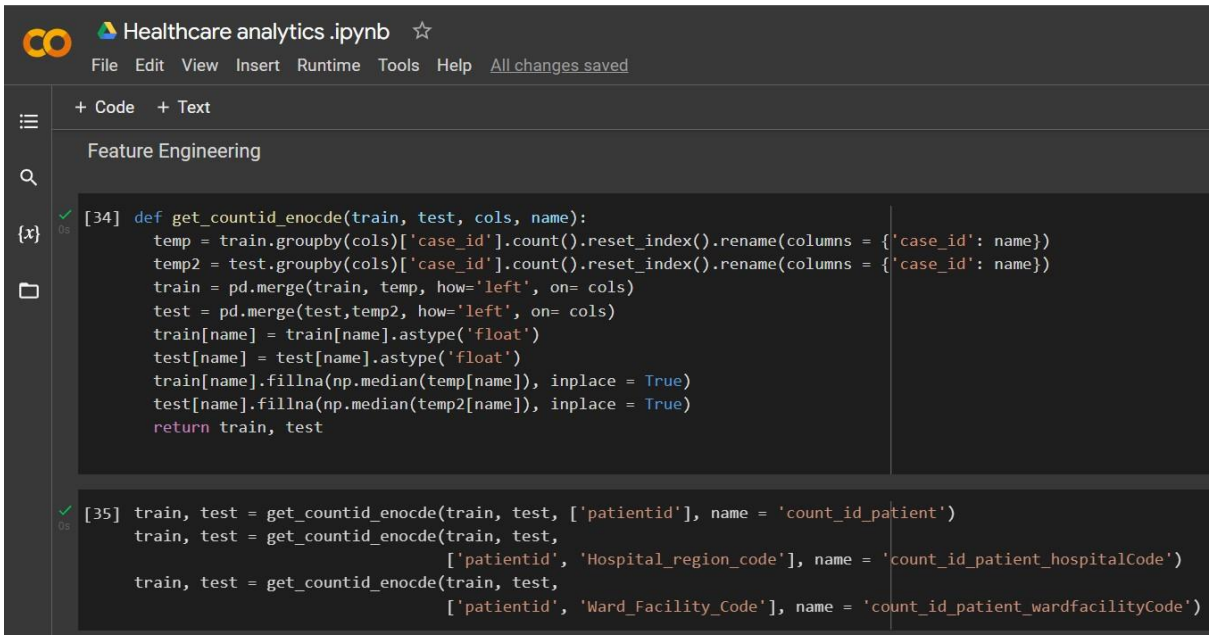
SPRINT 3

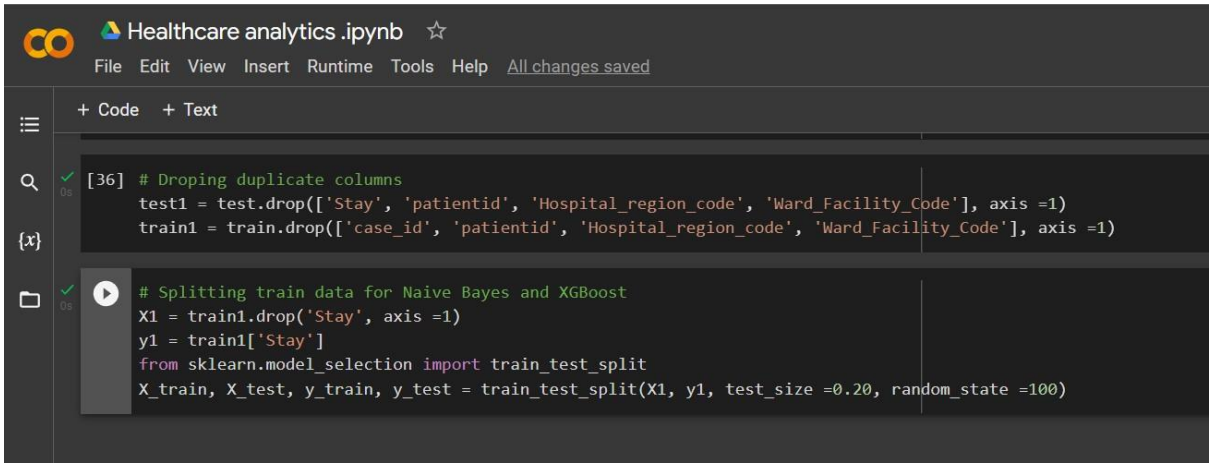
Date	13-11-2022
Team ID	PNT2022TMID23263
Project Title	Analytics for Hospitals' Health-Care Data
Team Members	Venkateshbabu B B, Pandidurai M, Udaya Shankar M, Abdul Rahuman R

Feature Engineering

Once the data is cleaned and prepared, we grouped patientid and case_id to extract the new column “count_id_patient”. This variable contains the count of multiple admits of a patient under different case_id. Further two more columns “Hospital_region_code” and “ward_facility_code” were grouped to patientid and case_id. These two new variables “count_id_patient_hospitalCode” and “count_id_patient_wardfacilityCode” contain the count of multiple admissions in a hospital region and the count of multiple wards allocated to a patient.

