IE 7374 MACHINE LEARNING OPERATIONS

DATA PIPELINE ASSIGNMENT

Parkinsons Disease Prediction using end-to end machine learning pipeline

Introduction:

This project aims to predict Parkinson's disease using machine learning algorithms and MLOPS techniques. By analyzing biomedical data such as demographic attributes, motor skills, and other relevant biomarkers this model can assist in early identification of Parkinson's disease symptoms. Parkinson's disease is a progressive neurological disorder with no known cure, but early detection can significantly improve patient outcomes by enabling earlier interventions. Predictive models help in identifying the disease at an early stage when treatments can be more effective in managing symptoms, thereby improving the quality of life for affected individuals. This project supports healthcare professionals by providing a tool for early detection, which can aid in timely diagnosis and treatment planning. Additionally, it can be a valuable resource for researchers studying Parkinson's disease, potentially contributing to the discovery of new biomarkers or insights into disease progression.

Dataset Overview:

Following are the features we have considered after performing regressive data analysis on the acquired data files:

Participant_status, Demographics, Biospecimen_Analysis, and Motor assessments.

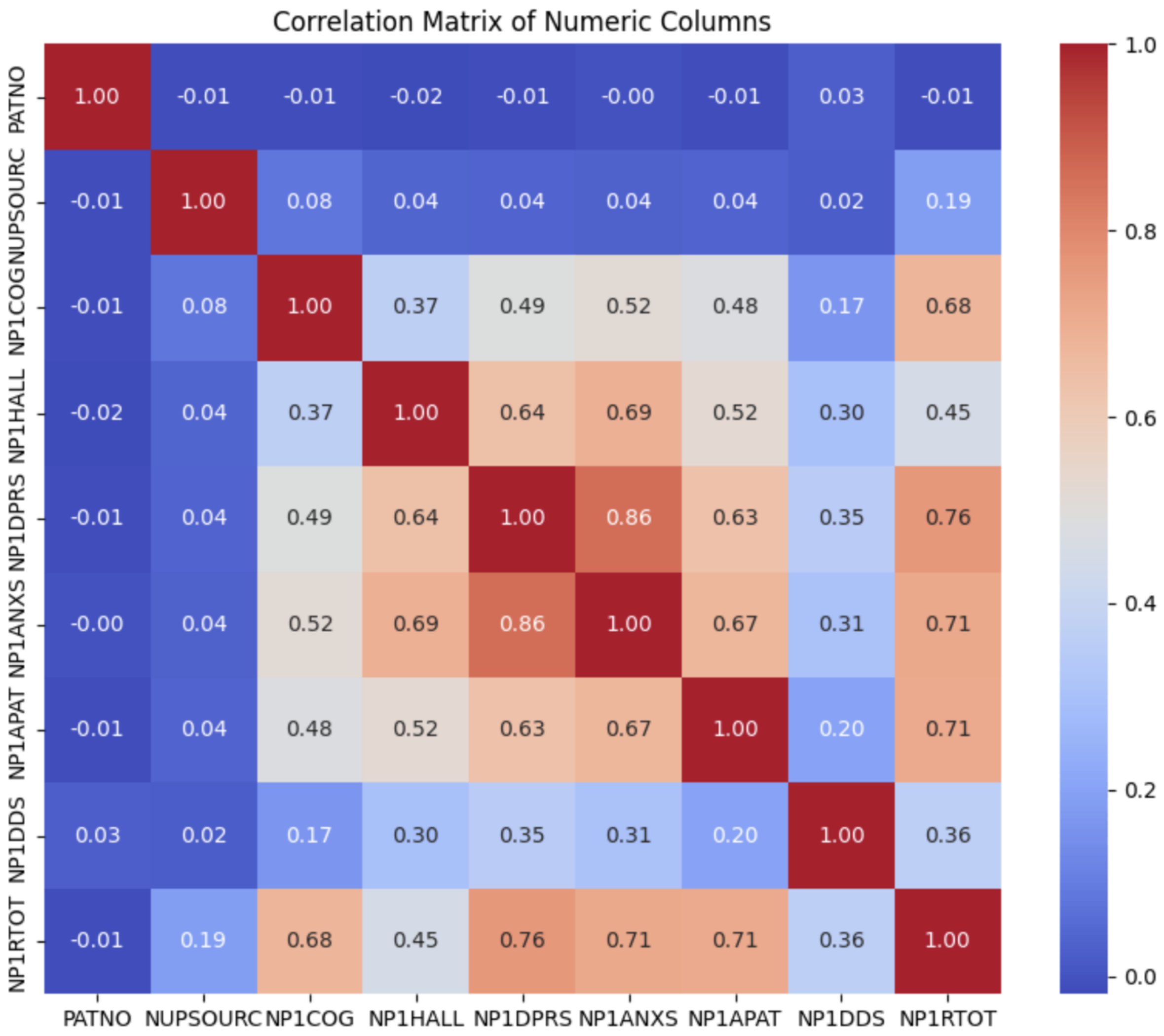
Column	Description
PATNO	Participant_ID
COHORT	Enrollment Cohort
ENROLL_DATE	Enrollment Date
ENROLL_AGE	Age at Enrollment
ENRLPRKN	Parkin Mutation at Enrollment
ENRLSRDC	Sporadic PD at Enrollment
ENRLHPSM	Hyposmia / Generalized Risk at Enrollment
ENRLRBD	RBD at Enrollment
ENRLSNCA	SNCA Mutation at Enrollment
ENRLGBA	GBA Mutation at Enrollment
SEX	Sex of participant at birth

