Medical Question Natural Language Processing

Elina Rankova

Data Scientist
Brooklyn, NY

NLP ANALYTICS

TOMORROW'S FUTURE

agenda

1. Business Objective

Define our primary goal of our NLP task

2. Data Understanding

Highlight the data specifics for our NLP task

3. Modeling

Showcase how NLP can provide insights to medical questions

4. Results

Summarize model performance

5. Next Steps

Advanced next steps to our NLP task

6. Contact

Contact information for data scientist

Business Objective

Healing Hands, a local medical practice wants to optimize patient Q&A process by predicting question type. Front desk needs to direct patients to the appropriate medical professional based on question type.

F1 Score focused analysis with attention to accuracy, recall, and precision for balanced predictions

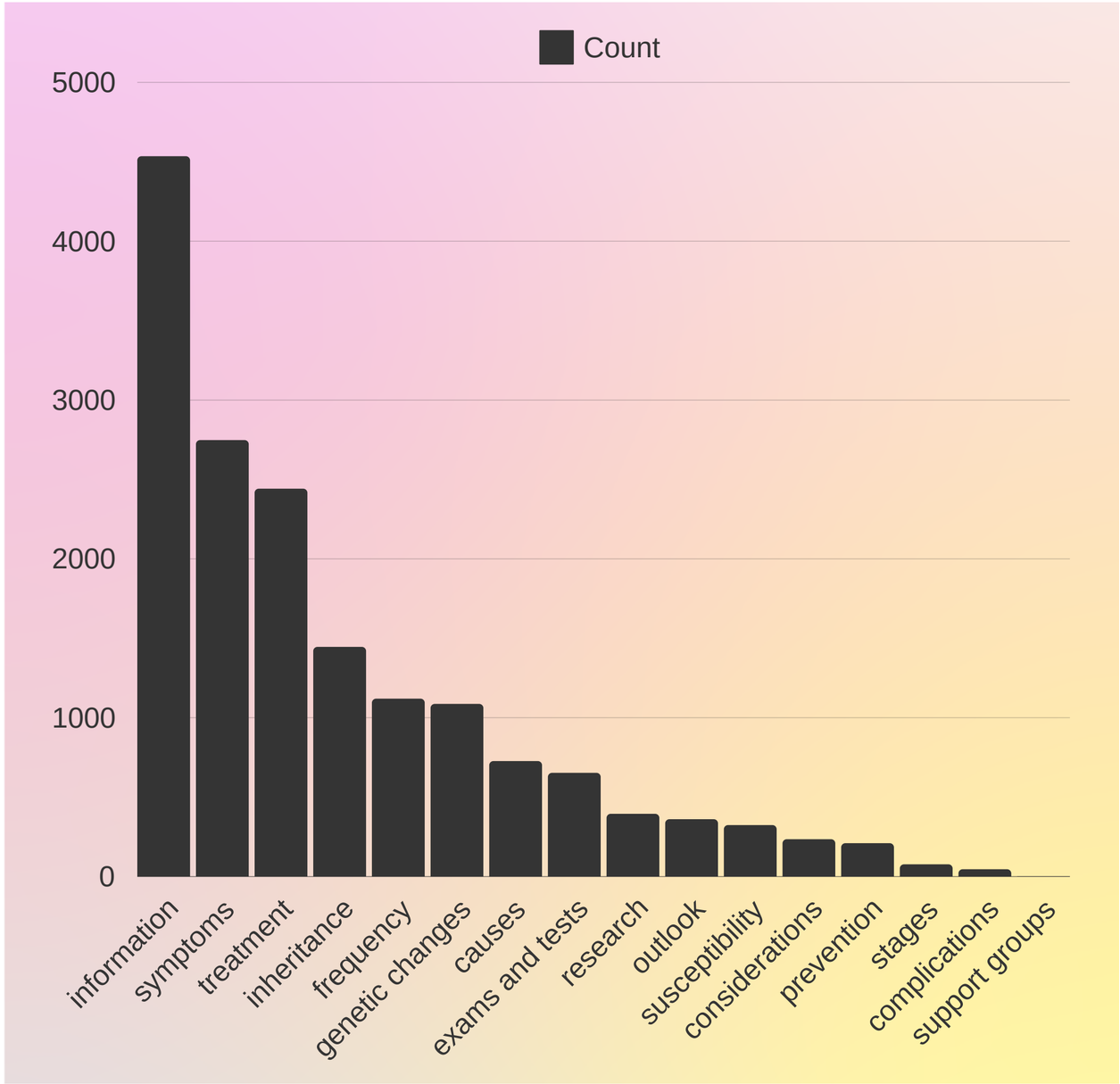
The Data

Sourced from Kaggle

Data Understanding

- No missingness
- Only 16407 total records
- 16 total labels
- Support Groups label contains *1 record*
- Apparent class imbalance
- Questions such as:
  - *exams and tests:* How to diagnose Lymphocytic Choriomeningitis (LCM) ?
  - *susceptibility:* Who is at risk for Parasites - Enterobiasis (also known as Pinworm Infection)?

