

Kehua Chu (uid: 806153163)

## 0.) Import and Clean data

```
In [52]: # pip install google
```

Collecting google

Obtaining dependency information for google from <https://files.pythonhosted.org/packages/ac/35/17c9141c4ae21e9a29a43acdfd848e3e468a810517f862cad07977bf8fe9/google-3.0.0-py2.py3-none-any.whl.metadata>

Downloading google-3.0.0-py2.py3-none-any.whl.metadata (627 bytes)

Requirement already satisfied: beautifulsoup4 in e:\..kacie\anacondakc\lib\site-packages (from google) (4.12.2)

Requirement already satisfied: soupsieve>1.2 in e:\..kacie\anacondakc\lib\site-packages (from beautifulsoup4->google) (2.4)

Downloading google-3.0.0-py2.py3-none-any.whl (45 kB)

```
----- 0.0/45.3 kB ? eta -:-:--
----- 10.2/45.3 kB ? eta -:-:--
----- 30.7/45.3 kB 262.6 kB/s eta 0:00:01
----- 41.0/45.3 kB 326.8 kB/s eta 0:00:01
----- 45.3/45.3 kB 280.1 kB/s eta 0:00:00
```

Installing collected packages: google

Successfully installed google-3.0.0

Note: you may need to restart the kernel to use updated packages.

```
In [1]: import pandas as pd
# from google.colab import drive
import matplotlib.pyplot as plt
import numpy as np
```

```
In [2]: from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import BaggingClassifier
from sklearn.datasets import make_classification
from sklearn.metrics import accuracy_score
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.tree import plot_tree
from sklearn.metrics import confusion_matrix
import seaborn as sns
```

```
In [ ]: #drive.mount('/content/gdrive/', force_remount = True)
```

Mounted at /content/gdrive/

```
In [10]: df = pd.read_csv('bank-additional-full.csv')
```

```
In [11]: # can use df = pd.read_csv('bank-additional-full.csv', sep = ';')
```

```
In [12]: df.head()
```

```
Out[12]:
```

	age	job	marital	education	default	housing	loan	contact	month	day_of_week	...	campaign	pdays	previous	poutcome	en
0	56	housemaid	married	basic.4y	no	no	no	telephone	may	mon	...	1	999	0	nonexistent	
1	57	services	married	high.school	unknown	no	no	telephone	may	mon	...	1	999	0	nonexistent	
2	37	services	married	high.school	no	yes	no	telephone	may	mon	...	1	999	0	nonexistent	
3	40	admin.	married	basic.6y	no	no	no	telephone	may	mon	...	1	999	0	nonexistent	
4	56	services	married	high.school	no	no	yes	telephone	may	mon	...	1	999	0	nonexistent	

5 rows × 21 columns



```
In [13]: df = df.drop(["default", "pdays", "previous", "poutcome", "emp.var.rate", "cons.price.idx", "cons.conf.idx", "eurib
```

```
Out[13]:
```

	age	job	marital	education	housing	loan	contact	month	day_of_week	duration	campaign	y
0	56	housemaid	married	basic.4y	no	no	telephone	may	mon	261	1	no
1	57	services	married	high.school	no	no	telephone	may	mon	149	1	no
2	37	services	married	high.school	yes	no	telephone	may	mon	226	1	no
3	40	admin.	married	basic.6y	no	no	telephone	may	mon	151	1	no
4	56	services	married	high.school	no	yes	telephone	may	mon	307	1	no
...	...	...	...	...	...	...	...	...	...	...	...	...
41183	73	retired	married	professional.course	yes	no	cellular	nov	fri	334	1	yes
41184	46	blue-collar	married	professional.course	no	no	cellular	nov	fri	383	1	no
41185	56	retired	married	university.degree	yes	no	cellular	nov	fri	189	2	no
41186	44	technician	married	professional.course	no	no	cellular	nov	fri	442	1	yes
41187	74	retired	married	professional.course	yes	no	cellular	nov	fri	239	3	no