CHLDBEAR	Female of childbearing potential
SAAMethod	Method used for analysis
SAA_Status	Qualitative status (Positive, Negative, or Inconclusive)
SAA_Type	SAA Type (NA, Type1, Type2, Undetermined)
InstrumentRep1	Instrument (Rep1)
PATNO	Participant ID
NUPSOURC	Primary Source of Information
NP1COG	COGNITIVE IMPAIRMENT
NP1HALL	HALLUCINATIONS AND PSYCHOSIS
NP1DPRS	DEPRESSED MOODS
NP1ANXS	ANXIOUS MOOD
NP1APAT	APATHY
NP1DDS	FEATURES-DOPAMINE DYSREGULATION SYNDROME
NP1SLPN	SLEEP PROBLEMS (NIGHT)
NP1SLPD	DAYTIME SLEEPINESS
NP1PAIN	PAIN AND OTHER SENSATIONS
NP1URIN	URINARY PROBLEMS
NP1CNST	CONSTIPATION PROBLEMS
NP1LTHD	LIGHTHEADEDNESS ON STANDING
NP1FATG	FATIGUE
NP2SPCH	SPEECH
NP2SALV	SALIVA + DROOLING
NP2SWAL	CHEWING AND SWALLOWING
NP2EAT	EATING TASKS
NP2DRES	DRESSING
NP2HYGN	HYGIENE
NP2HWRT	HANDWRITING
NP2HOBB	DOING HOBBIES AND OTHER ACTIVITIES
NP2TURN	TURNING IN BED
NP2TRMR	TREMOR
NP2RISE	GETTING OUT OF BED, CAR, OR DEEP CHAIR
NP2WALK	WALKING AND BALANCE
NP2FREZ	FREEZING

