

Unlocking Medical Assistance Through Machine Learning

Project 4

Authors: Leo Wheeler, Jacob Brown, Connor Beaton, Liliana Guevara-Guerrero





Disclaimer

This chatbot does *not* give medical advice, but rather suggestions on how soon you should visit a physician. Please see a professional if you are having serious health concerns.

Identify Data Sources

	A	B
1	text	label
2	I have been having migraines and headaches. I can't sleep. My whole body is shaking and shivering. I feel dizzy sometime	308
3	I have asthma and I get wheezing and breathing problems. I also have fevers, headaches, and I feel tired all the time.	35
4	Signs and symptoms of primary ovarian insufficiency are similar to those of menopause or estrogen deficiency. They inclu	798
5	cough,high_fever,breathlessness,family_history,mucoid_sputum	149
6	chills,vomiting,high_fever,sweating,headache,nausea,diarrhoea,muscle_pain	596
7	Posterior cortical atrophy symptoms vary among people. Symptoms also can vary over time. They tend to gradually get wo	785
8	I've been having back pain, a cough, and numbness in my arms and legs. My neck hurts too, and I've been feeling dizzy an	186
9	yellowish_skin,nausea,loss_of_appetite,yellowing_of_eyes,family_history	466
10	joint_pain,neck_pain,knee_pain,hip_joint_pain,swelling_joints,painful_walking	700

7033	skin_rash,joint_pain,skin_peeling,small_dents_in_nails,inflammatory_nails	822
7034	skin_rash,blackheads,scurring	7
7035	breathlessness,sweating,chest_pain	447
7036	chills,vomiting,fatigue,high_fever,headache,nausea,constipation,diarrhoea,toxic_look_(typhos),belly_pain	1026
7037	I have some veins on my legs that are swollen and protruding from my skin. They are very noticeable and make me self-co	1047
7038	Mosquito bites often happen on parts of the body that aren't covered by clothing. Symptoms include: An itchy, inflamed bu	649
7039	vomiting,headache,weakness_of_one_body_side,altered_sensorium	718
7040	itching,skin_rash,stomach_pain,burning_micturition,spotting_urination	308
7041	muscle_wasting,high_fever,extra_marital_contacts	27
7042	The signs and symptoms of gastritis include: Gnawing or burning ache or pain (indigestion) in your upper abdomen that i	401
7043	vomiting,indigestion,loss_of_appetite,abdominal_pain,internal_itching	735
7044	I've been having a lot of trouble breathing, and I feel like I'm constantly nauseous. I also have mild belly pain. It's been rea	1026

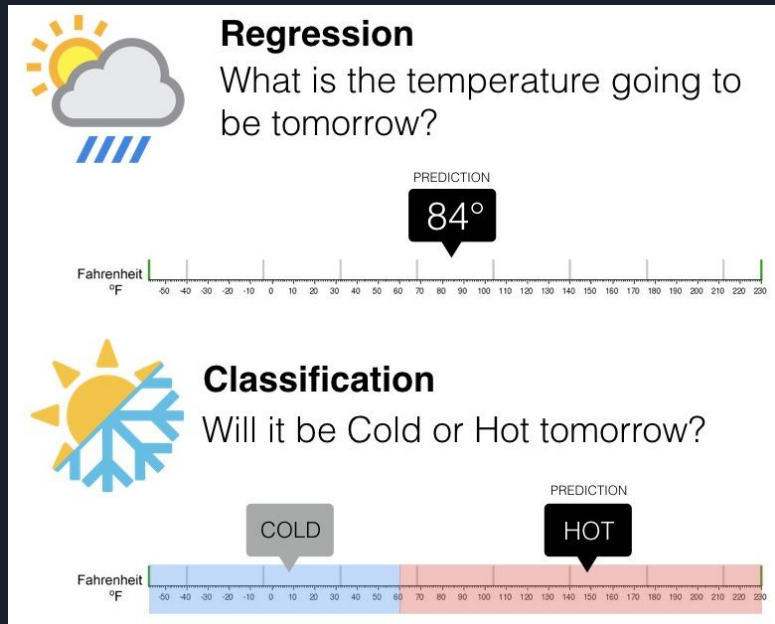
Source:

huggingface.co/datasets/dux-tecblic/symptom-disease-dataset

Regression or Classification?