41188 rows × 12 columns

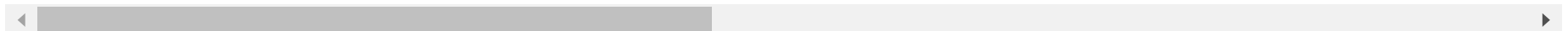
```
In [14]: df = pd.get_dummies(df, columns = ["loan", "job", "marital", "housing", "contact", "day_of_week", "campaign", "month", "education"], drop_first = T
```

```
In [8]: df.head()
```

```
Out[8]:
```

	age	duration	y	loan_unknown	loan_yes	job_blue-collar	job_entrepreneur	job_housemaid	job_management	job_retired	...	month_nov	month_
0	56	261	no	False	False	False	False	True	False	False	...	False	F
1	57	149	no	False	False	False	False	False	False	False	...	False	F
2	37	226	no	False	False	False	False	False	False	False	...	False	F
3	40	151	no	False	False	False	False	False	False	False	...	False	F
4	56	307	no	False	True	False	False	False	False	False	...	False	F

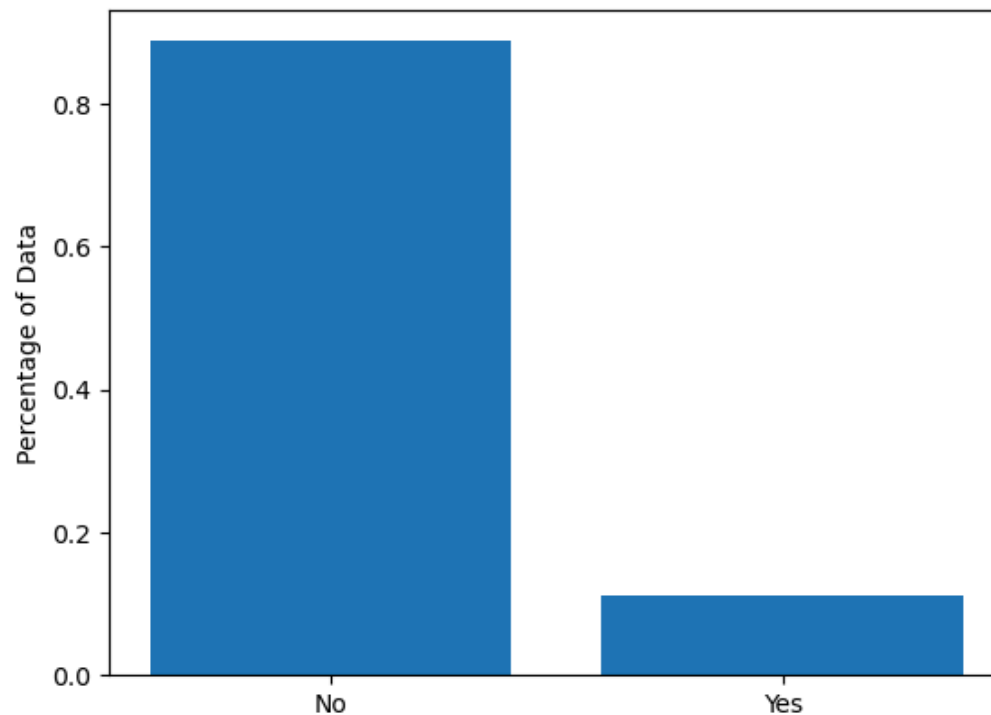
5 rows × 83 columns



```
In [15]: y = pd.get_dummies(df["y"], drop_first = True)
X = df.drop(["y"], axis = 1)
```

```
In [ ]:
```

```
In [16]: obs = len(y)
plt.bar(["No", "Yes"], [len(y[y.yes==0])/obs, len(y[y.yes==1])/obs])
plt.ylabel("Percentage of Data")
plt.show()
```



1.) Based on the visualization above, use your expert opinion to transform the data based on what we learned this quarter

```
In [17]: # Train Test Split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
```

```

scaler = StandardScaler().fit(X_train)

X_scaled = scaler.transform(X_train)
X_test = scaler.transform(X_test)

```

```

In [18]: #####
        ###TRANSFORM###
        #####
        #here we can notice that the data is highly imbalanced, therefore we need to use smote to balance the data.
        from imblearn.over_sampling import RandomOverSampler
        from imblearn.under_sampling import RandomUnderSampler
        from imblearn.over_sampling import SMOTE

        ros = RandomOverSampler()
        over_X, over_y = ros.fit_resample(X_train, y_train)

        X_scaled = over_X
        y_train = over_y

        #smote = SMOTE()
        #smote_X, smote_y = smote.fit_resample(X_train, y_train)

        #X_scaled = smote_X
        #y_train = smote_y