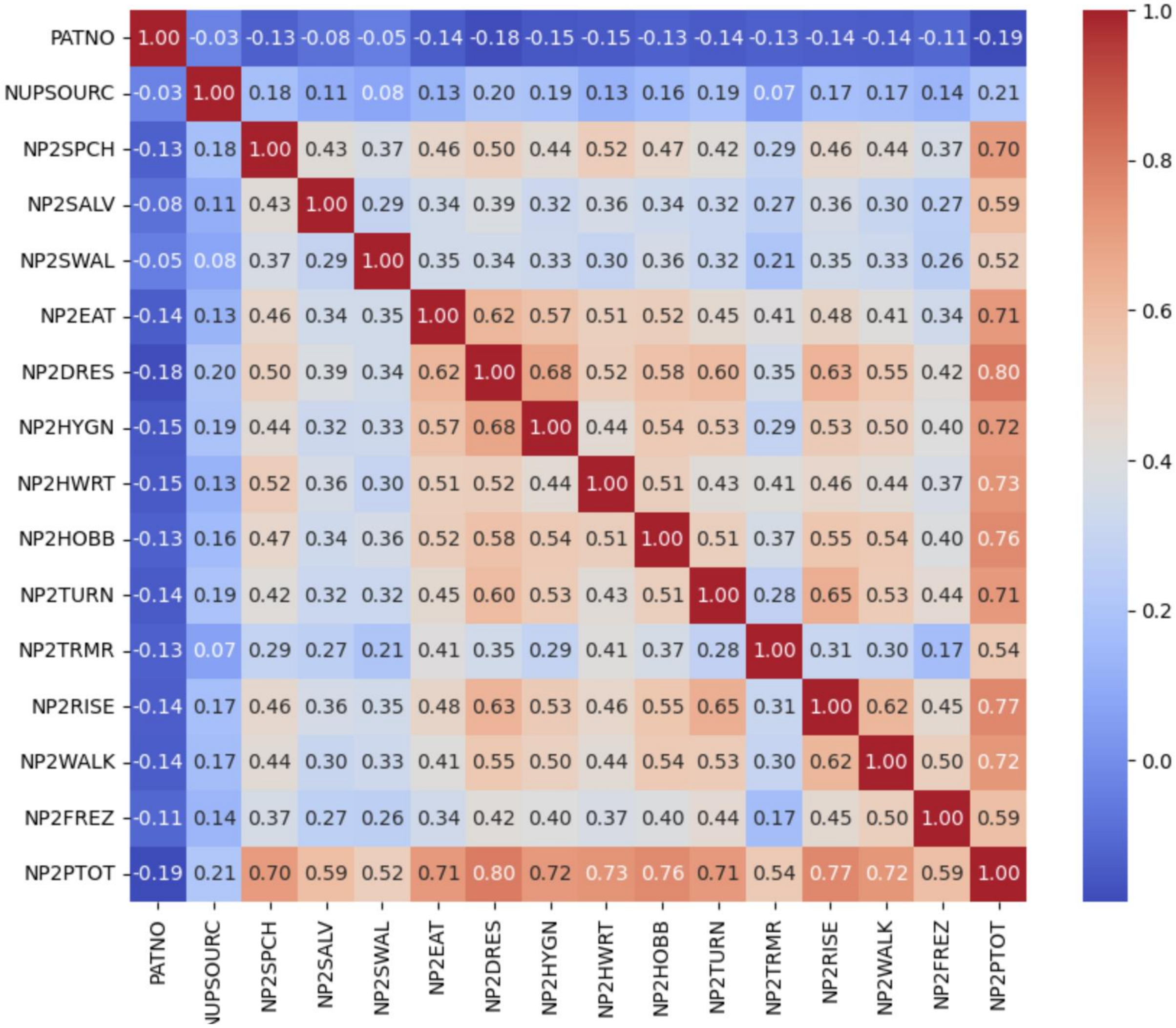
NP3SPCH	3.1 Speech
NP3FACXP	3.2 Facial expression
NP3RIGN	3.3a Rigidity - Neck
NP3RIGRU	3.3b Rigidity - RUE
NP3RIGLU	3.3c Rigidity - LUE
NP3RIGRL	3.3d Rigidity - RLE
NP3RIGLL	3.3e Rigidity - LLE
NP3FTAPR	3.4a Finger Tapping Right Hand
NP3FTAPL	3.4b Finger Tapping Left Hand
NP3HMOVR	3.5a Hand movements - Right Hand
NP3HMOVL	3.5b Hand movements - Left Hand
NP3PRSPR	3.6a Pronation-Supination - Right Hand
NP3PRSPL	3.6b Pronation-Supination - Left Hand
NP3TTAPR	3.7a Toe tapping - Right foot
NP3TTAPL	3.7b Toe tapping - Left foot
NP3LGAGR	3.8a Leg agility - Right leg
NP3LGAGL	3.8b Leg agility - Left leg
NP3RISNG	3.9 Arising from chair
NP3GAIT	3.10 Gait
NP3FRZGT	3.11 Freezing of gait
NP3PSTBL	3.12 Postural stability
NP3POSTR	3.13 Posture
NP3BRADY	3.14 Global spontaneity of movement
NP3PTRMR	3.15a Postural tremor - Right Hand
NP3PTRML	3.15b Postural tremor - Left hand
NP3KTRMR	3.16a Kinetic tremor - Right hand
NP3KTRML	3.16b Kinetic tremor - Left hand
NP3RTARU	3.17a Rest tremor amplitude - RUE
NP3RTALU	3.17b Rest tremor amplitude - LUE
NP3RTARL	3.17c Rest tremor amplitude - RLE
NP3RTALL	3.17d Rest tremor amplitude - LLE
NP3RTALJ	3.17e Rest tremor amplitude - Lip/jaw
NP3RTCON	3.18 Constancy of rest tremor