Target Labels

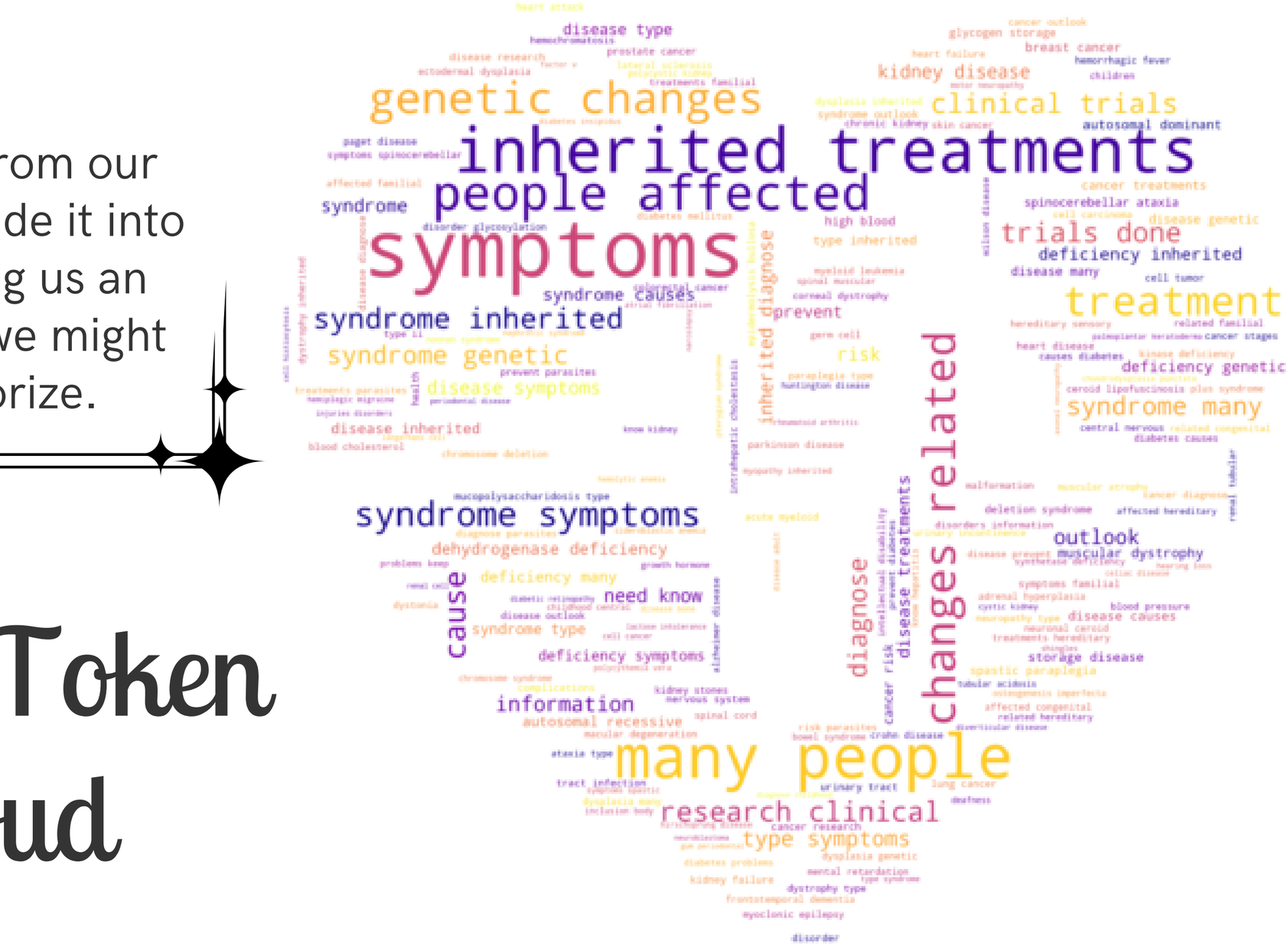


Question Type with Question Example

susceptibility	Who is at risk for Lymphocytic Choriomeningitis (LCM)? ?
symptoms	What are the symptoms of Lymphocytic Choriomeningitis (LCM) ?
exams and tests	How to diagnose Lymphocytic Choriomeningitis (LCM) ?
treatment	What are the treatments for Lymphocytic Choriomeningitis (LCM) ?
prevention	How to prevent Lymphocytic Choriomeningitis (LCM) ?
information	What is (are) Parasites - Cysticercosis ?
frequency	how common are these diseases for Marine Toxins ?
complications	are there complications from botulism?
causes	What causes Chronic Fatigue Syndrome (CFS) ?
research	what research is being done for Tuberculosis (TB) ?
outlook	What is the outlook for Striatonigral Degeneration ?
considerations	What to do for Lactose Intolerance ?
inheritance	Is Ovarian Epithelial, Fallopian Tube, and Primary Peritoneal Cancer inherited ?
stages	What are the stages of Ovarian Epithelial, Fallopian Tube, and Primary Peritoneal Cancer ?
genetic changes	What are the genetic changes related to Chronic Myelogenous Leukemia ?
support groups	Where to find support for people with Alcohol Use and Older Adults ?

Some of the words from our raw token analysis made it into the word cloud giving us an idea of the features we might see after we vectorize.

