Classification with Python

In this notebook we try to practice all the classification algorithms that we have learned in this course.

We load a dataset using Pandas library, and apply the following algorithms, and find the best one for this specific dataset by accuracy evaluation methods.

Let's first load required libraries:

In [1]:
```
import itertools
import numpy as np
import matplotlib.pyplot as plt
from matplotlib.ticker import NullFormatter
import pandas as pd
import numpy as np
import matplotlib.ticker as ticker
from sklearn import preprocessing
%matplotlib inline
```

About dataset

This dataset is about past loans. The **Loan_train.csv** data set includes details of 346 customers whose loan are already paid off or defaulted. It includes following fields:

Field	Description
Loan_status	Whether a loan is paid off on in collection
Principal	Basic principal loan amount at the
Terms	Origination terms which can be weekly (7 days), biweekly, and monthly payoff schedule
Effective_date	When the loan got originated and took effects
Due_date	Since it's one-time payoff schedule, each loan has one single due date
Age	Age of applicant
Education	Education of applicant
Gender	The gender of applicant

Let's download the dataset

In [2]:
```
!wget -O loan_train.csv https://cf-courses-data.s3.us.cloud-object-storage.appdomain.clo
```

```
'wget' is not recognized as an internal or external command,
operable program or batch file.
```

Load Data From CSV File

In [4]:
```
df = pd.read_csv('loan_train.csv')
```

```
df.head()
```

Out[4]:

	Unnamed: 0.1	Unnamed: 0	loan_status	Principal	terms	effective_date	due_date	age	education	Gender
0	0	0	PAIDOFF	1000	30	9/8/2016	10/7/2016	45	High School or Below	male
1	2	2	PAIDOFF	1000	30	9/8/2016	10/7/2016	33	Bechalor	female
2	3	3	PAIDOFF	1000	15	9/8/2016	9/22/2016	27	college	male
3	4	4	PAIDOFF	1000	30	9/9/2016	10/8/2016	28	college	female
4	6	6	PAIDOFF	1000	30	9/9/2016	10/8/2016	29	college	male

In [5]:

```
df.shape
```

Out[5]:

```
(346, 10)
```

Convert to date time object

In [6]:

```
df['due_date'] = pd.to_datetime(df['due_date'])
df['effective_date'] = pd.to_datetime(df['effective_date'])
df.head()
```

Out[6]:

	Unnamed: 0.1	Unnamed: 0	loan_status	Principal	terms	effective_date	due_date	age	education	Gender
0	0	0	PAIDOFF	1000	30	2016-09-08	2016-10-07	45	High School or Below	male
1	2	2	PAIDOFF	1000	30	2016-09-08	2016-10-07	33	Bechalor	female
2	3	3	PAIDOFF	1000	15	2016-09-08	2016-09-22	27	college	male
3	4	4	PAIDOFF	1000	30	2016-09-09	2016-10-08	28	college	female
4	6	6	PAIDOFF	1000	30	2016-09-09	2016-10-08	29	college	male

Data visualization and pre-processing

Let’s see how many of each class is in our data set

In [7]:

```
df['loan_status'].value_counts()
```

Out[7]:

```
PAIDOFF       260
COLLECTION     86
Name: loan_status, dtype: int64
```

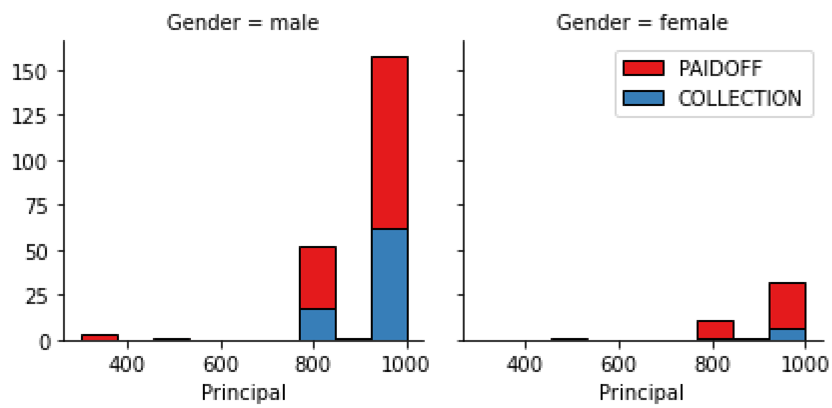
260 people have paid off the loan on time while 86 have gone into collection

