```
use predict = True
          output name = 'average.csv' if use average else 'normal.csv'
          output name2020 = '2020average.csv' if use average else '2020normal.csv'
In [98]: def read csv(file):
              df = pd.read csv(file)
              del df['Station Name']
              del df['Climate ID']
              del df['Longitude (x)']
              del df['Latitude (y)']
              del df['Year']
              del df['Month']
              del df['Day']
              del df['Time']
              df.columns = ['date', 'temp', 'dew', 'rel', 'wdir', 'wspeed', 'stn']
              df['date'] = pd.to_datetime(df['date'])
              return df
In [99]: if use predict:
              df1 = read csv('HALIFAX DOCKYARD.csv')
              df2 = read csv('HALIFAX KOOTENAY.csv')
          else:
              df1 = read_csv('2020-HALIFAX DOCKYARD.csv')
              df2 = read_csv('2020-HALIFAX KOOTENAY.csv')
          if use average:
              data = {'date': df1['date'],
                      'temp': (df1['temp'] + df2['temp'])/2.0,
                      'dew': (df1['dew'] + df2['dew'])/2.0,
                       'rel': (df1['rel'] + df2['rel'])/2.0,
                      'wdir': (df1['wdir'] + df2['wdir'])/2.0,
                      'wspeed': (df1['wspeed'] + df2['wspeed'])/2.0,
                      'stn': (df1['stn'] + df2['stn'])/2.0,
              data = {'date': df1['date'],
                      's1_temp': df1['temp'],
                      's2_temp': df2['temp'],
                      's1_dew': df1['dew'],
                      's2_dew': df2['dew'],
                      's1_rel': df1['rel'],
                      's2 rel': df2['rel'],
                      's1_wdir': df1['wdir'],
                      's2_wdir': df2['wdir'],
                      's1_wspeed': df1['wspeed'],
                      's2_wspeed': df2['wspeed'],
                      's1_stn': df1['stn'],
                      's2_stn': df2['stn'],
In [100]: | df = pd.DataFrame(data)
           def str2datetime(string):
               return datetime.datetime.strptime(string, '%b %d, %Y %I%p')
          if use_predict:
              headache = [
               [str2datetime('Sep 5, 2019 7AM'), str2datetime('Sep 5, 2019 10AM'), 8],
               [str2datetime('Sep 15, 2019 9AM'), str2datetime('Sep 15, 2019 3PM'), 7],
               [str2datetime('Sep 27, 2019 3PM'), str2datetime('Sep 27, 2019 9PM'), 8],
               [str2datetime('Oct 15, 2019 2PM'), str2datetime('OCT 15, 2019 6PM'), 7],
               [str2datetime('Oct 16, 2019 4AM'), str2datetime('Oct 16, 2019 6PM'), 10],
               [str2datetime('Oct 28, 2019 3PM'), str2datetime('OCT 28, 2019 9PM'), 6],
               [str2datetime('Nov 28, 2019 5PM'), str2datetime('Nov 28, 2019 6PM'), 6],
               [str2datetime('Dec 10, 2019 5AM'), str2datetime('Dec 10, 2019 10AM'), 4],
               [str2datetime('Dec 15, 2019 2AM'), str2datetime('Dec 15, 2019 10AM'), 7],
               [str2datetime('Dec 20, 2019 1PM'), str2datetime('Dec 20, 2019 6PM'), 9],
              ]
           workout = [
               [str2datetime('SEP 2, 2019 6PM'), str2datetime('SEP 2, 2019 7PM')],
               [str2datetime('SEP 4, 2019 6PM'), str2datetime('SEP 4, 2019 7PM')],
               [str2datetime('SEP 6, 2019 6PM'), str2datetime('SEP 6, 2019 7PM')],
               [str2datetime('SEP 27, 2019 6PM'), str2datetime('SEP 27, 2019 7PM')],
               [str2datetime('OCT 8, 2019 6PM'), str2datetime('OCT 8, 2019 7PM')],
               [str2datetime('OCT 20, 2019 8PM'), str2datetime('OCT 20, 2019 9PM')],
               [str2datetime('OCT 27, 2019 6AM'), str2datetime('OCT 27, 2019 7AM')],
               [str2datetime('OCT 28, 2019 6PM'), str2datetime('OCT 28, 2019 7PM')],
               [str2datetime('NOV 2, 2019 6PM'), str2datetime('NOV 2, 2019 7PM')],
               [str2datetime('NOV 7, 2019 6PM'), str2datetime('NOV 7, 2019 7PM')],
               [str2datetime('NOV 9, 2019 6PM'), str2datetime('NOV 9, 2019 7PM')],
               [str2datetime('NOV 12, 2019 6PM'), str2datetime('NOV 12, 2019 7PM')],
               [str2datetime('NOV 20, 2019 6PM'), str2datetime('NOV 20, 2019 7PM')],
               [str2datetime('DEC 15, 2019 6PM'), str2datetime('DEC 15, 2019 7PM')],
               [str2datetime('DEC 16, 2019 6PM'), str2datetime('DEC 16, 2019 7PM')],
               [str2datetime('SEP 2, 2020 6PM'), str2datetime('SEP 2, 2020 7PM')],
               [str2datetime('SEP 4, 2020 6PM'), str2datetime('SEP 4, 2020 7PM')],
               [str2datetime('SEP 6, 2020 6PM'), str2datetime('SEP 6, 2020 7PM')],
In [101]: | df['hour'] = pd.Series([0]* df['date'].size)
          df['dayweek'] = pd.Series([0]* df['date'].size)
           for i in range(df['date'].size):
               df['hour'][i] = df['date'][i].hour
              df['dayweek'][i] = df['date'][i].weekday()
          df['workout'] = pd.Series([0]* df['date'].size)
          workout index = []
          for w in workout:
              for i in range(df['date'].size):
                   if w[0] <= df['date'][i] < w[1]:</pre>
                      workout index.append(i)
          df['workout'][workout_index] = 1
          df['work'] = pd.Series([0]* df['date'].size)
           for i in range(df['date'].size):
              if df['date'][i].weekday() == 6 or df['date'][i].weekday() == 0:
                   continue
              if df['date'][i].hour < 9 or df['date'][i].hour > 17:
                   continue
              if str2datetime('OCT 27, 2019 1AM') < df['date'][i] < str2datetime('NOV 3, 2019 11PM'):</pre>
                   continue
              df['work'][i] = 1
          df['work hour'] = pd.Series([0] * df['date'].size)
          work sum = 0
          for i in range(df['date'].size):
              if df['work'][i] == 1:
                  work sum += 1
                  df['work hour'][i] = work sum
              else:
                  work sum = 0
          if use predict:
              df['headache'] = pd.Series([0]* df['date'].size)
               for w in headache:
                   for i in range(df['date'].size):
                       if df['date'][i] >= w[0] and df['date'][i] < w[1]:</pre>
                           df['headache'][i] = w[2]
           # del df['date']
          if use predict:
              df.to csv(output name)
          else:
              df.to_csv(output_name2020)
          <ipython-input-101-79c08b35a8fb>:4: SettingWithCopyWarning:
          A value is trying to be set on a copy of a slice from a DataFrame
          See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user guide/indexin
          g.html#returning-a-view-versus-a-copy
            df['hour'][i] = df['date'][i].hour
          <ipython-input-101-79c08b35a8fb>:5: SettingWithCopyWarning:
          A value is trying to be set on a copy of a slice from a DataFrame
          See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexin
          g.html#returning-a-view-versus-a-copy
            df['dayweek'][i] = df['date'][i].weekday()
          <ipython-input-101-79c08b35a8fb>:13: SettingWithCopyWarning:
          A value is trying to be set on a copy of a slice from a DataFrame
          See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user guide/indexin
          g.html#returning-a-view-versus-a-copy
            df['workout'][workout_index] = 1
          <ipython-input-101-79c08b35a8fb>:23: SettingWithCopyWarning:
          A value is trying to be set on a copy of a slice from a DataFrame
          See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexin
          g.html#returning-a-view-versus-a-copy
            df['work'][i] = 1
          <ipython-input-101-79c08b35a8fb>:29: SettingWithCopyWarning:
          A value is trying to be set on a copy of a slice from a DataFrame
          See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexin
          g.html#returning-a-view-versus-a-copy
            df['work hour'][i] = work sum
          <ipython-input-101-79c08b35a8fb>:38: SettingWithCopyWarning:
          A value is trying to be set on a copy of a slice from a DataFrame
          See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexin
          g.html#returning-a-view-versus-a-copy
            df['headache'][i] = w[2]
In [102]: if use predict and use average:
              data = pd.read csv('average.csv')
          elif use predict and not use average:
              data = pd.read csv('normal.csv')
          elif not use predict and use average:
              data = pd.read csv('2020average.csv')
              data = pd.read_csv('2020normal.csv')
          df = data.dropna(axis = 0 , how = 'any')
          checkFor nan = df.isnull().values.any()
          print (checkFor nan)
          False
In [103]: | if use_predict and use_average:
              df.to_csv('DataSetAverage.csv')
          elif use predict and not use average:
              df.to csv('DataSetNormal.csv')
          elif not use predict and use average:
              df.to csv('PredictionDataAverage.csv')
              df.to csv('PredictionDataNormal.csv')
In [104]: import matplotlib.pyplot as plt
          plt.matshow(df.corr())
          plt.show()
                                  10
                          6
                                      12
            6
            8
```

10

12

In [97]: import pandas as pd

import datetime

use average = True

	<pre>import pandas as pd import numpy as np import matplotlib.pyplot as plt from sklearn.tree import DecisionTreeClassifier from sklearn.model_selection import cross_val_score from sklearn.utils.class_weight import compute_class_weight data = pd.read csv('DataSetAverage.csv')</pre>
<pre>In [103]: Out[103]:</pre>	
In [104]: Out[104]:	2901 rows × 13 columns #Finding Data Correlation alpha = data.columns data.corr()
<pre>In [105]: Out[105]:</pre>	workout 0.022741 0.014283 -0.020008 -0.007188 -0.018984 0.005779 0.060581 0.000443 1.000000 -0.041690 -0.035789 0.051 work 0.068829 -0.025341 -0.238168 -0.025122 0.187953 -0.055323 0.123158 -0.003540 -0.041690 1.000000 0.858458 0.068 work_hour 0.075488 -0.025818 -0.251976 -0.012792 0.167101 -0.060976 0.201790 -0.003040 -0.035789 0.858458 1.000000 0.085 headache 0.082381 0.063103 -0.031769 0.014816 0.010035 -0.025244 0.004529 0.010346 0.051382 0.068859 0.085044 1.000
In [106]:	<pre>dtype='object') We checked for correlation between features to avoid redundunt feature. As you may see in correlation hitmap temp and dew temprature are highly correlated. alpha = data.columns fig = plt.figure() ax = fig.add_subplot(111) cax = ax.matshow(data.corr()) fig.colorbar(cax) ax.set_xticklabels(['']+alpha, rotation='vertical', verticalalignment='bottom') ax.set_yticklabels(['']+alpha)</pre>
	plt.show() plt.close()
	#Data with high correlation is deleted df = data.drop(['dew', 'work', 'dayweek'], 1) df.columns Index(['date', 'temp', 'rel', 'wdir', 'wspeed', 'stn', 'hour', 'workout',
In [122]:	<pre>dtype='object') #check and remove null record data by row. Rnan = [] for index, row in df.iterrows(): isnanValue = row.isnull() if isnanValue.any(): Rnan.append(index) print(Rnan) df = df.dropna(axis=0, how='any') df.isnull().values.any()</pre>
	<pre>False from collections import Counter X_train = df[['temp', 'rel', 'wdir', 'wspeed', 'stn', 'hour', 'workout', 'work_hour']] y_train = df['headache'] factor = compute_class_weight('balanced', list(Counter(y_train).keys()), y_train)</pre>
	<pre>class_weights = {k:v for k, v in zip(Counter(y_train).keys(), factor)} C:\Users\reza\anaconda3\lib\site-packages\sklearn\utils\validation.py:68: FutureWarning: Pass classes =[0, 8, 7, 10, 6, 4, 9], y=0</pre>
	2900 0 Name: headache, Length: 2901, dtype: int64 as keyword args. From version 0.25 passing these as positi onal arguments will result in an error warnings.warn("Pass {} as keyword args. From version 0.25 " class_weights {0: 0.14572031344183242, 8: 46.04761904761905, 7: 23.023809523809526, 10: 29.602040816326532, 6: 69.07142857142857, 4: 82.88571428571429,
In [158]:	<pre>from sklearn.model_selection import cross_val_score import numpy as np import matplotlib.pyplot as plt from sklearn import tree def run_cross_validation_on_trees(X, y, tree_depths, weights=None, cv=5, scoring='accuracy'): Slist = [] Mean = []</pre>
	<pre>InputCV = [] accuracy = [] for depth in tree_depths: tree_model = DecisionTreeClassifier(max_depth=depth, class_weight=weights) SC = cross_val_score(tree_model, X, y, cv=cv, scoring=scoring) Mean.append(SC.mean()) InputCV.append(SC.std()) Slist.append(SC) accuracy.append(tree_model.fit(X, y).score(X, y)) Mean = np.array(Mean) InputCV = np.array(InputCV) accuracy = np.array(accuracy) return Mean, InputCV, accuracy def plot_cross_validation_on_trees(depths, cv_scores_mean, cv_scores_std, accuracy_scores, title): fig, ax = plt.subplots(1,1, figsize=(15,5)) ax.plot(depths, cv_scores_mean, '-o', label='mean cross-validation accuracy', alpha=0.9) ax.fill_between(depths, cv_scores_mean-cv_scores_std, cv_scores_mean+cv_scores_std, alpha=0.2) ylim = plt.ylim() ax.plot(depths, accuracy_scores, '-*', label='train accuracy', alpha=0.9) ax.set_title(title, fontsize=16) ax.set_vlabel('Tree depth', fontsize=14) ax.set_ylabel('Accuracy', fontsize=14) ax.set_ylim(ylim) ax.set_ticks(depths)</pre>