Normalized Token Wordcloud



The Data

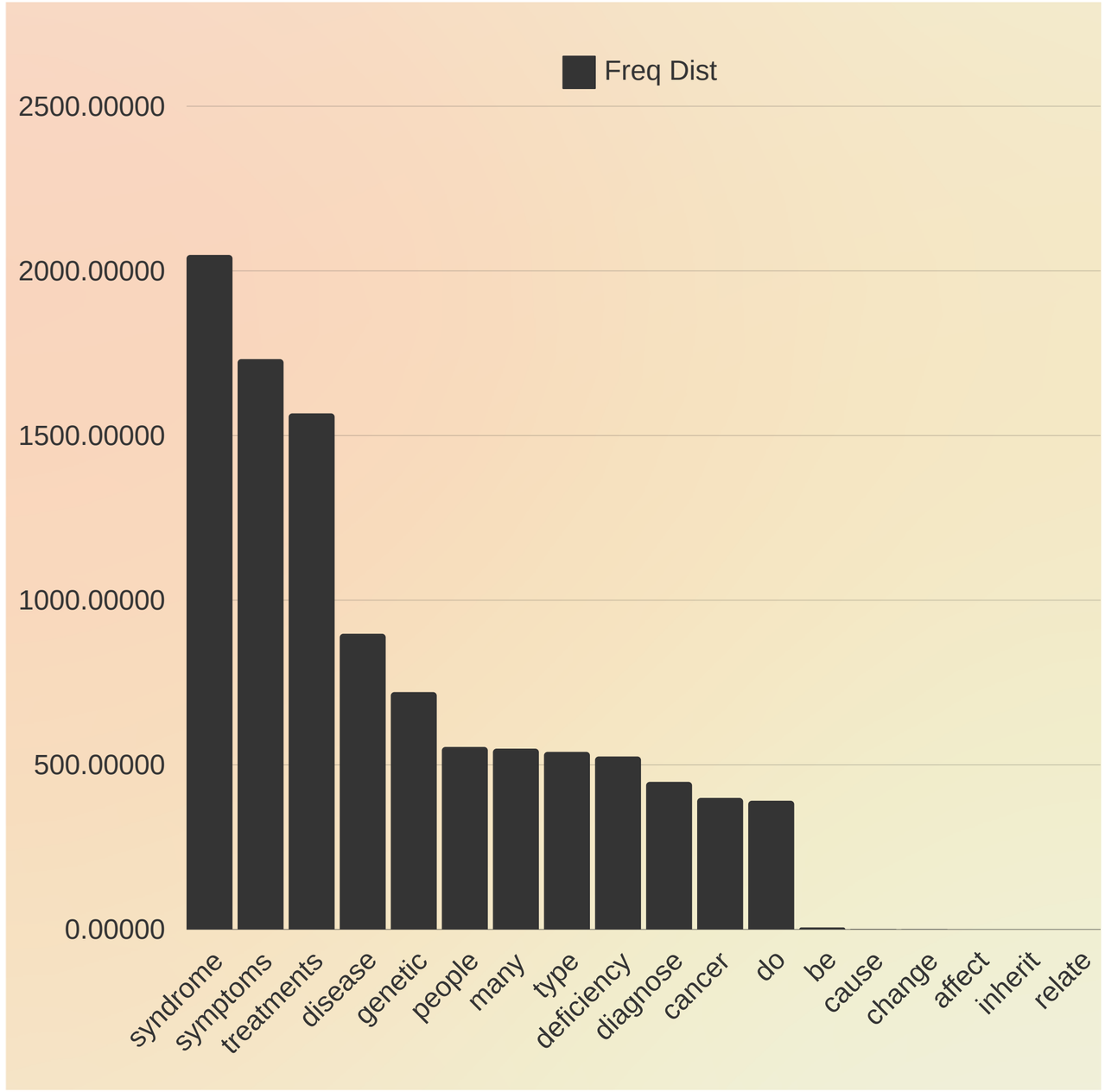
Dropped

Preparation

- Support Groups label contains *1 record*
  - Keeping our train/test stratified
- LabelEncoder on target prior to split
- Incorporate TextPreprocessor class into pipeline
- Use TfidVectorizer to keep class weights relevant
  - Adjust minimum and maximum document frequency
  - Left with 18 features
    - min_df = .03
    - max_df = .98

A lot of repeatative medical language !

Feature Space



Modeling

Naive Bayes

1 MultinomialNB

2 ComplementNB

RandomizedSearchCV

- Parameter tuning
- Cross validation
- Final scores on **best_estimator_**

Modeling

leading f1 score

Naive Bayes

1 MultinomialNB

2 ~~ComplementNB~~

RandomizedSearchCV

- Parameter tuning
- Cross validation
- Final scores on **best_estimator_**

Model Name	Train Score	Test Score
MultiNB	0.855172	0.853105
ComplementNB	0.733154	0.742919

Modeling

1

MultinomialNB

3

DecisionTree

4

RandomForest

RandomizedSearchCV

- Parameter tuning
- Cross validation
- Final scores on **best_estimator_**

Modeling

leading f1 score

1

~~MultinomialNB~~

3

DecisionTree

4

RandomForest

RandomizedSearchCV

- Parameter tuning
- Cross validation
- Final scores on **best_estimator_**

Model Name	Train Score	Test Score
Decision	0.873223	0.873457
MultiNB	0.855172	0.853105

