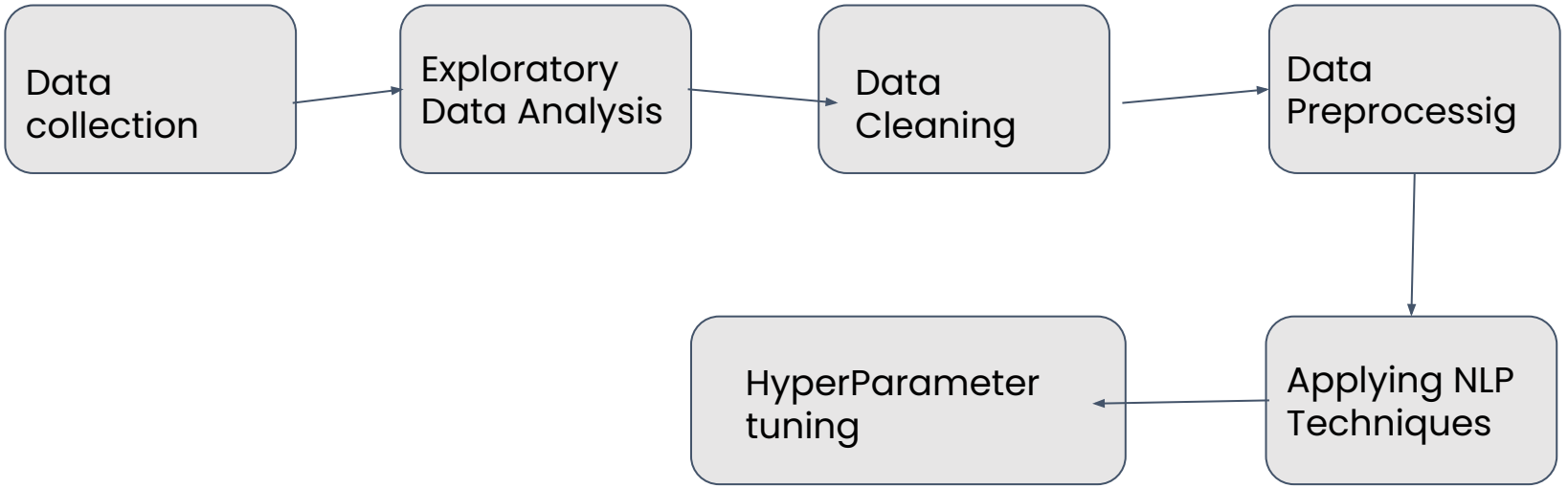
SAMSUNG

SmartMIDs

Outliers Team

Artificial Intelligence Course

Project Workflow



Exploratory Data Analysis (EDA)

After reading the data we find that many columns have null/duplicate and even out of range values

the EDA showed that our target ***condition*** is highly skewed towards the class of **Birth Control** indicating that the dataset is highly imbalanced. along with many Out of range values in the ***condition*** like “ user found this comment helpful”

and there is high value count for unlisted conditions (~1.3%)
the most challenging part is handling the similar data with different spelling. we had to fix that manually

EDA

since we have many object (string) columns we had to take a closer look to the data and we spotted a regex values along with many text noise especially in the ***review*** column

also we had to clean the html entities because the data was scraped from different websites from the web network

the EDA showed that after ***Date Parsing*** all years contributes equally to the dataset and all features almost fall equally into Different years making it useless columns

Data Preprocessing

- we handled manually over 100 unique condition and replace them with their appropriate spelling
- removing *regex, stop words, and sentence tokenization,after that we extracted the lemmatized text*
- ensuring that data is containing less zero values as some models tends to act randomly when plotting data with 0 values so our encoding covered the values [1,2,3]

Data Preprocessing

- There are features like NLTK Sentiment intensity score that give a sentiment about the review we get the score and then map it into 3 categories (positive, negative, neutral)

- We mapped both the ratings and useful count columns the same way to make it easier for the model to spot pattern and to decrease learning time by simplifying the features

- When handling the drug_name and condition (our target) columns we used LabelEncoder

Models used

- the modeling phase was the most challenging, we start searching for (2-5) promising models before start playing with the hyper-parameters tuning
- the models faced problems detecting the underlying pattern so we had made multiple drafts of our data and try different features
- our data contains more than 200k instance which made it hard for the models to train.and the nature of our data is hard to sample. so we run it in simpler models first

Models used

Model name	Sklearn model name	Hyperparameter
Random Forest	RandomForestClassifier()	Used with RandomizedSearchCV to find the best Max depth= 18 No. of estimators=255
Decision Tree	DecisionTreeClassifier	Criterion = entropy max depth = [10,20, 25,50] used the best parameters using gridsearchCV
KNN	KNeighborsClassifier	Using gridsearch with cross validation, the best n_neighbors was found to be n-neighbors=9

Results

Model _____ Scores	**Random Forest**	**KNN**	**Decision Tree**
Accuracy	0.645	0.68	0.71
Mean Absolute Error	89.34 degrees	-	-
F-1 score	-	-	0.72

In progress

Deployment:

We have 2 options for deployment,

1. LLMs

We start creating our own custom Chat GPT With our Data that can help diagnose your condition in a conversational way!

2. Website

We hope to finish a website that will have a drop-down menu with some categories to enter that will help the model find similar classes and return the best value