```

## 2.) Build and visualize a decision tree of Max Depth 3. Show the confusion matrix.

In [ ]:

```

In [19]: dtree_main = DecisionTreeClassifier(max_depth = 3)
        dtree_main.fit(X_scaled, y_train)

```

```

Out[19]: ▼ DecisionTreeClassifier
        DecisionTreeClassifier(max_depth=3)

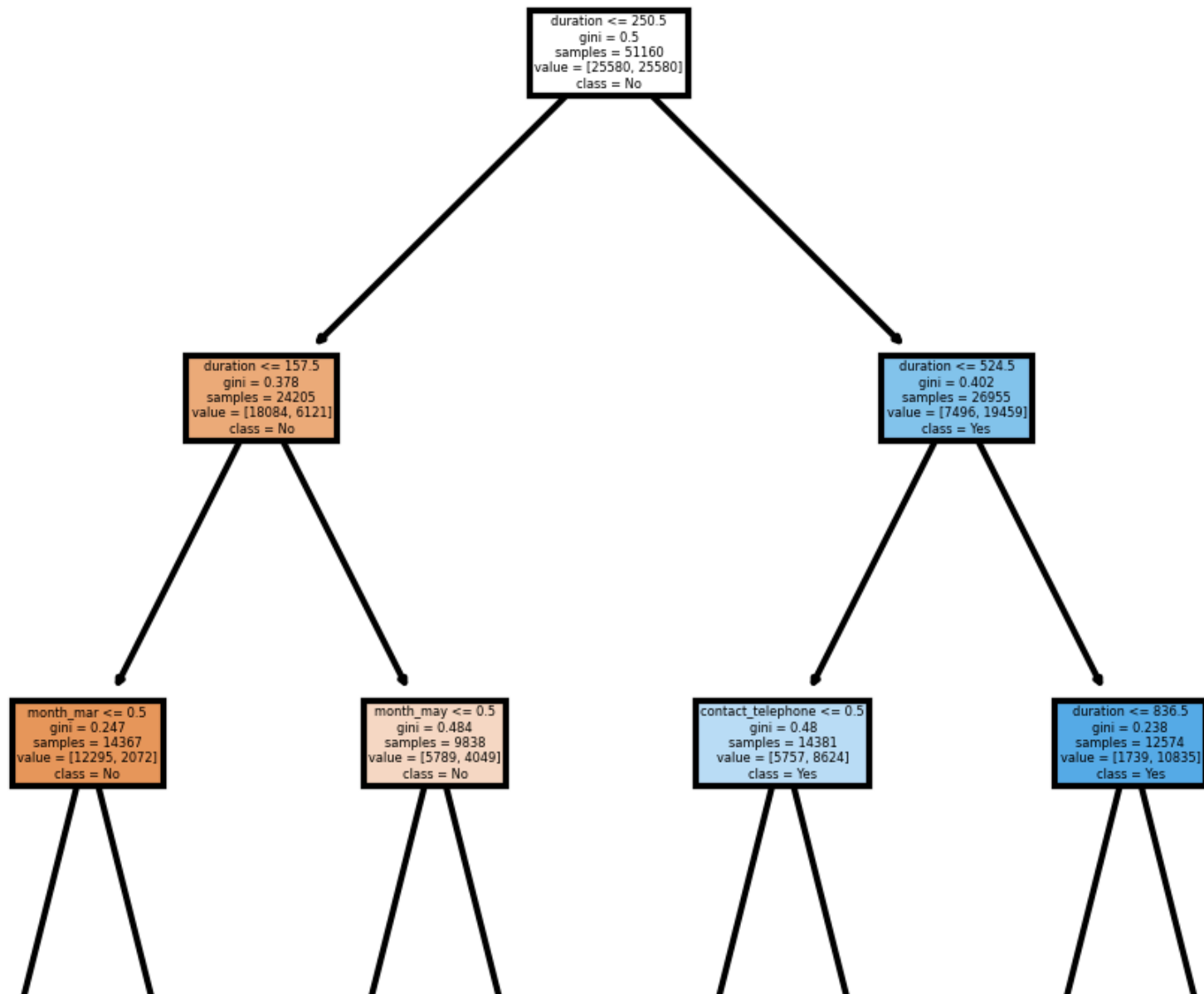
```