Before getting into analysis, the train data must be split into two parts, the first part with all the feature variables and the second part with a target variable (“Stay”). Then preprocessed into train and validation sets. So, here we are portioning the train set with 80% and validation set with 20% of the data for Naïve Bayes and XGBoost models.





Choosing ML models for analysis:

The goal is to predict Length of Stay i.e., "Stay" column (Target Variable) and it is classified into 11 levels. We must find the probability of each patient's length of stay using feature variables, which contain the patient's condition and hospital-level information.

After analysis of different ML models for classification, we have decided to choose three models:

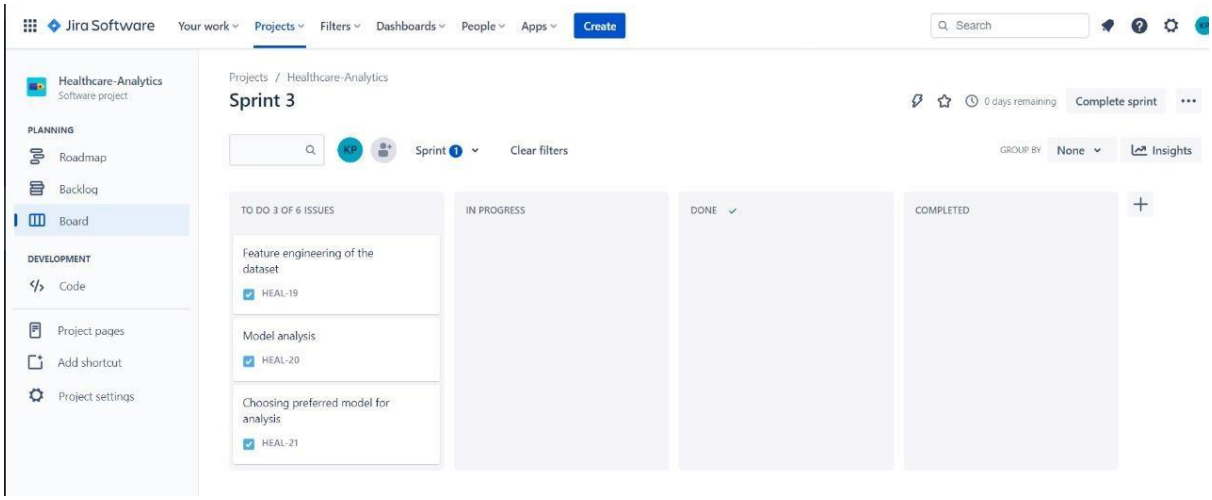
1) Naives Bayes Model
2) XGboost
3) Neural Networks

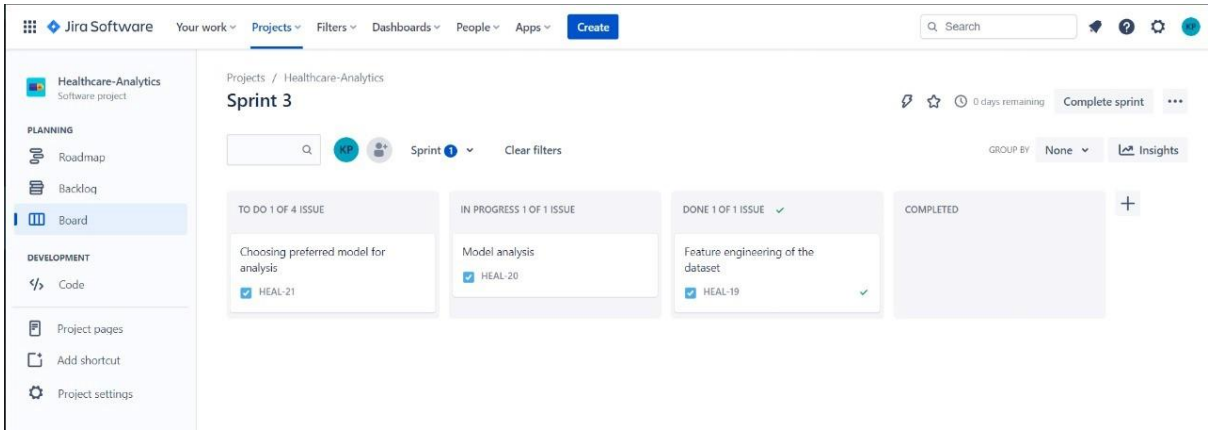
We have chosen Naïve Bayes model, as the feature variables are ordinal in nature and also Naïve Bayes Model is a perfect multilevel classifier.