In [159]:	<pre># fitting trees of depth 1 to 10 - unbalanced data sm_tree_depths = range(1,10) sm_cv_scores_mean, sm_cv_scores_std, sm_accuracy_scores = run_cross_validation_on_trees(X_train, y_tra in, sm_tree_depths) # plotting accuracy plot_cross_validation_on_trees(sm_tree_depths, sm_cv_scores_mean, sm_cv_scores_std, sm_accuracy_scores ,</pre>
	Accuracy per decision tree depth on training data-Unweighted classes 100
In [128]:	nee depen
	<pre>sm_tree_depths = range(1,10) sm_cv_scores_mean, sm_cv_scores_std, sm_accuracy_scores = run_cross_validation_on_trees(X_train, y_tra in, sm_tree_depths, class_weights) # plotting accuracy plot_cross_validation_on_trees(sm_tree_depths, sm_cv_scores_mean, sm_cv_scores_std, sm_accuracy_scores ,</pre>
	<pre>sm_tree_depths = range(1,10) sm_cv_scores_mean, sm_cv_scores_std, sm_accuracy_scores = run_cross_validation_on_trees(X_train, y_tra in, sm_tree_depths, class_weights) # plotting accuracy plot_cross_validation_on_trees(sm_tree_depths, sm_cv_scores_mean, sm_cv_scores_std, sm_accuracy_scores,</pre>
In [129]:	sm_tree_depths = range(1,10) sm_cv_scores_mean, sm_cv_scores_std, sm_accuracy_scores = run_cross_validation_on_trees(X_train, y_train, sm_tree_depths, class_weights) # plotting accuracy plot_cross_validation_on_trees(sm_tree_depths, sm_cv_scores_mean, sm_cv_scores_std, sm_accuracy_scores,
	<pre>sm tree_depths = range(1,10) sm cv_scores_mean, sm_cv_scores_std, sm_accuracy_scores = run_cross_validation_on_trees(X_train, y_train, sm_tree_depths, class_weights) # plotting accuracy plot_cross_validation_on_trees(sm_tree_depths, sm_cv_scores_mean, sm_cv_scores_std, sm_accuracy_scores /*Accuracy per decision tree depth on training data - weighted labels') Accuracy per decision tree depth on training data - weighted labels mean cross_validation accuracy</pre>
	<pre>sm tree depths = range(1,10) sm_cv_socres_mean, sm_cv_socres_std, sm_accuracy_scores = run_cross_validation_on_trees(X_train, y_train, sm_tree_depths, class_weights) # plotting accuracy plot_cross_validation_on_trees(sm_tree_depths, sm_cv_scores_mean, sm_cv_scores_std, sm_accuracy_scores</pre>
<pre>In [130]: In [131]: Out [131]:</pre>	page 1. See the proposed of th
<pre>In [130]: In [131]: Out [131]: In [133]: In [134]:</pre>	manufactures importances Accuracy per decision tree depth on training data - weighted labels The depth The
In [130]: In [131]: Out [131]: Out [133]: In [134]: In [135]:	in tock contain "range () If, as two depths, case weights) If as two depths, case weights as a contain now depth on training data — weighted labers? Accuracy per decision tree depth on training data — weighted labers? Accuracy per decision tree depth on training data — weighted labers? Accuracy per decision tree depth on training data — weighted labers? Accuracy per decision tree depth on training data — weighted labers? Accuracy per decision tree depth on training data — weighted labers? Accuracy per decision tree depth on training data — weighted labers? Accuracy per decision tree depth on training data — weighted labers? Accuracy per decision tree depth on training data — weighted labers? Accuracy per decision tree depth on training data — weighted labers? Accuracy per decision tree depth on training data — weighted labers? Accuracy per decision tree depth on training data — weighted labers? Accuracy per decision tree depth on training data — weighted labers? Accuracy per decision tree depth on training data — weighted labers? Accuracy per decision tree depth on training data — weighted labers? Accuracy per decision tree depth on training data — weighted labers? Accuracy per decision tree depth on training data— excused accuracy () 17- ()% on training dataset as a containing
In [130]: In [131]: Out [131]: Out [134]: In [135]: Out [135]:	### Company of the co
In [130]: In [131]: Out [131]: Out [134]: In [135]: Out [135]:	page 100 per
<pre>In [130]: In [131]: Out [131]: Out [133]: In [134]: Out [134]: Out [135]: Out [136]: In [137]:</pre>	### Comparison of the Comparis
<pre>In [130]: In [131]: Out [131]: Out [133]: In [134]: Out [134]: Out [135]: Out [136]: In [137]:</pre>	### Additional Processing Sections (1) 10 10 10 10 10 10 10