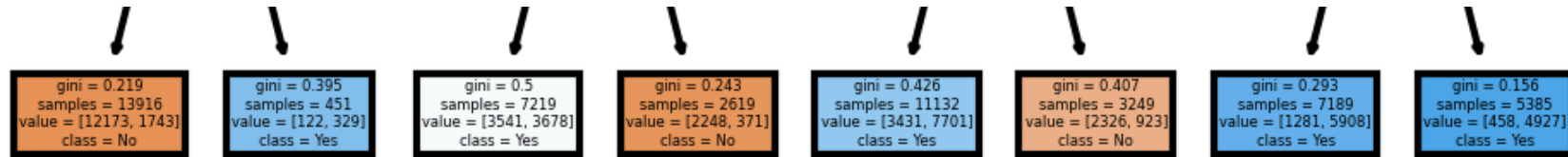
```

In [22]: fig, axes = plt.subplots(nrows = 1,ncols = 1,figsize = (4,4), dpi=300)
        feature_names = list(X.columns)
        plot_tree(dtree_main, filled=True, feature_names=feature_names, class_names=["No", "Yes"])

```

```
Out[22]: [Text(0.5, 0.875, 'duration <= 250.5\ngini = 0.5\nsamples = 51160\nvalue = [25580, 25580]\nclass = No'),
Text(0.25, 0.625, 'duration <= 157.5\ngini = 0.378\nsamples = 24205\nvalue = [18084, 6121]\nclass = No'),
Text(0.125, 0.375, 'month_mar <= 0.5\ngini = 0.247\nsamples = 14367\nvalue = [12295, 2072]\nclass = No'),
Text(0.0625, 0.125, 'gini = 0.219\nsamples = 13916\nvalue = [12173, 1743]\nclass = No'),
Text(0.1875, 0.125, 'gini = 0.395\nsamples = 451\nvalue = [122, 329]\nclass = Yes'),
Text(0.375, 0.375, 'month_may <= 0.5\ngini = 0.484\nsamples = 9838\nvalue = [5789, 4049]\nclass = No'),
Text(0.3125, 0.125, 'gini = 0.5\nsamples = 7219\nvalue = [3541, 3678]\nclass = Yes'),
Text(0.4375, 0.125, 'gini = 0.243\nsamples = 2619\nvalue = [2248, 371]\nclass = No'),
Text(0.75, 0.625, 'duration <= 524.5\ngini = 0.402\nsamples = 26955\nvalue = [7496, 19459]\nclass = Yes'),
Text(0.625, 0.375, 'contact_telephone <= 0.5\ngini = 0.48\nsamples = 14381\nvalue = [5757, 8624]\nclass = Yes'),
Text(0.5625, 0.125, 'gini = 0.426\nsamples = 11132\nvalue = [3431, 7701]\nclass = Yes'),
Text(0.6875, 0.125, 'gini = 0.407\nsamples = 3249\nvalue = [2326, 923]\nclass = No'),
Text(0.875, 0.375, 'duration <= 836.5\ngini = 0.238\nsamples = 12574\nvalue = [1739, 10835]\nclass = Yes'),
Text(0.8125, 0.125, 'gini = 0.293\nsamples = 7189\nvalue = [1281, 5908]\nclass = Yes'),
Text(0.9375, 0.125, 'gini = 0.156\nsamples = 5385\nvalue = [458, 4927]\nclass = Yes')]
<Figure size 1000x800 with 0 Axes>
```





