

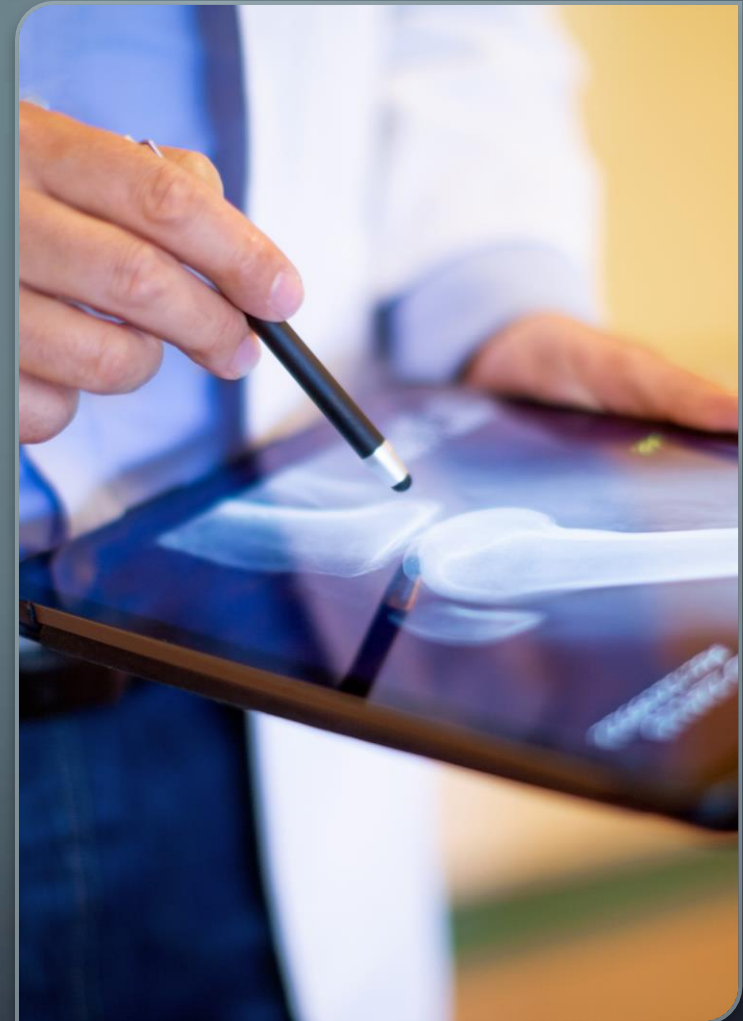
A decorative graphic on the left side of the slide, consisting of a network of white lines and small circles on a dark blue background, resembling a circuit board or a neural network diagram.

PREDICTING GOUT IN ER PATIENTS

CLASSIFY PATIENTS 'CHIEF COMPLAINT' AS 'GOUT' OR 'NOT GOUT'

WHY? should we build a model that predicts the diagnosis of gout?

- 10 out of 100 patients were mislabeled by ER physicians
- Studies show a cause of misdiagnosis result from vocabulary differences between patient and physician. due to vocabulary
- Vocabulary patterns can be recognized by ML models and aid in predicting 'Gout' based on the patients' complaint, bridging the gap that causes these errors.





DATA SOURCE

- Data was obtained from the MIMIC III database, Beth Israel Deaconess Medical Center:

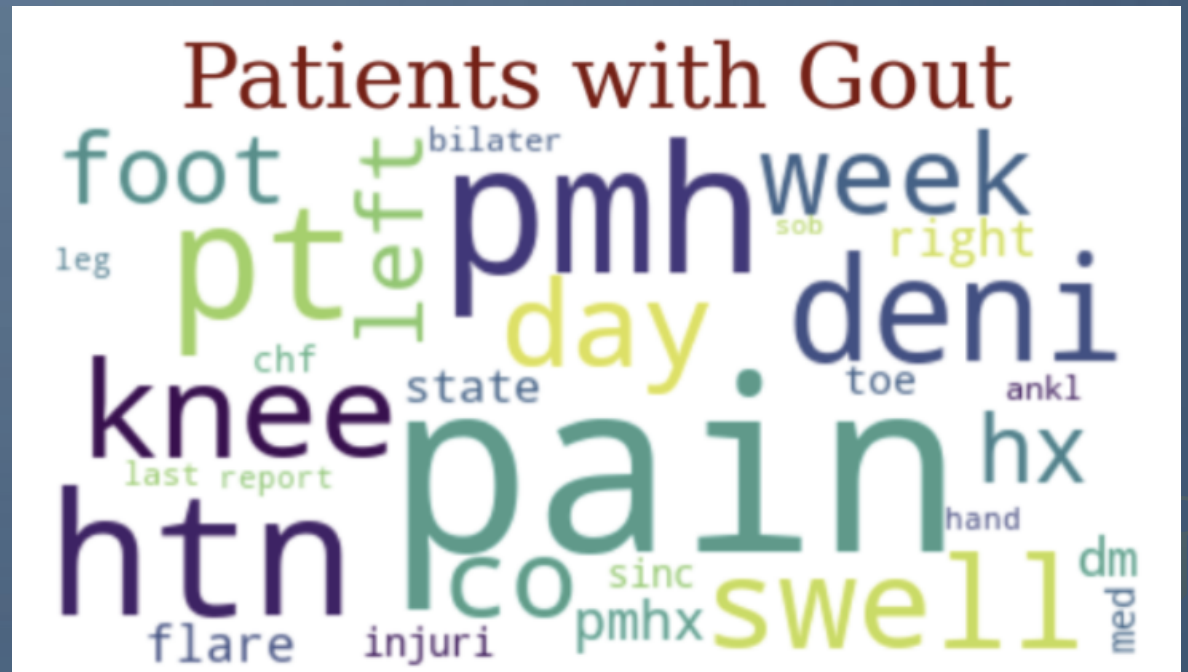
<https://physionet.org/content/emer-complaint-gout/1>.

NLP DATA PRE-PROCESSING

Pre-processed Data

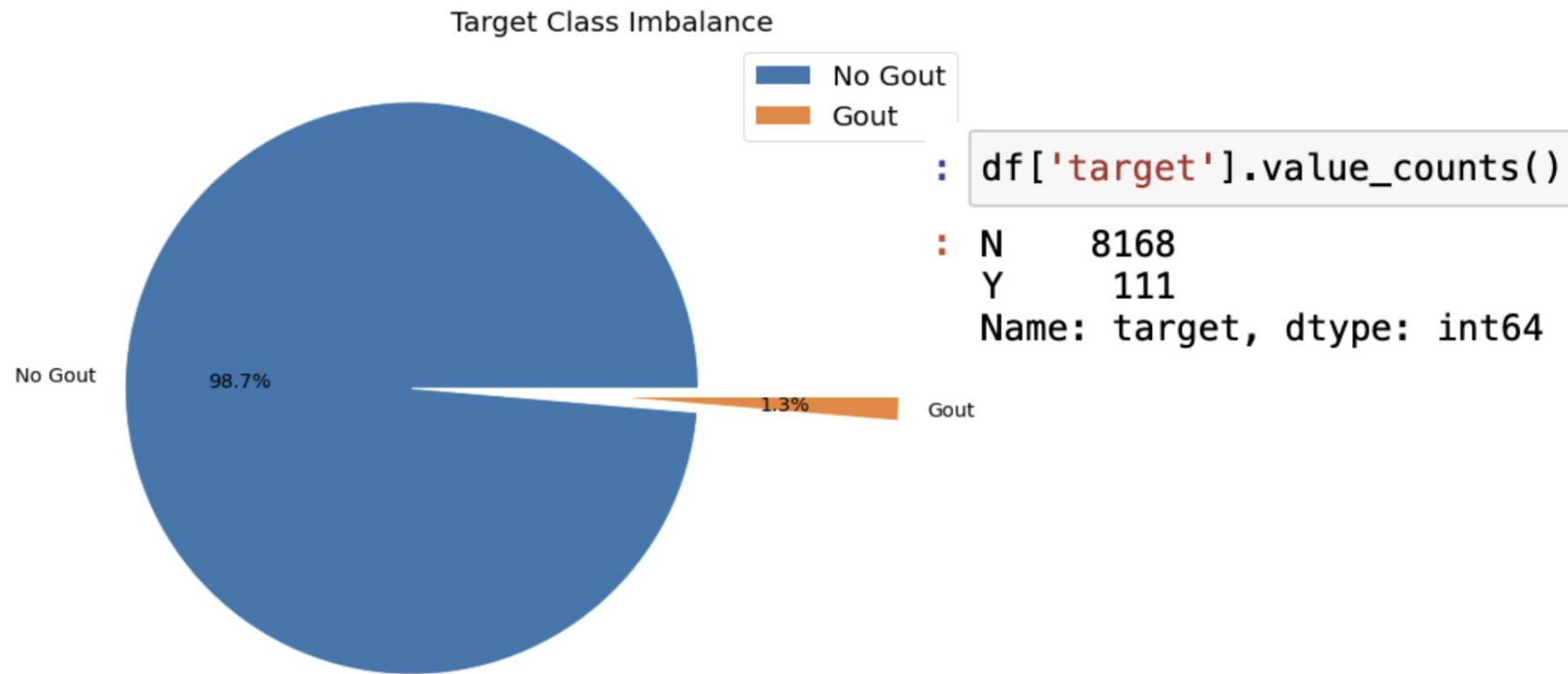
	corpus	target
0	"been feeling bad" last 2 weeks & switched BP ...	N
1	"can't walk", reports onset at 0830 am. orient...	Y
2	"dehydration" Chest hurts, hips hurt, cramps P...	Y

WORD CLOUD



IMBALANCED DATA

Class Distribution is significantly imbalanced.



MODELING

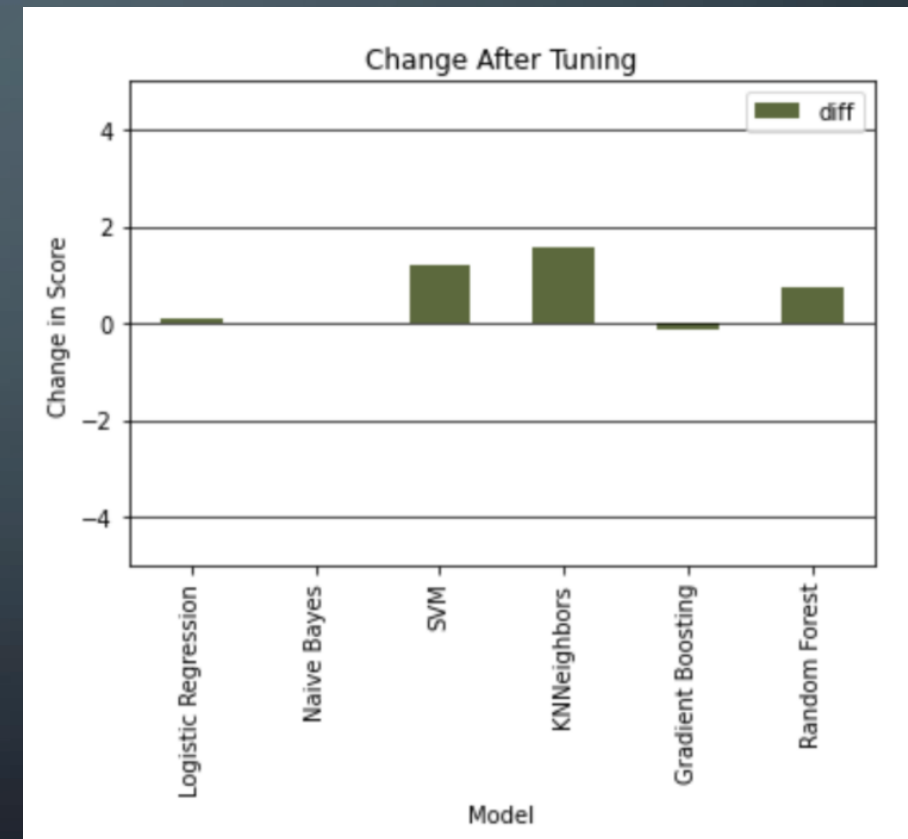
- CLASSIFICATION MODELS USING SCI-KIT LEARN
 - * Logistic Regression
 - * Naive Bayes
 - * Support Vector Machine
 - * KNneighbors
 - * Gradient Boost
 - * Random Forest
- BERT PRE-Trained Models



ROC AUC SCORES before and after hyperparameter tuning

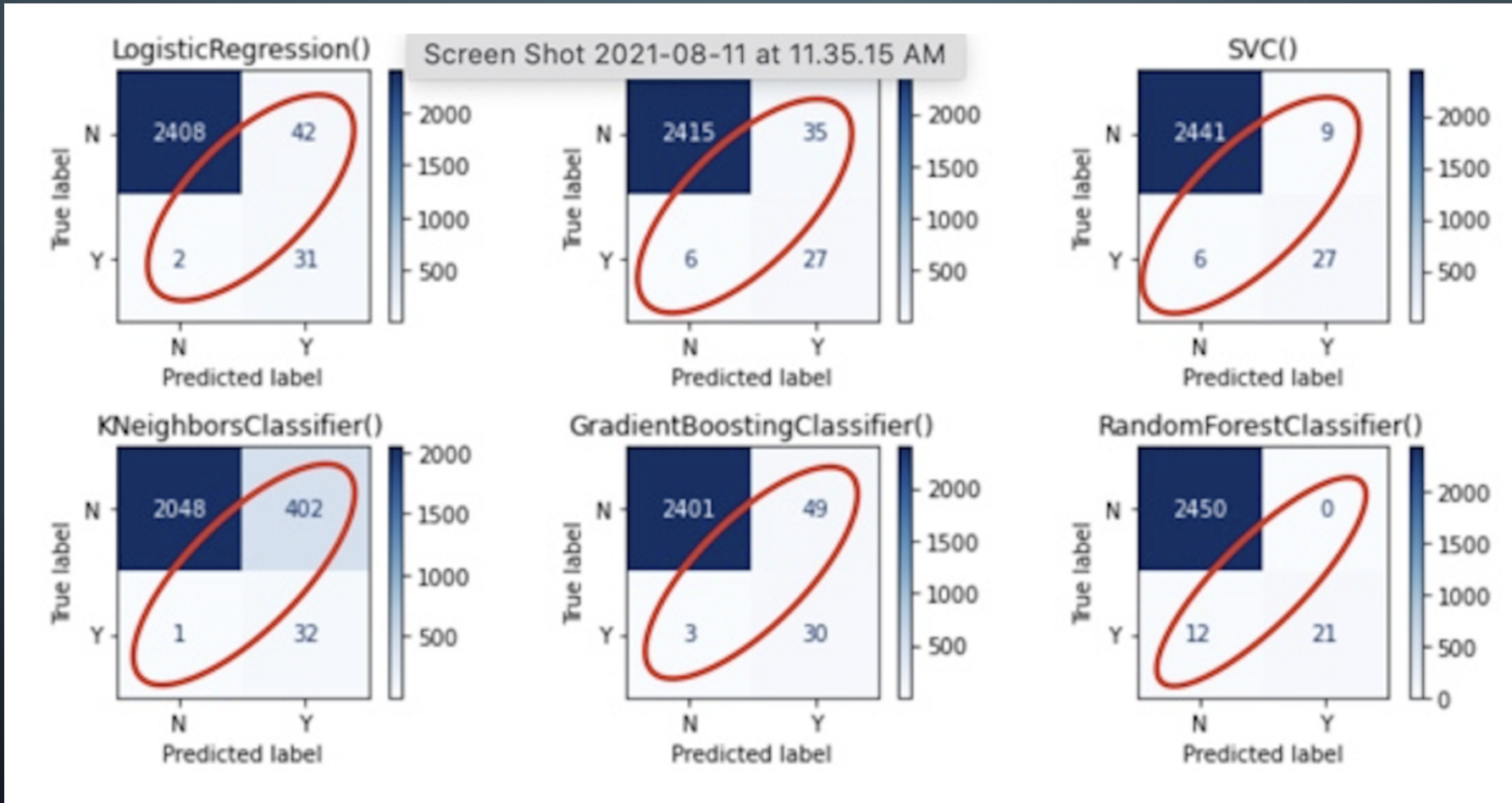
model	tuned score	untuned score
Logistic Regression	93.897341	93.794682
Naive Bayes	92.226345	92.226345
SVM	93.209647	91.996289
KNNighbors	86.552257	84.988868
Gradient Boosting	91.137910	91.289425
Random Forest	91.288806	90.529375

Top 3: Logistic Regression 93.90%
Naive Bayes 92.23%
SVM 93.21%



CONFUSION MATRIX RESULTS

NOTE THE BALANCE BETWEEN FALSE NEGATIVES AND OVERALL ERRORS



ROC AUC BERT PRE-TRAINED ACCURACY RESULTS

PREDICTION ACCURACY = 98.67%

Epoch: 0%| | 0/2 [00:00<?, ?it/s]

Train loss: 0.026148697721712938

Epoch: 50%|██████ | 1/2 [1:05:15<1:05:15, 3915.78s/it]

Validation Accuracy: 0.9867788461538461

Train loss: 0.026181237024088717

Epoch: 100%|██████████ | 2/2 [5:36:35<00:00, 10097.72s/it]

Validation Accuracy: 0.9867788461538461

CONSLUSION

FIRST CHOICE: BERT PRE-TRAINED

SECOND CHOICE: SVC

- BERT Pre-Trained is based on large data sets and is preferable. The nature of BERT however requires API connectivity, opening potential security concerns.
- For In-House modeling sci-kit learn's SVC model reduced the false negatives without compromising overall error rate.

FURTHER RESEARCH

1

Explore clinical
BERT pre-
trained
models

2

Improve
model with
additional
labeled
training data

3

Include a
statistical
probability in
results