Let's plot some columns to underestand data better:

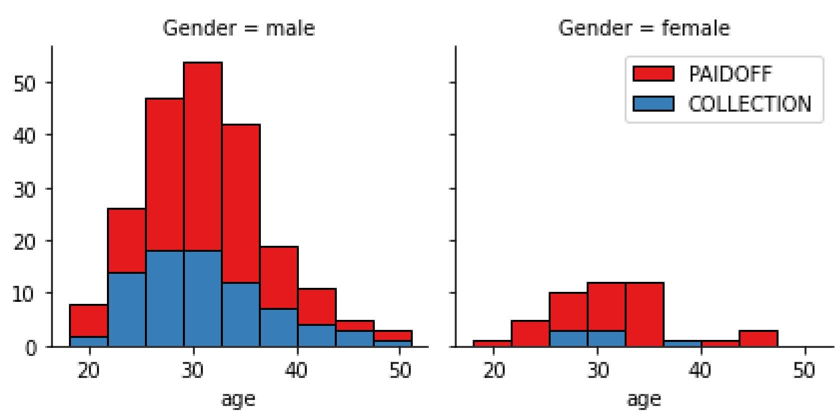
In [ ]:

```
# notice: installing seaborn might takes a few minutes
#!conda install -c anaconda seaborn -y
```

In [8]:
```
import seaborn as sns

bins = np.linspace(df.Principal.min(), df.Principal.max(), 10)
g = sns.FacetGrid(df, col="Gender", hue="loan_status", palette="Set1", col_wrap=2)
g.map(plt.hist, 'Principal', bins=bins, ec="k")

g.axes[-1].legend()
plt.show()
```



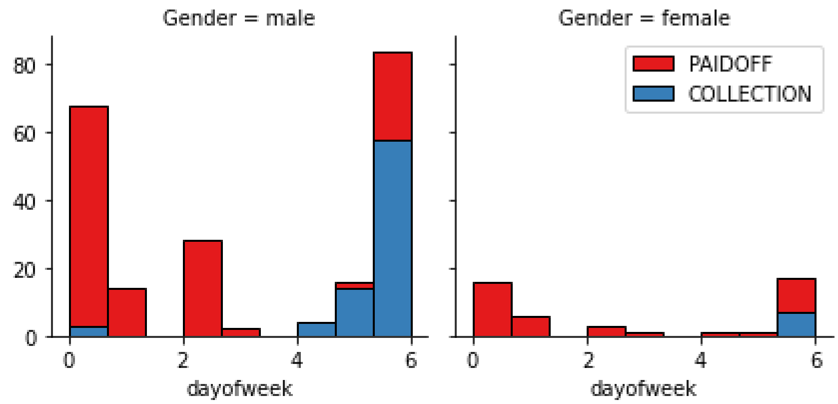
In [9]:
```
bins = np.linspace(df.age.min(), df.age.max(), 10)
g = sns.FacetGrid(df, col="Gender", hue="loan_status", palette="Set1", col_wrap=2)
g.map(plt.hist, 'age', bins=bins, ec="k")

g.axes[-1].legend()
plt.show()
```



Pre-processing: Feature selection/extraction

Let's look at the day of the week people get the loan

In [11]:
```
df['dayofweek'] = df['effective_date'].dt.dayofweek
bins = np.linspace(df.dayofweek.min(), df.dayofweek.max(), 10)
g = sns.FacetGrid(df, col="Gender", hue="loan_status", palette="Set1", col_wrap=2)
g.map(plt.hist, 'dayofweek', bins=bins, ec="k")
g.axes[-1].legend()
plt.show()
```



We see that people who get the loan at the end of the week don't pay it off, so let's use Feature binarization to set a threshold value less than day 4

In [14]:
```
df.head()
```

Out[14]:

	Unnamed: 0.1	Unnamed: 0	loan_status	Principal	terms	effective_date	due_date	age	education	Gender	dayofv
0	0	0	PAIDOFF	1000	30	2016-09-08	2016-10-07	45	High School or Below	male	
1	2	2	PAIDOFF	1000	30	2016-09-08	2016-10-07	33	Bechalor	female	
2	3	3	PAIDOFF	1000	15	2016-09-08	2016-09-22	27	college	male	
3	4	4	PAIDOFF	1000	30	2016-09-09	2016-10-08	28	college	female	
4	6	6	PAIDOFF	1000	30	2016-09-09	2016-10-08	29	college	male	

In [15]:
```
df['weekend'] = df['dayofweek'].apply(lambda x: 1 if (x>3)  else 0)
df.head()
```

Out[15]:

	Unnamed: 0.1	Unnamed: 0	loan_status	Principal	terms	effective_date	due_date	age	education	Gender	dayofv
0	0	0	PAIDOFF	1000	30	2016-09-08	2016-10-07	45	High School or Below	male	
1	2	2	PAIDOFF	1000	30	2016-09-08	2016-10-07	33	Bechalor	female	
2	3	3	PAIDOFF	1000	15	2016-09-08	2016-09-22	27	college	male	
3	4	4	PAIDOFF	1000	30	2016-09-09	2016-10-08	28	college	female	
4	6	6	PAIDOFF	1000	30	2016-09-09	2016-10-08	29	college	male	

Convert Categorical features to numerical values

Let's look at gender:

In [16]:
```
df.groupby(['Gender'])['loan_status'].value_counts(normalize=True)
```
Out[16]:
```
Gender  loan_status
female  PAIDOFF        0.865385
        COLLECTION     0.134615
male    PAIDOFF        0.731293
        COLLECTION     0.268707
Name: loan_status, dtype: float64
```

86 % of female pay there loans while only 73 % of males pay there loan

Let's convert male to 0 and female to 1:

In [17]:
```
df['Gender'].replace(to_replace=['male','female'], value=[0,1],inplace=True)
df.head()
```
Out[17]:

	Unnamed: 0.1	Unnamed: 0	loan_status	Principal	terms	effective_date	due_date	age	education	Gender	dayofw
0	0	0	PAIDOFF	1000	30	2016-09-08	2016-10-07	45	High School or Below	0	
1	2	2	PAIDOFF	1000	30	2016-09-08	2016-10-07	33	Bechalor	1	
2	3	3	PAIDOFF	1000	15	2016-09-08	2016-09-22	27	college	0	
3	4	4	PAIDOFF	1000	30	2016-09-09	2016-10-08	28	college	1	
4	6	6	PAIDOFF	1000	30	2016-09-09	2016-10-08	29	college	0	

One Hot Encoding

How about education?

In [18]:
```
df.groupby(['education'])['loan_status'].value_counts(normalize=True)
```
Out[18]:
```
education             loan_status
Bechalor              PAIDOFF        0.750000
                      COLLECTION     0.250000
High School or Below  PAIDOFF        0.741722
                      COLLECTION     0.258278
Master or Above       COLLECTION     0.500000
                      PAIDOFF        0.500000
college               PAIDOFF        0.765101
                      COLLECTION     0.234899
Name: loan_status, dtype: float64
```

Features before One Hot Encoding

In [19]:
```
df[['Principal','terms','age','Gender','education']].head()
```
Out[19]:

	Principal	terms	age	Gender	education
0	1000	30	45	0	High School or Below
1	1000	30	33	1	Bechalor

2	1000	15	27	0	college
3	1000	30	28	1	college
4	1000	30	29	0	college

Use one hot encoding technique to convert categorical varables to binary variables and append them to the feature Data Frame

In [21]:

```
Feature = df[['Principal','terms','age','Gender','weekend']]
Feature = pd.concat([Feature,pd.get_dummies(df['education'])], axis=1)
Feature.drop(['Master or Above'], axis = 1,inplace=True)
Feature.head()
```

Out[21]:

	Principal	terms	age	Gender	weekend	Bechalor	High School or Below	college
0	1000	30	45	0	0	0	1	0
1	1000	30	33	1	0	1	0	0
2	1000	15	27	0	0	0	0	1
3	1000	30	28	1	1	0	0	1
4	1000	30	29	0	1	0	0	1

Feature Selection

Let's define feature sets, X:

In [22]:

```
X = Feature
X[0:5]
```

Out[22]:

	Principal	terms	age	Gender	weekend	Bechalor	High School or Below	college
0	1000	30	45	0	0	0	1	0
1	1000	30	33	1	0	1	0	0
2	1000	15	27	0	0	0	0	1
3	1000	30	28	1	1	0	0	1
4	1000	30	29	0	1	0	0	1

What are our lables?