## 1b.) Confusion matrix on out of sample data. Visualize and store as variable

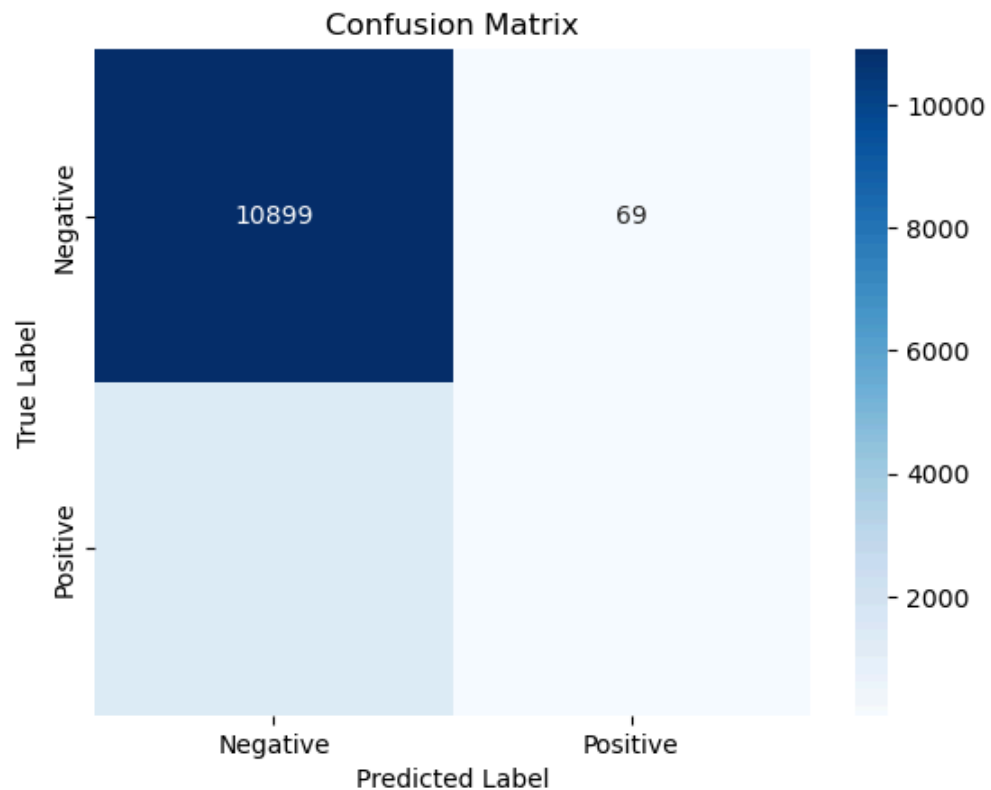
```
In [23]: y_pred = dtree_main.predict(X_test)
y_true = y_test
cm_raw = confusion_matrix(y_true, y_pred)
```

e:\..Kacie\AnacondaKC\Lib\site-packages\sklearn\base.py:464: UserWarning: X does not have valid feature names, but DecisionTreeClassifier was fitted with feature names  
warnings.warn(

```
In [24]: class_labels = ['Negative', 'Positive']

# Plot the confusion matrix as a heatmap
sns.heatmap(cm_raw, annot=True, fmt='d', cmap='Blues', xticklabels=class_labels, yticklabels=class_labels)
plt.title('Confusion Matrix')
plt.xlabel('Predicted Label')
plt.ylabel('True Label')
plt.show()
```





### 3.) Use bagging on your decision tree

```
In [25]: dtree = DecisionTreeClassifier(max_depth=3)
```

```
In [26]: bagging = BaggingClassifier(estimator = dtree,  
                                   n_estimators=100,  
                                   max_samples=0.5,  
                                   max_features=1.0  
                                   )
```

```
In [27]: bagging.fit(X_scaled, y_train)  
y_pred = bagging.predict(X_test)
```

```
y_true = y_test
cm_raw = confusion_matrix(y_true, y_pred)
```

e:\..Kacie\AnacondaKC\Lib\site-packages\sklearn\ensemble\\_bagging.py:802: DataConversionWarning: A column-vector y was passed when a 1d array was expected. Please change the shape of y to (n\_samples, ), for example using ravel().

