933	1	18856.281306	328	2	1730	3	2
15211	62	1	0	5	104	33648	16012	3	33643.327287	265	1	109	1	1
6476	76	0	1	5	1233	37828	5473	0	27955.096079	205	1	1247	0	2
26935	28	0	6	3	560	22511	12317	3	37909.782410	450	0	589	1	0
26241	43	0	2	2	1230	21259	33598	0	14238.317814	458	2	1249	4	0

❖ Accuracy

The decision tree algorithm, while popular for its simplicity and interpretability, can sometimes exhibit weak accuracy. Here are several reasons for this:

1) **Overfitting:** Decision trees can become overly complex, capturing noise in the training data rather than the underlying patterns.
2) **Underfitting:** If the tree is too shallow or has insufficient splits, it may not capture the complexities of the data.
3) **Imbalanced Data:** If one class is significantly more frequent than others, the tree may become biased toward that class.
4) **Feature Selection:** Irrelevant or redundant features can lead to suboptimal splits and reduce accuracy.
5) **Data Quality:** Noisy data, outliers, or missing values can negatively impact the model's performance.
6) **Bias in Data:** If the training data is not representative of the actual problem domain, the model may not generalize well.
7) **Lack of Ensemble Methods:** Single decision trees may not capture all patterns effectively compared to ensemble methods.
8) **Interpretability vs. Accuracy Trade-off:** Simpler models may sacrifice accuracy for interpretability. Decision trees tend to favor transparency over performance.