سوال هفتم

نمونه داده های فایل US Presidential Data را باز کنید. این فایل حاوی آمار برد و باخت انتخابات ریاست جمهوری آمریکا است که دارای دو دسته و ۱۳ ویژگی است. ۱۳ ویژگی شامل موارد زیر در متن سخنرانی انتخاباتی است:

۱. نسبت کلمات نشان دهنده:
   خوش بینی
   بدبینی
   گذشته
   حال
   آینده
۲. تعداد دفعاتی که نامزد از حزب خود نام میبرد
۳. تعداد دفعاتی که نامزدی از حزب رقیب نام میبرد
۴. معیاری از محتوای نشان دهنده:
   صراحت
   وجدان
   برون گرایی
   موافق بودن
   عصبیت
   احساسی بودن

داده ها را به مجموعه Test و Train تقسیم کنید

a

داده های مجموعه Train را با الگوریتم 1NN دسته بندی کنید. خطای دسته بندی بر روی داده Train را گزارش دهید.

خطای محاسبه شده صفر است. با توجه به اینکه تنها یک همسایگی برای این دسته بندی انتخاب میشود در نتیجه دقت باید ۱۰۰ درصد باشد. البته ممکن است که دقت روی داده های تست ، مانند داده های آموزشی مناسب نباشد. به همین دلیل است که خطای دسته بند روی داده های تست حدود ۳۰ درصد بدست آمده است.

b

بخش ۱ را با الگوریتم KNN با مقادیر مختلف برای k امتحان کنید. نمودار خطای الگوریتم را بر اساس kهای مختلف به دست آورید. تاثیر مقدار k را در دقت الگوریتم KNN چگونه ارزیابی میکنید؟

Run_Me

فراخوانی کتابخانه های مورد نیاز

```
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import classification_report, confusion_matrix
import matplotlib.pyplot as plt
```

خواندن دیتاست

```
data_pd = pd.read_csv('US Presidential Data.csv')
```

مخلوط کردن داده ها

```
data_pd = data_pd.sample(frac=1)
```

تقسیم بندی داده به دو دسته آموزشی و آزمایشی

```
train, test = train_test_split(data_pd, test_size=0.2)
```

جداسازی داده ها و برچسب ها

```
X_tr = train[train.columns[1:train.shape[1]]]
Y_tr = train[train.columns[0]]
X_ts = test[test.columns[1:test.shape[1]]]
Y_ts = test[test.columns[0]]
```

ساخت دسته بندی نزدیک ترین همسایگی

```
classifier = KNeighborsClassifier(n_neighbors=1)
```

آموزش دسته بند

```
classifier.fit(X_tr,Y_tr)
```

ارزیابی دسته بند روی داده های آموزشی

```
y_pred = classifier.predict(X_tr)
```

نمایش ماتریس تداخل و خطای دسته بندی روی داده های آموزشی

```
print(confusion_matrix(Y_tr, y_pred))
print("Classfication Error rate = {}".format(np.mean(y_pred != Y_tr)*100))
```

بررسی تاثیر مقدار k روی خطای دسته بندی

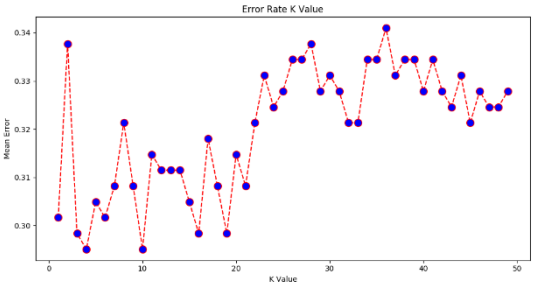
```
error = []

# Calculating error for K values between 1 and 50
Max_K = 50
for i in range(1, Max_K):
    knn = KNeighborsClassifier(n_neighbors=i)
    knn.fit(X_tr, Y_tr)
    pred_i = knn.predict(X_ts)
    error.append(np.mean(pred_i != Y_ts))


plt.figure(figsize=(12, 6))
plt.plot(range(1, Max_K), error, color='red', linestyle='dashed', marker='o',
         markerfacecolor='blue', markersize=10)
```

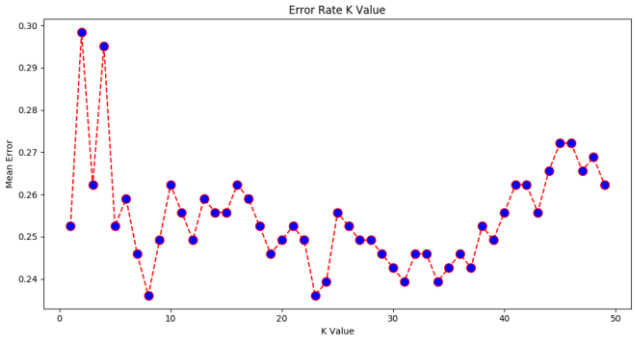
```
plt.title('Error Rate K Value')
plt.xlabel('K Value')
plt.ylabel('Mean Error')
plt.show()
```

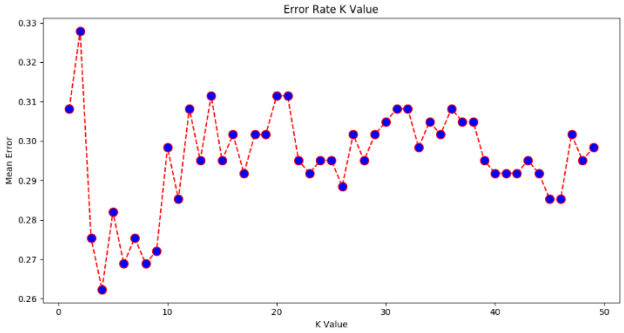
Output

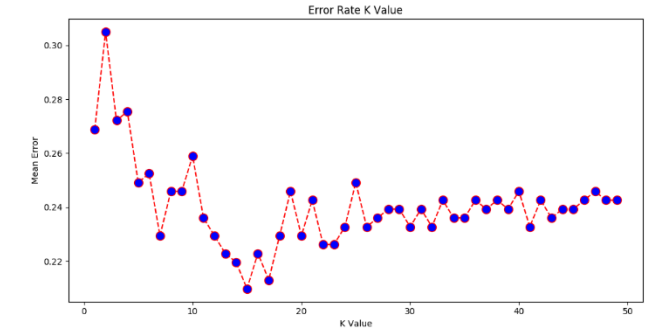
```
[[479   0]
 [  0 740]]
Train Classfication Error rate = 0.0
Test Classfication Error rate = 29.508196721311474
```

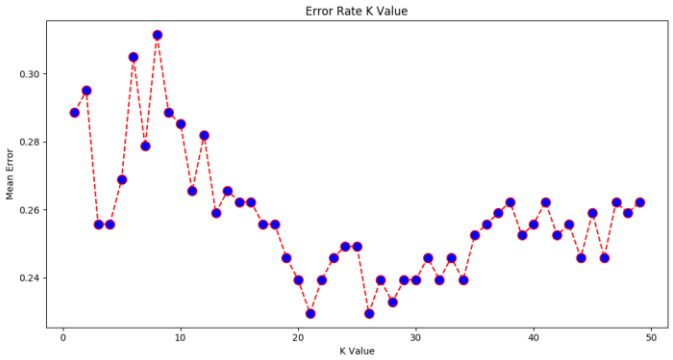
باید متذکر شد که با توجه به اینکه داده ها در هر مرحله مخلوط میشوند، در نتیجه نمودارهای یکسانی در هر آزمایش نخواهیم داشت. نمونه ای از نمودارها را در زیر مشاده میفرمایید.











تنها نقطه مشترک در نمودارهای بالا این است که مقدار K بین ۱ تا ۴ نمیتواند مناسب باشد و خطای بالایی دارد. از طرفی مشاهده میشود که در اکثر آزمایشها مقادیر بین ۱۰ تا ۲۰ ، خطای کمتری را داشته اند. اما مقادیر بالاتر از ۲۰ ، این دقت را تا حد زیادی بالا برده اند. این موضوع نشان میدهد که مقدار K ، نه باید زیاد کوچک و نه زیاد بزرگ باشد. مقدار متعادل K یک چالش است.