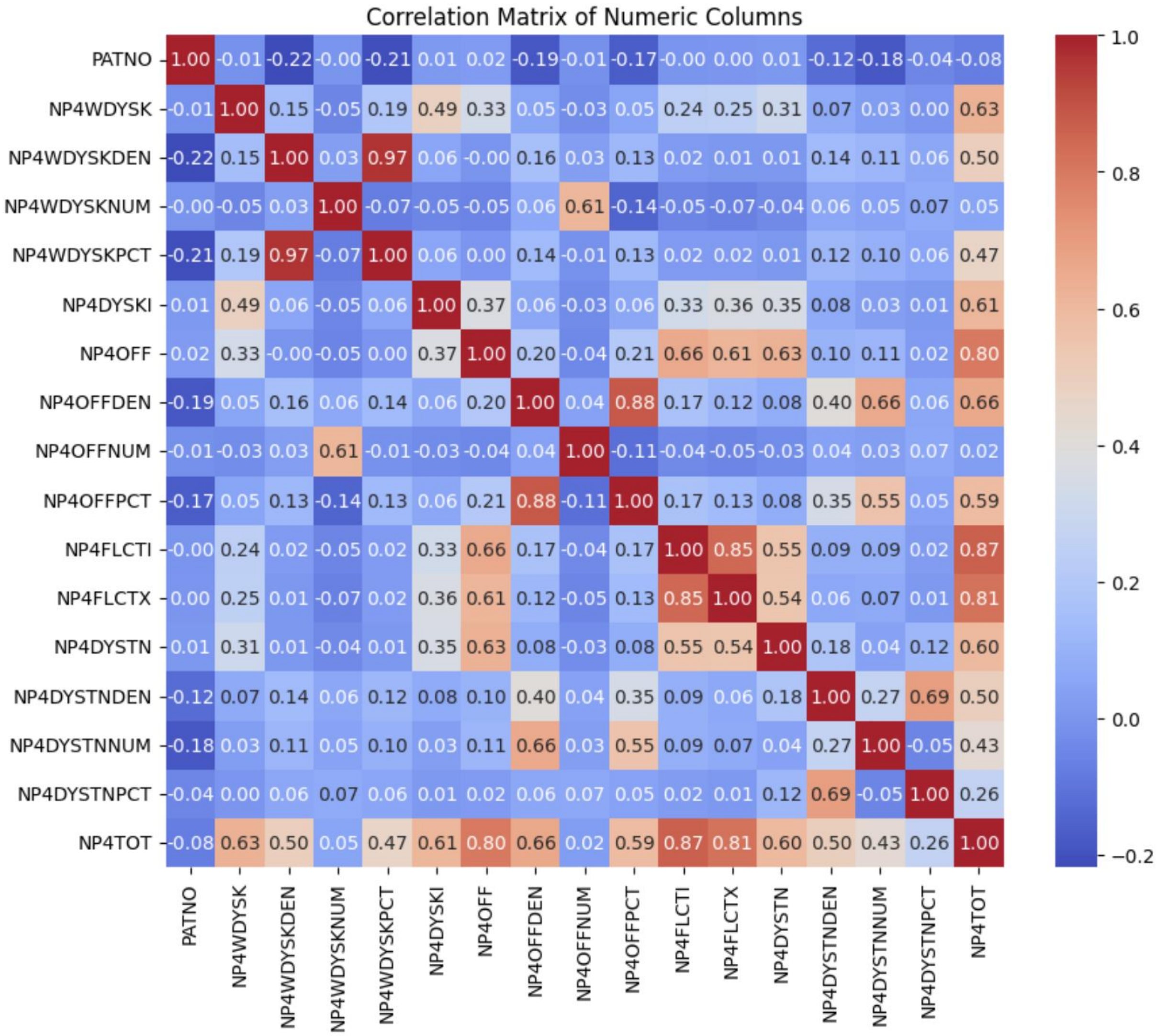
DYSKPRES	3.19 Were dyskinesias present
DYSKIRAT	3.20 Did movements interfere with rating
NHY	3.21 Hoehn and Yahr Stage
NP4WDYSK	4.1 Time spent with dyskinesias
NP4WDYSKDEN	4.1 Total Hours with Dyskinesia
NP4WDYSKNUM	4.1 Total Hours Awake
NP4WDYSKPCT	4.1 % Dyskinesia
NP4DYSKI	4.2 Functional impact of dyskinesias
NP4OFF	4.3 Time spent in the OFF state
NP4OFFDEN	4.3 Total Hours OFF
NP4OFFNUM	4.3 Total Hours Awake
NP4OFFPCT	4.3 % OFF
NP4FLCTI	4.4 Functional impact of fluctuations
NP4FLCTX	4.5 Complexity of motor fluctuations
NP4DYSTN	4.6 Painful OFF-state dystonia
NP4DYSTNDEN	4.6 Total Hours OFF with Dystonia
NP4DYSTNNUM	4.6 Total Hours OFF
NP4DYSTNPCT	4.6 % OFF Dystonia