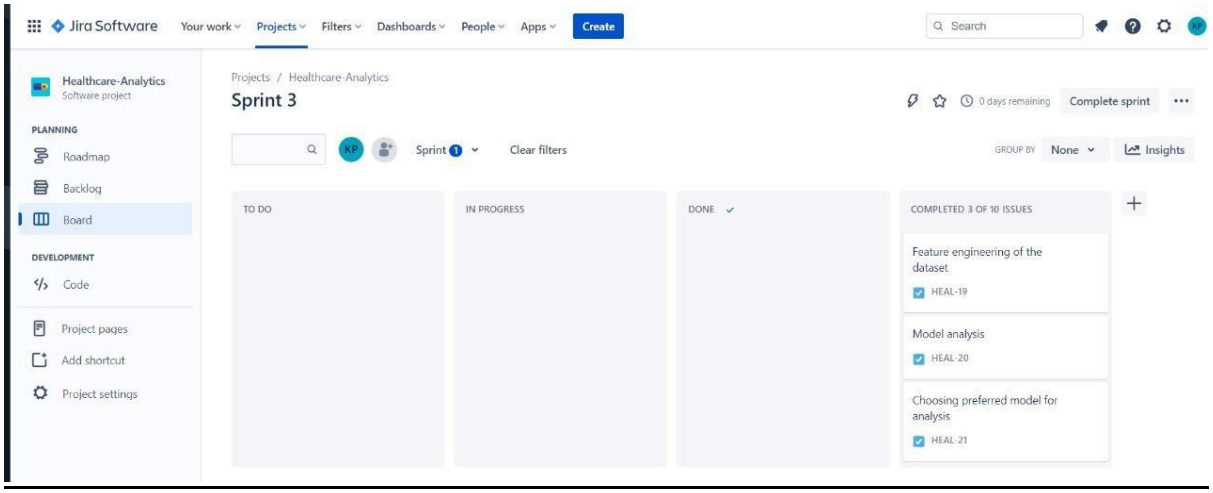
So, the model 2 we have decided to use XGBoosting. Boosting is a sequential technique that works on the principle of an ensemble model. It combines the set of weak learners and improves prediction accuracy. The final prediction score of the model is calculated by summing up each and individual score.

The 3rd model we have chosen to be neural network because in NN, increasing the number of neurons from each layer to the other layer, will increase the hypothetical space of the model and try to learn more patterns from the data. So, this can better predict the Length of stay of patients.

Jira Sprint 3 Tracking:


Jira Software
Your work
Projects
Filters
Dashboards
People
Apps
Create
Search
Healthcare-Analytics
Software project
PLANNING
Roadmap
Backlog
Board
DEVELOPMENT
Code
Project pages
Add shortcut
Project settings
Projects / Healthcare-Analytics
Sprint 3
0 days remaining
Complete sprint
Sprint 1
Clear filters
GROUP BY
None
Insights
TO DO 3 OF 6 ISSUES
IN PROGRESS
DONE
COMPLETED
Feature engineering of the dataset
HEAL-19
Model analysis
HEAL-20
Choosing preferred model for analysis
HEAL-21


Jira Software
Your work
Projects
Filters
Dashboards
People
Apps
Create
Search
Healthcare-Analytics
Software project
PLANNING
Roadmap
Backlog
Board
DEVELOPMENT
Code
Project pages
Add shortcut
Project settings
Projects / Healthcare-Analytics
Sprint 3
0 days remaining
Complete sprint
Sprint 1
Clear filters
GROUP BY
None
Insights
TO DO 1 OF 4 ISSUE
IN PROGRESS 1 OF 1 ISSUE
DONE 1 OF 1 ISSUE
COMPLETED
Choosing preferred model for analysis
HEAL-21
Model analysis
HEAL-20
Feature engineering of the dataset
HEAL-19


Jira Software
Your work
Projects
Filters
Dashboards
People
Apps
Create
Search
Healthcare-Analytics
Software project
PLANNING
Roadmap
Backlog
Board
DEVELOPMENT
Code
Project pages
Add shortcut
Project settings
Projects / Healthcare-Analytics
Sprint 3
0 days remaining
Complete sprint
Sprint 1
Clear filters
GROUP BY
None
Insights
TO DO
IN PROGRESS
DONE
COMPLETED 3 OF 10 ISSUES
Feature engineering of the dataset
HEAL-19
Model analysis
HEAL-20
Choosing preferred model for analysis
HEAL-21