In [23]:

```
y = df['loan_status'].values
y[0:5]
```

Out[23]:

```
array(['PAIDOFF', 'PAIDOFF', 'PAIDOFF', 'PAIDOFF', 'PAIDOFF'],
      dtype=object)
```

Normalize Data

Data Standardization give data zero mean and unit variance (technically should be done after train test split)

In [24]:

```
X= preprocessing.StandardScaler().fit(X).transform(X)
X[0:5]
```

```
array([[ 0.51578458,  0.92071769,  2.33152555, -0.42056004, -1.20577805,
```

```
Out[24]:        -0.38170062,  1.13639374, -0.86968108],
               [ 0.51578458,  0.92071769,  0.34170148,  2.37778177, -1.20577805,
                 2.61985426, -0.87997669, -0.86968108],
               [ 0.51578458, -0.95911111, -0.65321055, -0.42056004, -1.20577805,
                -0.38170062, -0.87997669,  1.14984679],
               [ 0.51578458,  0.92071769, -0.48739188,  2.37778177,  0.82934003,
                -0.38170062, -0.87997669,  1.14984679],
               [ 0.51578458,  0.92071769, -0.3215732 , -0.42056004,  0.82934003,
                -0.38170062, -0.87997669,  1.14984679]])
```

Classification

Now, it is your turn, use the training set to build an accurate model. Then use the test set to report the accuracy of the model You should use the following algorithm:

- K Nearest Neighbor(KNN)
- Decision Tree
- Support Vector Machine
- Logistic Regression

__ Notice:__

- You can go above and change the pre-processing, feature selection, feature-extraction, and so on, to make a better model.
- You should use either scikit-learn, Scipy or Numpy libraries for developing the classification algorithms.
- You should include the code of the algorithm in the following cells.

K Nearest Neighbor(KNN)

Notice: You should find the best k to build the model with the best accuracy.\ **warning:** You should not use the **loan_test.csv** for finding the best k, however, you can split your train_loan.csv into train and test to find the best **k**.

In [161…
```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split( X, y, test_size=0.2, random_state=4
print ('Train set:', X_train.shape,  y_train.shape)
print ('Test set:', X_test.shape,  y_test.shape)
```

```
Train set: (276, 8) (276,)
Test set: (70, 8) (70,)
```

In [162…
```
from sklearn.neighbors import KNeighborsClassifier
from sklearn import metrics
Ks = 10
std_acc = np.zeros(Ks-1)
mean_acc = np.zeros(Ks-1)

for i in range(1, Ks):
    kneighbors = KNeighborsClassifier(n_neighbors = i).fit(X_train, y_train)
    yhat = kneighbors.predict(X_test)
    mean_acc[i-1] = metrics.accuracy_score(y_test, yhat)
    std_acc[i-1]=np.std(yhat==y_test)/np.sqrt(yhat.shape[0])

print("The best accuracy was", mean_acc.max(), "with k =", mean_acc.argmax()+1)
```

```
The best accuracy was 0.7857142857142857 with k = 7
```

In [163...

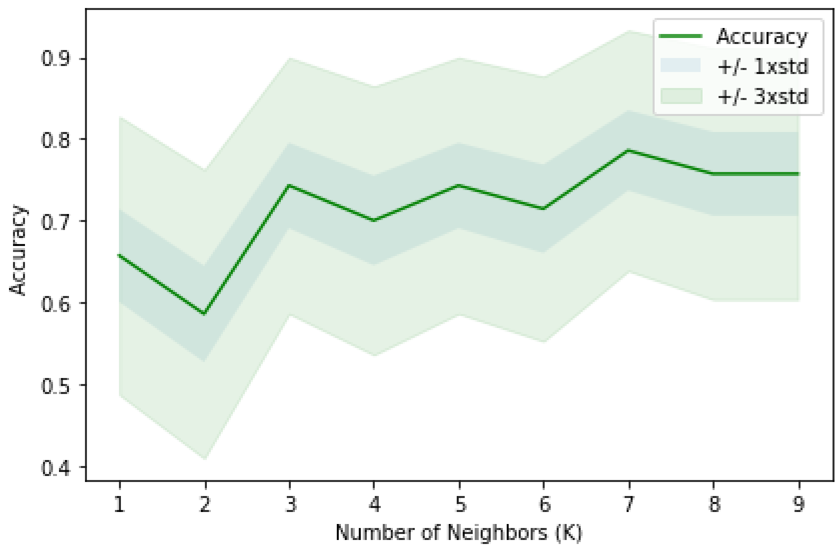
```
mean_acc
```

Out[163]:

```
array([0.65714286, 0.58571429, 0.74285714, 0.7       , 0.74285714,
       0.71428571, 0.78571429, 0.75714286, 0.75714286])
```