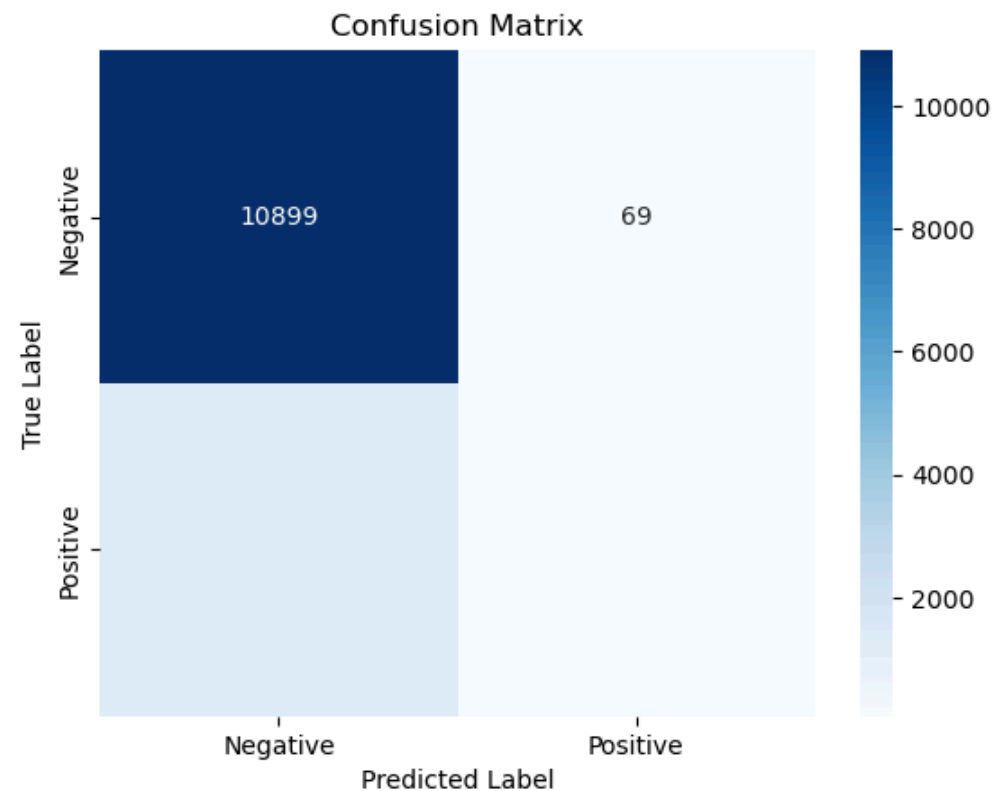
```
y = column_or_1d(y, warn=True)
```

e:\..Kacie\AnacondaKC\Lib\site-packages\sklearn\base.py:464: UserWarning: X does not have valid feature names, but BaggingClassifier was fitted with feature names

```
warnings.warn(
```

```
In [28]: class_labels = ['Negative', 'Positive']

# Plot the confusion matrix as a heatmap
sns.heatmap(cm_raw, annot=True, fmt='d', cmap='Blues', xticklabels=class_labels, yticklabels=class_labels)
plt.title('Confusion Matrix')
plt.xlabel('Predicted Label')
plt.ylabel('True Label')
plt.show()
```



```
In [ ]:
```

```
In [ ]:
```

## 4.) Boost your tree

```
In [29]: from sklearn.ensemble import AdaBoostClassifier
```

```
In [30]: dtree=DecisionTreeClassifier(max_depth=3)
```

```
In [31]: boost=AdaBoostClassifier(estimator= dtree,
                                n_estimators = 100,
                                learning_rate= 0.1)
```

```
In [32]: boost.fit(X_scaled, y_train)
y_pred=bagging.predict(X_test)
y_true = y_test
cm_raw = confusion_matrix(y_true, y_pred)
```

e:\..Kacie\AnacondaKC\Lib\site-packages\sklearn\utils\validation.py:1184: DataConversionWarning: A column-vector y was passed when a 1d array was expected. Please change the shape of y to (n\_samples, ), for example using ravel().

y = column\_or\_1d(y, warn=True)

e:\..Kacie\AnacondaKC\Lib\site-packages\sklearn\base.py:464: UserWarning: X does not have valid feature names, but BaggingClassifier was fitted with feature names

warnings.warn(

```
In [33]: class_labels = ['Negative', 'Positive']
```

```
# Plot the confusion matrix as a heatmap
```