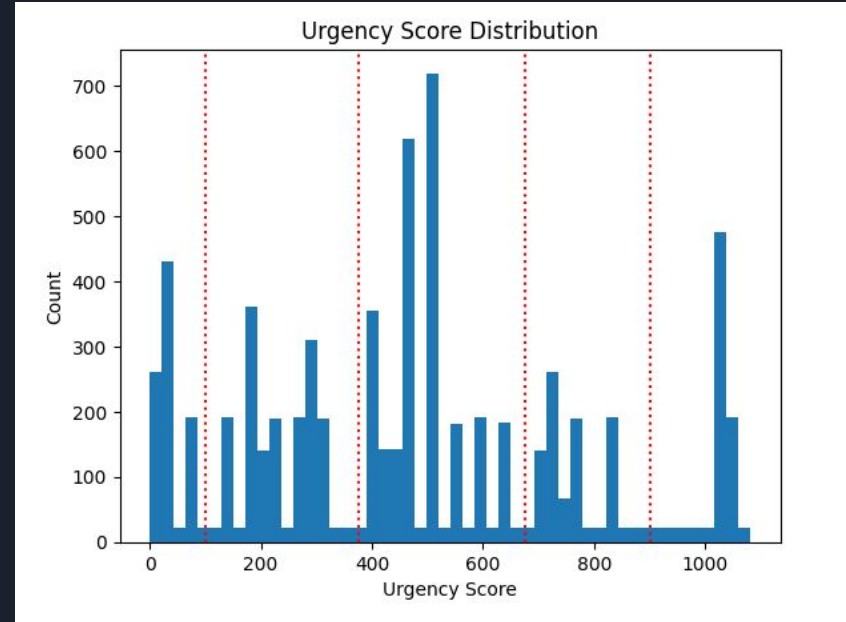
Regression: Data with Numeric Target

Classification: Data with Categorical Target



Regression or Classification?

Data splits nicely into 5 categories,
we will approach as a classification
problem



Regression or Classification?

```
12 def binner(scores):
13     output = []
14     for score in scores:
15         if score > 900:
16             output.append(1)
17         elif score > 675:
18             output.append(2)
19         elif score > 375:
20             output.append(3)
21         elif score > 100:
22             output.append(4)
23         else:
24             output.append(5)
25     return output
26
27
28 # Binning the dataframes and renaming them
29 data = data.rename(columns = {"text": "Text", "label" : "Label"})
30
31 data["Label"] = binner(data['Label'])
```

	A	B
1	text	label
2	I have been having migr	308
3	I have asthma and I get	35
4	Signs and symptoms of	798
5	cough,high_fever,breat	149
6	chills,vomiting,high_fev	596
7	Posterior cortical atrop	785
8	I've been having back p	186
9	yellowish_skin,nausea,	466
10	joint_pain,neck_pain,k	700
11	An unusual sensation (971
12	vomiting,yellowish_skin	33
13	vomiting,fatigue,anxiet	504
14	fatigue,mood_swings,w	502
15	joint_pain,vomiting,yell	464



	A	B	C
1	text		target
2	0 I have been having migr		4
3	1 I have asthma and I get		5
4	2 Signs and symptoms of		2
5	3 cough,high_fever,breat		5
6	4 chills,vomiting,high_fev		3
7	5 Posterior cortical atrop		2
8	6 I've been having back p		5
9	7 yellowish_skin,nausea,		3
10	8 joint_pain,neck_pain,k		2
11	9 An unusual sensation (1
12	10 vomiting,yellowish_skin		5
13	11 vomiting,fatigue,anxiet		3
14	12 fatigue,mood_swings,w		3
15	13 joint_pain,vomiting,yel		3



Using Machine Learning Models

Machine Learning models are fantastic for NLP but can't directly take text input. Some kind of processing is required:

Vectorizer	Pros	Cons
One Hot Encoding	Simple	Bloated dataframe and lack of value encoded into data
Count Vectorizer	Easy to understand	Less applicable or appropriate with nonlinear or complex text forms
TFIDF (Text Frequency Inverse Document Frequency) Vectorizer	High computational power, great for complex text	Computationally complex, much harder to understand



Machine Learning Models

- Ridge
- K Nearest Neighbors
- Naive Bayes
- Tensorflow Neural Network
- Decision Tree
- **Random Forest**

Ending Test Accuracy on Random Forest: 0.8847



Engineering Bottleneck

For our machine learning engineering models we had:

- Accuracy means of approximately 86% across all models
- Standard deviation of approximately 2% across all models

If we were to drill into hyperparameter space would we be able to breach 90%?