We performed correlation analysis on all the data and selected features that had 50% or more variance compared to the total variance.


Correlation Matrix of Numeric Columns
PATNO
NUPSOURC
NP1COG
NP1HALL
NP1DPRS
NP1ANXS
NP1APAT
NP1DDS
NP1RTOT
1.00 -0.01 -0.01 -0.02 -0.01 -0.00 -0.01 0.03 -0.01
-0.01 1.00 0.08 0.04 0.04 0.04 0.04 0.02 0.19
-0.01 0.08 1.00 0.37 0.49 0.52 0.48 0.17 0.68
-0.02 0.04 0.37 1.00 0.64 0.69 0.52 0.30 0.45
-0.01 0.04 0.49 0.64 1.00 0.86 0.63 0.35 0.76
-0.00 0.04 0.52 0.69 0.86 1.00 0.67 0.31 0.71
-0.01 0.04 0.48 0.52 0.63 0.67 1.00 0.20 0.71
0.03 0.02 0.17 0.30 0.35 0.31 0.20 1.00 0.36
-0.01 0.19 0.68 0.45 0.76 0.71 0.71 0.36 1.00
PATNO
NUPSOURC
NP1COG
NP1HALL
NP1DPRS
NP1ANXS
NP1APAT
NP1DDS
NP1RTOT
1.0
0.8
0.6
0.4
0.2
0.0