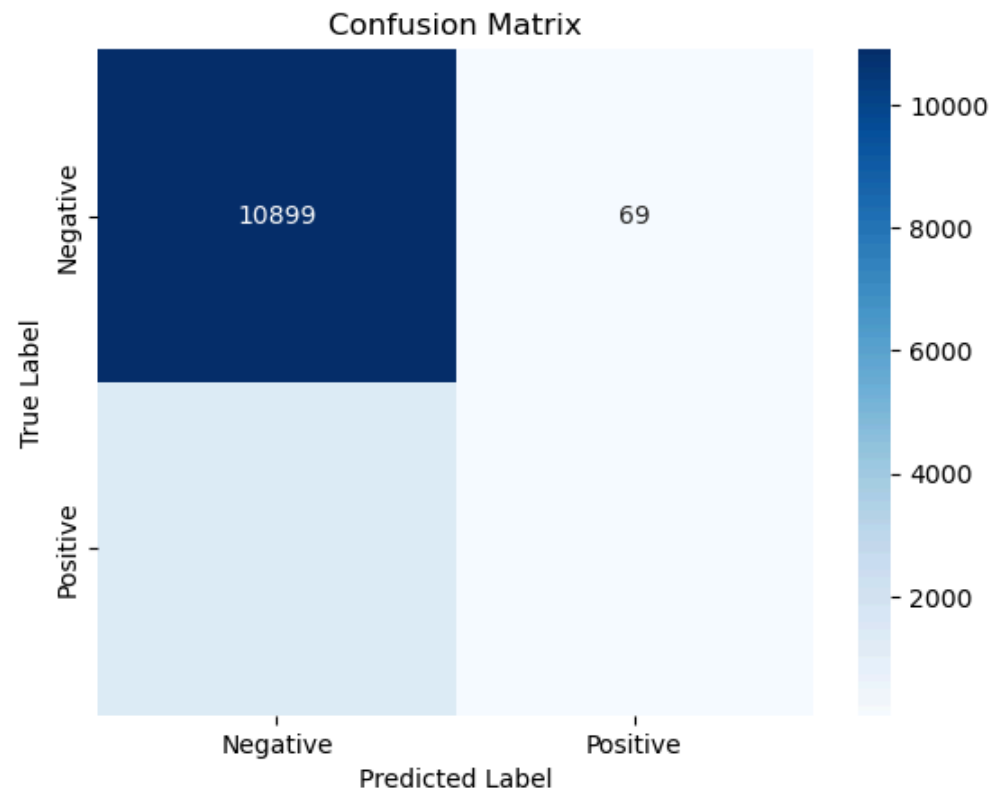
```
sns.heatmap(cm_raw, annot=True, fmt='d', cmap='Blues', xticklabels=class_labels, yticklabels=class_labels)
```

```
plt.title('Confusion Matrix')
```

```
plt.xlabel('Predicted Label')
```

```
plt.ylabel('True Label')
```

```
plt.show()
```



5.) Create a superlearner with at least 4 base learner models. Use a logistic reg for your metalearner. Interpret your coefficients and save your CM.

```
In [42]: # pip install mlens
```

Collecting mlens

Downloading mlens-0.2.3-py2.py3-none-any.whl (227 kB)

```
----- 0.0/227.7 kB ? eta -:--:--  
----- 10.2/227.7 kB ? eta -:--:--  
----- 41.0/227.7 kB 326.8 kB/s eta 0:00:01  
----- 225.3/227.7 kB 1.5 MB/s eta 0:00:01  
----- 227.7/227.7 kB 1.3 MB/s eta 0:00:00
```

Requirement already satisfied: numpy>=1.11 in e:\..kacie\anacondakc\lib\site-packages (from mlens) (1.24.3)

Requirement already satisfied: scipy>=0.17 in e:\..kacie\anacondakc\lib\site-packages (from mlens) (1.11.4)

Installing collected packages: mlens

Successfully installed mlens-0.2.3

Note: you may need to restart the kernel to use updated packages.

```
In [34]: from sklearn.linear_model import LogisticRegression  
# from sklearn.ensemble import RandomForestClassifier  
# from sklearn.neighbors import KNeighborsClassifier
```

```
In [46]: # from mlens.ensemble import SuperLearner
```

```
In [35]: X_base_learners=[list(bagging.predict(X_scaled)),list(boost.predict(X_scaled)),list(dtrees_main.predict(X_scaled))]  
super_learner=LogisticRegression()  
super_learner.fit(np.column_stack(X_base_learners), y_train)
```

e:\..Kacie\AnacondaKC\Lib\site-packages\sklearn\utils\validation.py:1184: DataConversionWarning: A column-vector y was passed when a 1d array was expected. Please change the shape of y to (n\_samples, ), for example using ravel().

y = column\_or\_1d(y, warn=True)

```
Out[35]: ▼ LogisticRegression  
LogisticRegression()
```

```
In [36]: super_learner.coef_
```

```
Out[36]: array([[ -0.06749077,  2.84611709,  1.0514362 ]])
```

```
In [ ]:
```

```
In [ ]:
```

```
In [ ]:
```

```
In [ ]:
```

6.)

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



```
/usr/local/lib/python3.8/dist-packages/sklearn/ensemble/_base.py:166: FutureWarning: `base_estimator` was renamed to `estimator` in version 1.2 and will be removed in 1.4.
```

```
warnings.warn(
```

```
Best Hyperparameters: {'base_estimator': LogisticRegression(), 'max_features': 0.7, 'max_samples': 0.5, 'n_estimators': 50}
```

```
Best Accuracy Score: 0.9289999999999999
```

In [ ]: