MACHINE LEARNING CHALLANGE

About the contest:

Thyroid is a gland in our body responsible for producing thyroid hormone, which is essential for regulating breathing, body weight, heart rate, and muscle strength. Any irregularity in the production of this hormone can be fatal. There are four disorders associated with the thyroid gland:

- Hyperthyroid
- Hypothyroid
- Euthyroid-sick
- Euthyroid

However, the human body reacts differently to the above irregularities resulting in diversified symptoms, and the disease goes undiagnosed in many cases. This challenge aims to train a machine learning model to predict whether a patient has a thyroid-related disorder or not.

Dataset:

We have provided a *thyroid.txt* file for this challenge which you can download from –

	age	sex	on thyroxine	on antithyroid medication	sick	pregnant	thyroid surgery	I131 treatment	lithium	goitre	tumor	hypopituitary	psych	TSH	T3	TT4	T4U	FTI	Thyroid
0	41	F	f	f	f	f	f	f	f	f	f	f	f	1.3	2.5	125	1.14	109	P
1	23	F	f	f	f	f	f	f	f	f	f	f	f	4.1	2	102	?	?	P
2	46	M	f	f	f	f	f	f	f	f	f	f	f	0.98	?	109	0.91	120	P
3	70	F	t	f	f	f	f	f	f	f	f	f	f	0.16	1.9	175	?	?	P
4	70	F	f	f	f	f	f	f	f	f	f	f	f	0.72	1.2	61	0.87	70	P

Figure1: Glimpse of the dataset

This is a data of 3772 patients and the target parameter to predict here is *“Thyroid”* where *“P”* and *“N”* refers to the Thyroid positive and negative patients respectively.

Submission Guidelines:

For completing this challenge, the following **two files** needs to be submitted:

- A Jupyter notebook file (i.e., .ipynb file) used for code implementation. Name the file as ***geekschallange_Firstname_Lastname.ipynb***. For example, if the participant’s name is John Smith, then the submission file will be named as ***geekschallange_John_Smith.ipynb***

- The trained machine learning model used for the data challenges as a pickle file [follow the appendix for better understanding]. Name the file as ***geekschallange_Firstname_Lastname.pickle***. For example, if the participant’s name is John Smith, then the submission file will be named as ***geekschallange_John_Smith.pickle***

Make sure to follow the following guidelines for naming the datafile. Do not zip these files into one zip archive, submit two independent files.

We will only accept Jupyter Notebook submissions. Please make sure that the code is properly commented, and the jupyter notebook should include the reporting about the following:

- A brief description of the data set and the problem.
- Details of the data exploration and pre-processing steps.
- Analysis of the exploration plots and insights.
- An explanation of the model selection process and the chosen model.
- Final model evaluation metrics and a discussion of model performance.

Evaluation Criteria:

Submissions will be evaluated based on the following criteria:

- Data Pre-processing techniques
- Data exploration techniques and analysis
- Model selection and performance
- Final model evaluation and discussion
- Coding architecture and commenting

Resources:

- How to properly comment a Python code: https://www.geeksforgeeks.org/pep-8-coding-style-guide-python/

Appendix:

What is a pickle file?

A pickle file is a serialized object that can be used to store the state of a Python object, such as a machine learning model. It is used in machine learning to save a trained model, along with any pre-processing steps that have been applied to the data, so that it can be reloaded later and used to make predictions on new data.

When you train a machine learning model, the resulting model is just a set of weights and biases that have been learned from the training data. To use the model later, you need to save these weights and biases in a file so that they can be loaded into memory and used to make predictions on new data.

Pickle files provide a way to save Python objects to disk, so that they can be reloaded later. This is useful in machine learning because it allows you to save a trained model, along with any pre-processing steps that have been applied to the data, in a single file. This makes it easy to share and distribute trained models, and to use them in production environments.

How to save the machine learning model in a pickle file

Let's suppose you have trained a logistic regression machine learning model on the dataset as show in the figure 2 below:

```
from sklearn.linear_model import LogisticRegression
from sklearn.datasets import load_iris

# Load the iris dataset
iris = load_iris()

# Train a logistic regression model on the dataset
X, y = iris.data, iris.target
linear_model = LogisticRegression()
linear_model.fit(X, y)
```

Figure 2

Now, to save the *"linear_model"* you can use the following code:

```
import pickle
# Save the trained model to a file
with open('model.pickle', 'wb') as f:
    pickle.dump(linear_model, f)
```

Or

```
import pickle
# Save the trained model to a file
pickle_out = open("model.pkl", "wb")
pickle.dump(linear_model, pickle_out)
pickle_out.close()
```

Figure 3: Code snippet options for saving the ML model.

To load the saved model back into memory later, you can use the pickle.load() function, like this:

```
# Load the saved model from a file
with open('model.pickle', 'rb') as f:
    saved_model = pickle.load(f)

# Use the loaded model to make predictions
X_new = [[5.0, 3.6, 1.3, 0.25]]
y_pred = saved_model.predict(X_new)
print(y_pred)
```

Or

```
# Load the saved model from a file
pickle_in = open('model.pkl', 'rb')
saved_model = pickle.load(pickle_in)

# Use the loaded model to make predictions
X_new = [[5.0, 3.6, 1.3, 0.25]]
y_pred = saved_model.predict(X_new)
print(y_pred)
```

Figure 4: Code snippet options for loading the pickle file.