Correlation Matrix of Numeric Columns





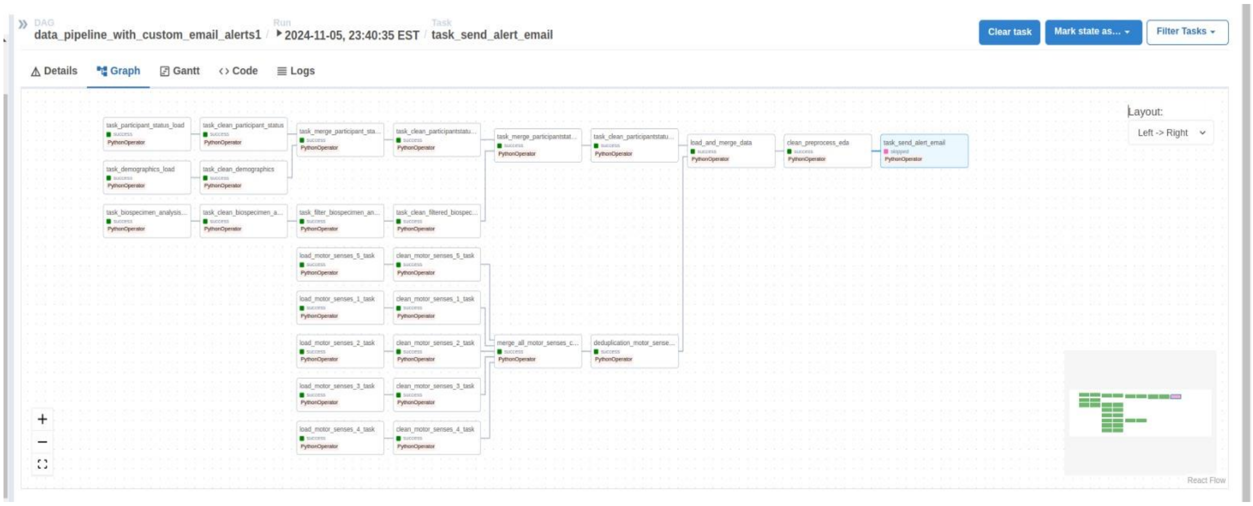
	PATNO	NP4WDYSK	NP4WDYSKDEN	NP4WDYSKNUM	NP4WDYSKPCT	NP4DYSKI	NP4OFF	NP4OFFDEN	NP4OFFNUM	NP4OFFPCT	NP4FLCTI	NP4FLCTX	NP4DYSTN	NP4DYSTNDEN	NP4DYSTNNUM	NP4DYSTNPCT	NP4TOT
PATNO	1.00	-0.01	-0.22	-0.00	-0.21	0.01	0.02	-0.19	-0.01	-0.17	-0.00	0.00	0.01	-0.12	-0.18	-0.04	-0.08
NP4WDYSK	-0.01	1.00	0.15	-0.05	0.19	0.49	0.33	0.05	-0.03	0.05	0.24	0.25	0.31	0.07	0.03	0.00	0.63
NP4WDYSKDEN	-0.22	0.15	1.00	0.03	0.97	0.06	-0.00	0.16	0.03	0.13	0.02	0.01	0.01	0.14	0.11	0.06	0.50
NP4WDYSKNUM	-0.00	-0.05	0.03	1.00	-0.07	-0.05	-0.05	0.06	0.61	-0.14	-0.05	-0.07	-0.04	0.06	0.05	0.07	0.05
NP4WDYSKPCT	-0.21	0.19	0.97	-0.07	1.00	0.06	0.00	0.14	-0.01	0.13	0.02	0.02	0.01	0.12	0.10	0.06	0.47
NP4DYSKI	0.01	0.49	0.06	-0.05	0.06	1.00	0.37	0.06	-0.03	0.06	0.33	0.36	0.35	0.08	0.03	0.01	0.61
NP4OFF	0.02	0.33	-0.00	-0.05	0.00	0.37	1.00	0.20	-0.04	0.21	0.66	0.61	0.63	0.10	0.11	0.02	0.80
NP4OFFDEN	-0.19	0.05	0.16	0.06	0.14	0.06	0.20	1.00	0.04	0.88	0.17	0.12	0.08	0.40	0.66	0.06	0.66
NP4OFFNUM	-0.01	-0.03	0.03	0.61	-0.01	-0.03	-0.04	0.04	1.00	-0.11	-0.04	-0.05	-0.03	0.04	0.03	0.07	0.02
NP4OFFPCT	-0.17	0.05	0.13	-0.14	0.13	0.06	0.21	0.88	-0.11	1.00	0.17	0.13	0.08	0.35	0.55	0.05	0.59
NP4FLCTI	-0.00	0.24	0.02	-0.05	0.02	0.33	0.66	0.17	-0.04	0.17	1.00	0.85	0.55	0.09	0.09	0.02	0.87
NP4FLCTX	0.00	0.25	0.01	-0.07	0.02	0.36	0.61	0.12	-0.05	0.13	0.85	1.00	0.54	0.06	0.07	0.01	0.81
NP4DYSTN	0.01	0.31	0.01	-0.04	0.01	0.35	0.63	0.08	-0.03	0.08	0.55	0.54	1.00	0.18	0.04	0.12	0.60
NP4DYSTNDEN	-0.12	0.07	0.14	0.06	0.12	0.08	0.10	0.40	0.04	0.35	0.09	0.06	0.18	1.00	0.27	0.69	0.50
NP4DYSTNNUM	-0.18	0.03	0.11	0.05	0.10	0.03	0.11	0.66	0.03	0.55	0.09	0.07	0.04	0.27	1.00	-0.05	0.43
NP4DYSTNPCT	-0.04	0.00	0.06	0.07	0.06	0.01	0.02	0.06	0.07	0.05	0.02	0.01	0.12	0.69	-0.05	1.00	0.26
NP4TOT	-0.08	0.63	0.50	0.05	0.47	0.61	0.80	0.66	0.02	0.59	0.87	0.81	0.60	0.50	0.43	0.26	1.00

Data Sources:

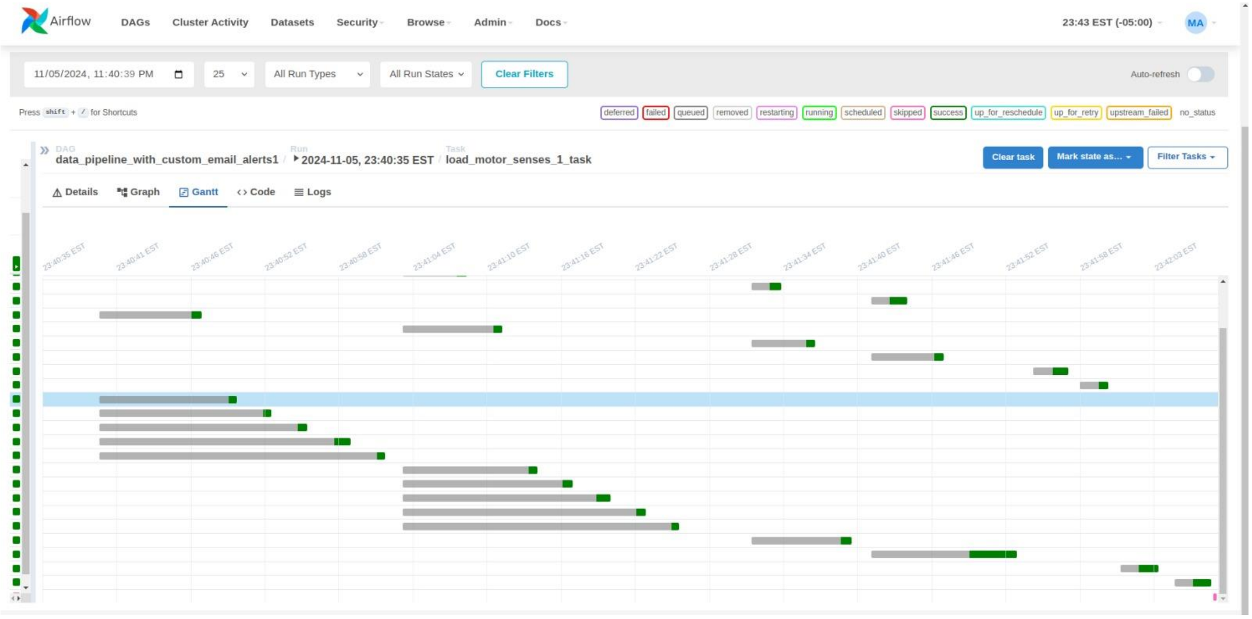
The data is taken from Parkinson's Progression Markers Initiative (PPMI).

Airflow DAG:

Data Pipeline Components:



DAG
data_pipeline_with_custom_email_alerts1
Run
2024-11-05, 23:40:35 EST
Task
task_send_alert_email
Clear task
Mark state as...
Filter Tasks
Details
Graph
Gantt
Code
Logs
Layout:
Left -> Right
task_participant_status_load
task_clean_participant_status
task_demographics_load
task_clean_demographics
task_biospecimen_analysis...
task_clean_biospecimen_a...
task_merge_participant_sta...
task_clean_participantstatu...
task_filter_biospecimen_an...
task_clean_filtered_biospec...
task_merge_participantstat...
task_clean_participantstatu...
load_and_merge_data
clean_preprocess_eda
task_send_alert_email
skipped
load_motor_senses_5_task
clean_motor_senses_5_task
load_motor_senses_1_task
clean_motor_senses_1_task
load_motor_senses_2_task
clean_motor_senses_2_task
load_motor_senses_3_task
clean_motor_senses_3_task
load_motor_senses_4_task
clean_motor_senses_4_task
merge_all_motor_senses_c...
deduplication_motor_sense...
success
PythonOperator
React Flow



Description of the Data Pipeline Components:

- **send_custom_alert_email**: Sends a custom alert email if a task fails or is retried, with details about the task and DAG.
- **participant_status_load**: Loads the "Participant_Status" CSV file into a DataFrame.
- **demographics_load**: Loads the "Demographics" CSV file into a DataFrame.
- **clean_participant_status**: Cleans the "Participant_Status" DataFrame by converting enrollment dates, renaming a column, and dropping unnecessary columns.
- **clean_demographics**: Cleans the "Demographics" DataFrame by dropping columns that are not needed.