Modeling

3

DecisionTree

4

Random Forest

5

GradientBoost

RandomizedSearchCV

- Parameter tuning
- Cross validation
- Final scores on **best_estimator_**

Modeling

leading f1 score

3

DecisionTree

4

~~Random Forest~~

5

GradientBoost

RandomizedSearchCV

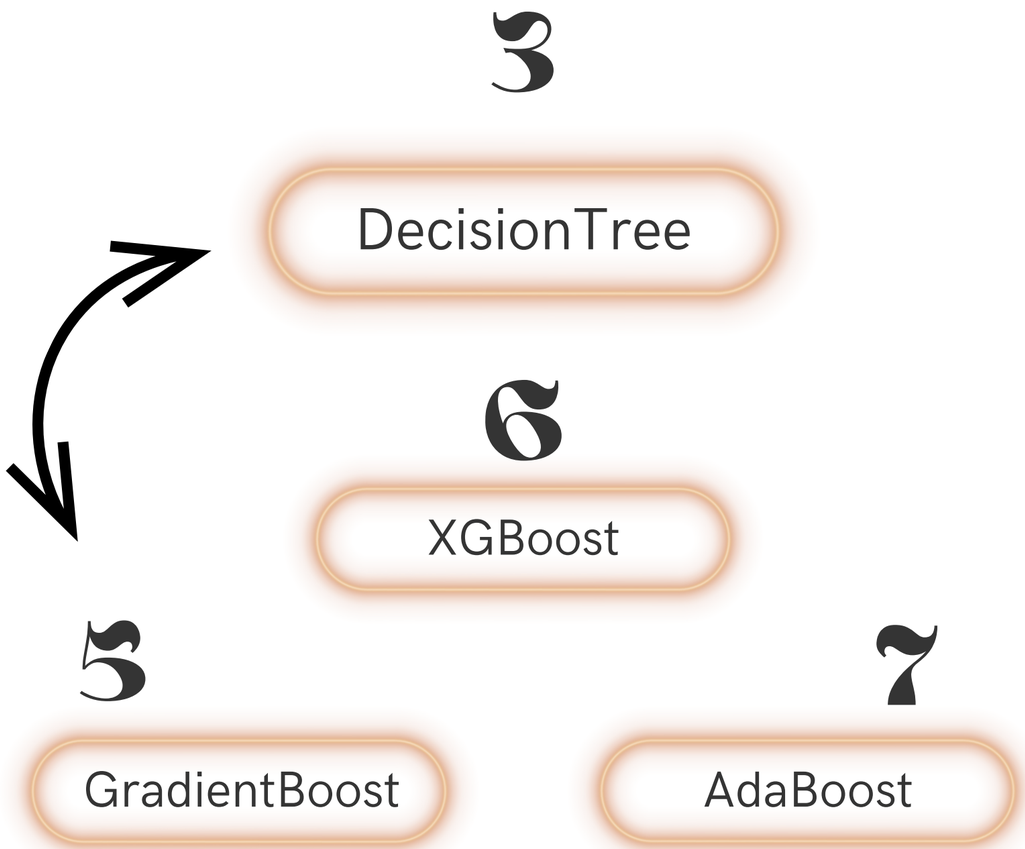
- Parameter tuning
- Cross validation
- Final scores on **best_estimator_**

Model Name	Train Score	Test Score
Decision	0.873223	0.873457
MultiNB	0.855172	0.853105
Forest	0.749046	0.757757