In [164...

```
plt.plot(range(1,Ks),mean_acc,'g')
plt.fill_between(range(1,Ks),mean_acc - 1 * std_acc,mean_acc + 1 * std_acc, alpha=0.1)
plt.fill_between(range(1,Ks),mean_acc - 3 * std_acc,mean_acc + 3 * std_acc, alpha=0.10,c
plt.legend(('Accuracy ', '+/- 1xstd','+/- 3xstd'))
plt.ylabel('Accuracy ')
plt.xlabel('Number of Neighbors (K)')
plt.tight_layout()
plt.show()
```



Decision Tree

In [165...

```
from sklearn.tree import DecisionTreeClassifier
loantree = DecisionTreeClassifier(criterion = 'entropy', max_depth = 6)
loantree.fit(X_train, y_train)
```

Out[165]:

```
▾ DecisionTreeClassifier
DecisionTreeClassifier(criterion='entropy', max_depth=6)
```

In [166...

```
yhattree = loantree.predict(X_test)
print(yhattree[0:5])
print(y_test[0:5])
```

```
['PAIDOFF' 'PAIDOFF' 'PAIDOFF' 'PAIDOFF' 'PAIDOFF']
['PAIDOFF' 'PAIDOFF' 'PAIDOFF' 'PAIDOFF' 'PAIDOFF']
```

In [167...

```
print(yhattree[0:20])
print(y_test[0:20])
```

```
['PAIDOFF' 'PAIDOFF' 'PAIDOFF' 'PAIDOFF' 'PAIDOFF' 'PAIDOFF' 'COLLECTION'
 'COLLECTION' 'PAIDOFF' 'PAIDOFF' 'PAIDOFF' 'PAIDOFF' 'PAIDOFF' 'PAIDOFF'
 'PAIDOFF' 'PAIDOFF' 'PAIDOFF' 'PAIDOFF' 'PAIDOFF' 'PAIDOFF']
['PAIDOFF' 'PAIDOFF' 'PAIDOFF' 'PAIDOFF' 'PAIDOFF' 'PAIDOFF' 'COLLECTION'
 'PAIDOFF' 'PAIDOFF' 'PAIDOFF' 'COLLECTION' 'COLLECTION' 'PAIDOFF'
 'PAIDOFF' 'PAIDOFF' 'PAIDOFF' 'PAIDOFF' 'PAIDOFF' 'PAIDOFF' 'PAIDOFF']
```

In [168...

```
print("The accuracy of the decision tree:", metrics.accuracy_score(y_test, yhattree))
```

```
The accuracy of the decision tree: 0.7714285714285715
```

Support Vector Machine

In [169…
```
from sklearn import svm
svmmodel = svm.SVC(kernel = 'rbf')
svmmodel.fit(X_train, y_train)
```

Out[169]:
```
▾ SVC
SVC()
```

In [170…
```
yhatsvm = svmmodel.predict(X_test)
```

In [171…
```
from sklearn.metrics import f1_score
print("f1 score:", f1_score(y_test, yhatsvm, average='weighted'))
from sklearn.metrics import jaccard_score
print("jaccard score:", jaccard_score(y_test, yhatsvm, pos_label = 'PAIDOFF'))
```

```
f1 score: 0.7275882012724117
jaccard score: 0.7272727272727273
```

Logistic Regression

In [172…
```
from sklearn.linear_model import LogisticRegression

logreg = LogisticRegression(C = 0.7, solver = 'liblinear').fit(X_train, y_train)
```

In [173…
```
yhatlogreg = logreg.predict(X_test)
yhat_prob = logreg.predict_proba(X_test)
```

In [174…
```
print("jaccard score:", jaccard_score(y_test, yhatlogreg, pos_label = 'PAIDOFF'))
from sklearn.metrics import log_loss
print("Logarithmic Loss:", log_loss(y_test, yhat_prob))
from sklearn.metrics import classification_report
print("Calssification Report:")
print(classification_report(y_test, yhatlogreg))
```

```
jaccard score: 0.7205882352941176
Logarithmic Loss: 0.497688785268822663
Calssification Report:
              precision    recall  f1-score   support

  COLLECTION       0.25      0.13      0.17        15
     PAIDOFF       0.79      0.89      0.84        55

    accuracy                           0.73        70
   macro avg       0.52      0.51      0.51        70
weighted avg       0.67      0.73      0.70        70
```