- **merge_participant_status_and_demographics**: Merges the cleaned "Participant_Status" and "Demographics" DataFrames on the participant ID and filters rows with valid enrollment statuses.
- **clean_participantstatus_demographic**: Further cleans the merged "Participant_Status" and "Demographics" DataFrame by dropping additional unnecessary columns.
- **biospecimen_analysis_load**: Loads the "SAA_Biospecimen_Analysis_Results" CSV file into a DataFrame.
- **clean_biospecimen_analysis**: Cleans the "Biospecimen_Analysis" DataFrame by formatting dates and dropping irrelevant columns.
- **filter_biospecimen_analysis**: Filters the "Biospecimen_Analysis" DataFrame to keep only records with the earliest run date for baseline clinical events.
- **clean_filtered_biospecimen_analysis**: Further cleans the filtered "Biospecimen_Analysis" DataFrame by dropping additional columns.
- **merge_participantstatus_demographics_biospecimen_analysis**: Merges the cleaned "Participant_Status", "Demographics", and "Biospecimen_Analysis" DataFrames.
- **clean_participantstatus_demographics_biospecimen_analysis**: Final cleanup of the merged DataFrame by dropping remaining unnecessary columns.
- **load_motor_senses_1**: Loads the first motor senses CSV file into a DataFrame.
- **load_motor_senses_2**: Loads the second motor senses CSV file into a DataFrame.
- **load_motor_senses_3**: Loads the third motor senses CSV file into a DataFrame.
- **load_motor_senses_4**: Loads the fourth motor senses CSV file into a DataFrame.
- **load_motor_senses_5**: Loads the fifth motor senses CSV file into a DataFrame.
- **clean_motor_senses_1**: Cleans the first motor senses DataFrame by dropping unnecessary columns after retrieving it from XCom.
- **clean_motor_senses_2**: Cleans the second motor senses DataFrame by dropping unnecessary columns after retrieving it from XCom.
- **clean_motor_senses_3**: Cleans the third motor senses DataFrame by dropping unnecessary columns after retrieving it from XCom.
- **clean_motor_senses_4**: Cleans the fourth motor senses DataFrame by dropping unnecessary columns after retrieving it from XCom.
- **clean_motor_senses_5**: Cleans the fifth motor senses DataFrame by dropping unnecessary columns after retrieving it from XCom.
- **merge_all_motor_senses_csvs**: Merges all cleaned motor senses DataFrames into a single DataFrame and pushes the merged DataFrame to XCom.
- **drop_duplicate_motor_senses_columns**: Removes duplicate columns from the merged DataFrame and saves the final deduplicated DataFrame to a CSV file.

Cleaning Data:

In this phase, the dataset undergoes various cleaning and preprocessing steps to ensure data quality and readiness for analysis. The following modules are involved in this process:

- motor_dag.py: Contains the clean_csv function, responsible for refining raw data by implementing standard cleaning operations such as handling missing values or incorrect data formats.
- preprocessing.py: Detects and handles outliers, a key part of data cleaning to ensure data quality by removing or adjusting extreme values.
- cleaned_data.csv: Stores data post-cleaning, serving as the base for further analysis and modeling.
- participantstatus_demographics_biospecimen_dag.py: Handles participant demographic and biospecimen data, ensuring uniformity and quality in demographic data processing.

These scripts and notebooks collectively ensured that the data is clean and consistent before moving on to feature engineering.

Feature Engineering:

In this step, we perform feature engineering to analyze and modify the features to further improve the training and improve the results and evaluation metrics. The following modules are created for feature engineering:

- correlation.py - Analyzes correlations between features, helping to identify relationships and select the most relevant features for the model.
- resampling.py - Handles class imbalance, which can be an essential step in feature engineering, particularly for classification tasks.
- pca.py - Although primarily for dimensionality reduction, it prepares the data by identifying and selecting the most significant features, which is a vital step in preprocessing.

These files help in refining and creating meaningful features to enhance the performance of your Parkinson’s Disease prediction model.