Modeling

RandomizedSearchCV

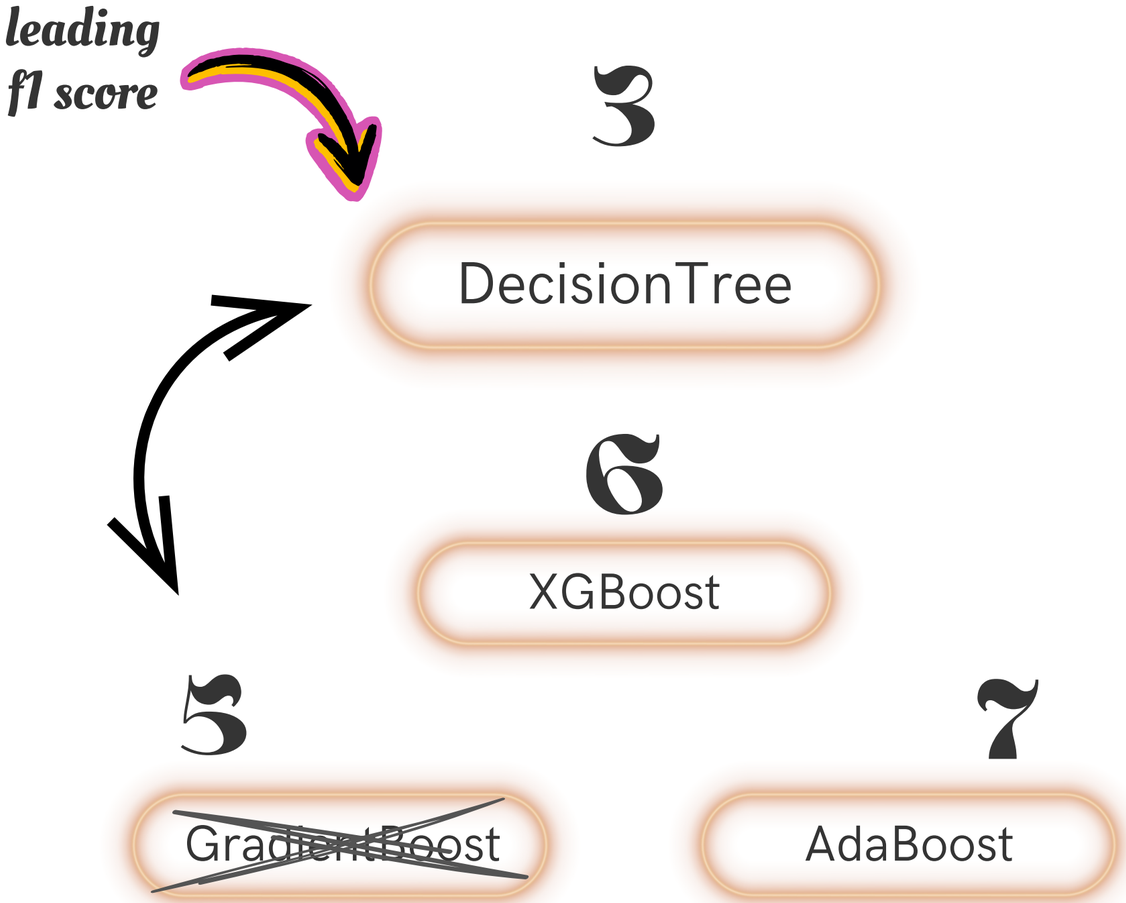
- Parameter tuning
- Cross validation
- Final scores on **best_estimator_**