Model Evaluation using Test set

In [160…
```
from sklearn.metrics import jaccard_score
from sklearn.metrics import f1_score
from sklearn.metrics import log_loss
```

First, download and load the test set:

In [ ]:
```
!wget -O loan_test.csv https://s3-api.us-geo.objectstorage.softlayer.net/cf-courses-data
```

Load Test set for evaluation

In [159…
```
test_df = pd.read_csv('loan_test.csv')
test_df.head()
```

Out[159]:

	Unnamed: 0.1	Unnamed: 0	loan_status	Principal	terms	effective_date	due_date	age	education	Gender
0	1	1	PAIDOFF	1000	30	9/8/2016	10/7/2016	50	Bechalor	female
1	5	5	PAIDOFF	300	7	9/9/2016	9/15/2016	35	Master or Above	male
2	21	21	PAIDOFF	1000	30	9/10/2016	10/9/2016	43	High School or Below	female
3	24	24	PAIDOFF	1000	30	9/10/2016	10/9/2016	26	college	male
4	35	35	PAIDOFF	800	15	9/11/2016	9/25/2016	29	Bechalor	male

In [175…
```
print("Jaccard-score for KNN", jaccard_score(y_test, yhat, pos_label = 'PAIDOFF'))
print("Jaccard-score for Decision Tree", jaccard_score(y_test, yhattree, pos_label = 'PA
print("Jaccard-score for SVM", jaccard_score(y_test, yhatsvm, pos_label = 'PAIDOFF'))
print("Jaccard-score for Losigstic Regression", jaccard_score(y_test, yhatlogreg, pos_la
```

```
Jaccard-score for KNN 0.7424242424242424
Jaccard-score for Decision Tree 0.7681159420289855
Jaccard-score for SVM 0.7272727272727273
Jaccard-score for Losigstic Regression 0.7205882352941176
```

In [176…
```
print("F1-score for KNN", f1_score(y_test, yhat, average='weighted'))
print("F1-score for Decision Tree", f1_score(y_test, yhattree, average='weighted'))
print("F1-score for SVM", f1_score(y_test, yhatsvm, average='weighted'))
print("F1-score for Losigstic Regression", f1_score(y_test, yhatlogreg, average='weighte
```

```
F1-score for KNN 0.7381366459627329
F1-score for Decision Tree 0.7064793130366899
F1-score for SVM 0.7275882012724117
F1-score for Losigstic Regression 0.6953867388649997
```

In [177…
```
print("Logarithmic Loss for Logistic Regression", log_loss(y_test, yhat_prob))
```

```
Logarithmic Loss for Logistic Regression 0.49768878526822663
```

Report

You should be able to report the accuracy of the built model using different evaluation metrics:

Algorithm	Jaccard	F1-score	LogLoss
KNN	0.724	0.738	NA
Decision Tree	0.768	0.706	NA
SVM	0.727	0.7275	NA
LogisticRegression	0.7205	0.695	0.49768

Want to learn more?

IBM SPSS Modeler is a comprehensive analytics platform that has many machine learning algorithms. It has been designed to bring predictive intelligence to decisions made by individuals, by groups, by systems – by your enterprise as a whole. A free trial is available through this course, available here: SPSS Modeler

Also, you can use Watson Studio to run these notebooks faster with bigger datasets. Watson Studio is IBM's leading cloud solution for data scientists, built by data scientists. With Jupyter notebooks, RStudio, Apache Spark and popular libraries pre-packaged in the cloud, Watson Studio enables data scientists to collaborate on their projects without having to install anything. Join the fast-growing community of Watson Studio users today with a free account at Watson Studio

Thanks for completing this lesson!

Author: Saeed Aghabozorgi

Saeed Aghabozorgi, PhD is a Data Scientist in IBM with a track record of developing enterprise level applications that substantially increases clients' ability to turn data into actionable knowledge. He is a researcher in data mining field and expert in developing advanced analytic methods like machine learning and statistical modelling on large datasets.

Change Log

Date (YYYY-MM-DD)	Version	Changed By	Change Description
2020-10-27	2.1	Lakshmi Holla	Made changes in import statement due to updates in version of sklearn library
2020-08-27	2.0	Malika Singla	Added lab to GitLab