- Engineering Bottleneck
 - Possible fixes include:
 - Adding data, including possibly more breadth and depth to each label type
 - Doing advanced ETL on the data
 - Changing methodology (LLM instead of Vectorizer)

Model Tuning

Tuning parameter: n_estimators

Max_depth: little to no effect if
n_estimators is changing

```
def rfTuner(X_train, y_train, X_test, y_test, epochs=100):  
    print('Initializing Tuner...')  
    output = {} # saving info on models  
    best_acc = 0 # tracking highest accuracy  
    start = time.time() # timing  
    for i in range(epochs): # user-defined loop length  
        model = RandomForestClassifier(n_estimators=i+1) # changing estimators  
        model.fit(X_train, y_train)  
        preds = model.predict(X_test)  
        acc = accuracy_score(y_test, preds)  
        output[i] = {'estimators':i+1, # output appending  
                    'acc':acc}  
        print(f'model complete: {i}', end='\r')  
        if acc > best_acc: # updating highest accuracy  
            print(f'New best Accuracy! Model {i}')  
            print(f'{round(acc*100,2)}%')  
            best_acc = acc  
    print(f'Total runtime: {(time.time() - start)/60} minutes')  
    return output
```

Highest Scoring Models:

Random Forest: 199 estimators, 0.89% accuracy

XGBoost: 200 estimators, 0.88% accuracy

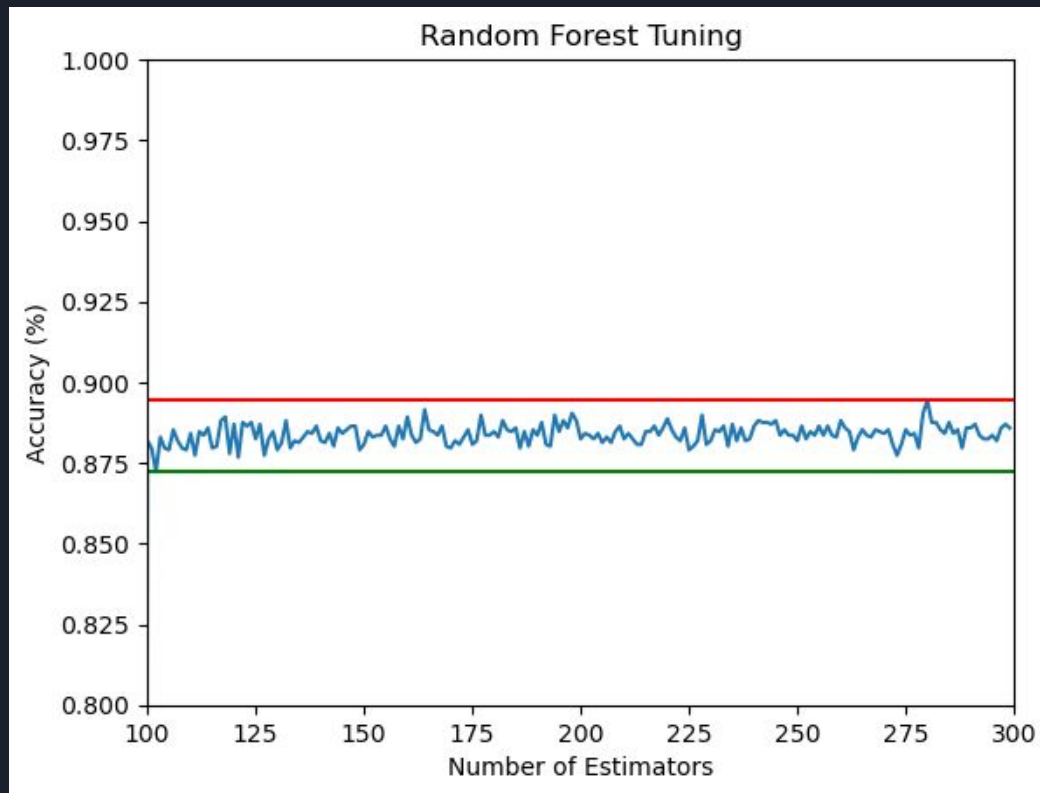
Model Tuning

All tuning values fall between 87.2% and 89.4%

Slight uptick on right side vs left side

200 seems consistent, so we landed there

Note: this graph is from a different test than the previous slide's tuner



Model Saving - Pickle

Pickle was used to save the models (.pkl files)

```
def rfPacker(n_estimators, file):  
    model = RandomForestClassifier(n_estimators=n_estimators)  
    model.fit(X_train, y_train)  
    with open(file, 'wb') as file:  
        pickle.dump(model, file)
```

```
rfPacker(199, 'trained_mlmodels/rfHealthBot.pkl')
```

Models can be read back in with Pickle as well

```
loaded_rf = pickle.load(open('trained_mlmodels/rfHealthBot.pkl', 'rb'))  
# loaded_xgb = pickle.load(open('trained_mlmodels/xgbHealthBot.pkl', 'rb'))  
result1 = loaded_rf.predict(X_test)  
# result2 = loaded_xgb.predict(X_test)  
print("Saved Model Scores:")  
print(f'RandomForest: {accuracy_score(y_test, result1)}')  
# ytest_mod = [item-1 for item in y_test] # XGBoost likes 0-4 instead of 1-5  
# print(f'XGBoost: {accuracy_score(ytest_mod, result2)}')
```

Saved Model Scores:

RandomForest: 0.8847245883021011

Model Saving - Pickle - Important Note

```
def rfPacker(n_estimators, file):  
    model = RandomForestClassifier(n_estimators=n_estimators)  
    model.fit(X_train, y_train)  
    with open(file, 'wb') as file:  
        pickle.dump(model, file)
```

```
rfPacker(199, 'trained_mlmodels/rfHealthBot.pkl')
```

Notice the "b" in the files...

```
loaded_rf = pickle.load(open('trained_mlmodels/rfHealthBot.pkl', 'rb'))  
# loaded_xgb = pickle.load(open('trained_mlmodels/xgbHealthBot.pkl', 'rb'))  
result1 = loaded_rf.predict(X_test)  
# result2 = loaded_xgb.predict(X_test)  
print("Saved Model Scores:")  
print(f'RandomForest: {accuracy_score(y_test, result1)}')  
# ytest_mod = [item-1 for item in y_test] # XGBoost likes 0-4 instead of 1-5  
# print(f'XGBoost: {accuracy_score(ytest_mod, result2)}')
```

Binary File encoding!

```
Saved Model Scores:  
RandomForest: 0.8847245883021011
```

Flask Backend

Read in ML model and
Vectorizer

Use the model to make a
prediction based on user
input

Using the chatbot app,
return text to specify
severity

```
1 from flask import Flask, render_template, request, jsonify
2 import pickle
3 import numpy as np
4 from utils.chatbot import chatbot
5
6 # Initialize Flask
7 app = Flask(__name__)
8
9 # Load pre-trained RandomForest model
10 with open("mlmodels/trained_mlmodels/rfHealthBot.pkl", "rb") as model_file:
11     model = pickle.load(model_file)
12
13 # Load pre-trained TF-IDF vectorizer
14 with open("mlmodels/trained_mlmodels/vectorizer.pkl", "rb") as vec_file:
15     vectorizer = pickle.load(vec_file)
16
17 @app.route("/")
18 def index():
19     return render_template("index.html")
20
21 @app.route("/predict", methods=["POST"])
22 def predict():
23     data = request.get_json()
24     user_input = data.get("text", "")
25
26     if not user_input:
27         return jsonify({"error": "No input provided"}), 400
28
29     # Vectorize and predict
30     vector = vectorizer.transform([user_input]).toarray()
31     prediction = chatbot(int(model.predict(vector)[0]))
32
33     return jsonify({"prediction": prediction})
34
35 if __name__ == "__main__":
36     app.run(debug=True)
```

End Results

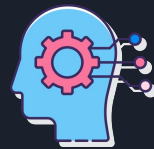
With the incorporation of Flask, we developed a “Symptoms Checker” chatbot that functions through a web page. The chatbot predicts the type of medical assistance needed when symptoms are entered by the user.

The symptoms are grouped by severity as follows:

- **Level 5 – Safe** 😊
- **Level 4 – Mild** 💤
- **Level 3 – Moderate** 🩺
- **Level 2 – Immediate Care** 🏥
- **Level 1 – Emergency** 🚑



Symptom Checker Demo



Let's walk through how the chatbot works!

A screenshot of a web browser window. The address bar shows '127.0.0.1:5000'. The page title is 'Symptom Severity Predictor'. The main content area is white with a light gray grid background. In the center is a white rounded rectangle containing the title 'Symptom Severity Checker' in bold. Below the title is a text input field with the placeholder text 'Describe your symptoms here...'. At the bottom right of the rounded rectangle is a blue button with the text 'Run Module' in white.

Thank You!