Modeling

RandomizedSearchCV

- Parameter tuning
- Cross validation
- Final scores on **best_estimator_**

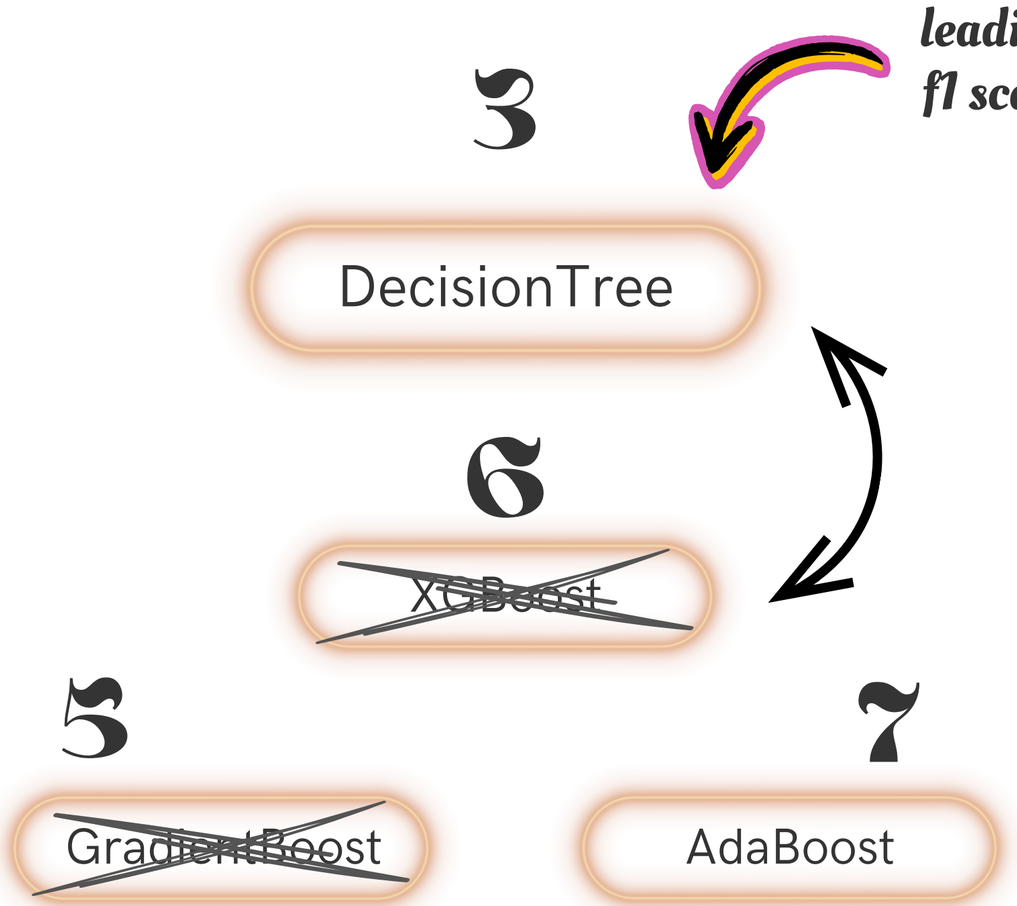


Model Name	Train Score	Test Score
Decision	0.873223	0.873457
AdaBoost	0.872402	0.873213
GradientBoost	0.872656	0.871478
XGBoost	0.872656	0.871117

Modeling

RandomizedSearchCV

- Parameter tuning
- Cross validation
- Final scores on **best_estimator_**



Model Name	Train Score	Test Score
Decision	0.873223	0.873457
AdaBoost	0.872402	0.873213
GradientBoost	0.872656	0.871478
XGBoost	0.872656	0.871117

Results

On Holdout Data

DecisionTree	AdaBoost
F1: .87	F1: .87
Accuracy: 90	Accuracy: 91
Precision: 93	Precision: 93
Recall: 90	Recall: 91

Results

On Holdout Data

AdaBoost

Base Metrics	*Tuned Metrics*
F1: .77	F1: .87
Accuracy: 82	Accuracy: 91
Precision: 81	Precision: 93
Recall: 82	Recall: 91

Looking Ahead

More Data

Collecting more data to satisfy all classes will help create wholistic predictions

Predicting Answer on Questions

Predicting common answer to common questions helps create a better experience for medical professionals and their patients

Chat Bot!

Medical offices can implement internal chatbot to assist with wholistic next steps efficiently

Thank you!

elinarankova@gmail.com
Github | Blog | LinkedIn