# گزارش فاز دوم پروژه‌ی بازیابی اطلاعات

میلاد آقاجوهری ، شماره دانشجویی: ۹۴۱۰۵۴۷۴

۹ آذر ۱۳۹۷

## ۱ گزارش نتایج

همانطور که در ابتدای pdf تمرین خواسته شده است در این قسمت به گزارش نتایج می‌پردازیم. این‌ها نتایج بر روی داده‌های جدید هستند.

### ۱.۱ Naive Bayes

نتایج این قسمت به شرح زیر هستند:

```
Accuracy is 0.781.
Category0 : Precision is 0.76 & Rcall is 0.76.
Category1 : Precision is 0.882 & Rcall is 0.853.
Category2 : Precision is 0.780 & Rcall is 0.686.
Category3 : Precision is 0.716 & Rcall is 0.826.
```

### ۲.۱ KNN

نتایج این قسمت به شرح زیر هستند:

```
For K = 1:
Accuracy is 0.835.
Category0 : Precision is 0.78 & Rcall is 0.826 & F1 is 0.803.
Category1 : Precision is 0.913 & Rcall is 0.881 & F1 is 0.8967.
Category2 : Precision is 0.831 & Rcall is 0.841 & F1 is 0.836.
Category3 : Precision is 0.824 & Rcall is 0.790 & F1 is 0.807.
Overall F1 score is 0.836.
For K = 5:
Accuracy is 0.856.
Category0 : Precision is 0.835 & Rcall is 0.878 & F1 is 0.85625.
```

```
Category1 : Precision is 0.882 & Rcall is 0.944 & F1 is 0.912.
Category2 : Precision is 0.828 & Rcall is 0.854& F1 is 0.841.
Category3 : Precision is 0.891 & Rcall is 0.748 & F1 is 0.813.
Overall F1 score is 0.855.
For K = 10:
Accuracy is 0.863.
Category0 : Precision is 0.875 & Rcall is 0.858 & F1 is 0.867.
Category1 : Precision is 0.925 & Rcall is 0.958 & F1 is 0.941.
Category2 : Precision is 0.809 & Rcall is 0.860 & F1 is 0.834.
Category3 : Precision is 0.847 & Rcall is 0.776 & F1 is 0.810.
Overall F1 score is 0.863.
```

و هم‌چنین حال در مورد بهترین پارامتر داریم که نتایج روی داده‌های تست این هستند:

```
Accuracy is 0.781.
Category0 : Precision is 0.76 & Rcall is 0.76.
Category1 : Precision is 0.882 & Rcall is 0.853.
Category2 : Precision is 0.780 & Rcall is 0.686.
Category3 : Precision is 0.716 & Rcall is 0.826.
```

### SVM ۳.۱

نتایج این قسمت به شرح زیر هستند(قسمت پارامترهای مختلف):

```
C = 2:
Accuracy is 0.87.
Category0 : Precision is 0.838 & Rcall is 0.844 & F1 is 0.841.
Category1 : Precision is 0.922& Rcall is 0.962 & F1 is 0.942.
Category2 : Precision is 0.855& Rcall is 0.840 & F1 is 0.8477611940298507.
Category3 : Precision is 0.853 & Rcall is 0.822 & F1 is 0.837
Overall F1 score is 0.867.
C = 1.5:
Accuracy is 0.875.
Category0 : Precision is 0.857& Rcall is 0.844 & F1 is 0.850.
Category1 : Precision is 0.928 & Rcall is 0.962 & F1 is 0.945.
Category2 : Precision is 0.857 & Rcall is 0.857& F1 is 0.857.
Category3 : Precision is 0.847 & Rcall is 0.822& F1 is 0.834.
Overall F1 score is 0.872.
C = 1:
Accuracy is 0.876.
Category0 : Precision is 0.870 & Rcall is 0.844& F1 is 0.857.
```

```
Category1 : Precision is 0.928 & Rcall is 0.962 & F1 is 0.945.
Category2 : Precision is 0.852 & Rcall is 0.857 & F1 is 0.855.
Category3 : Precision is 0.848& Rcall is 0.829& F1 is 0.838.
Overall F1 score is 0.874.
C = 0.5:
Accuracy is 0.881.
Category0 : Precision is 0.898 & Rcall is 0.851& F1 is 0.874.
Category1 : Precision is 0.928 & Rcall is 0.962 & F1 is 0.945.
Category2 : Precision is 0.863 & Rcall is 0.863& F1 is 0.863.
Category3 : Precision is 0.830 & Rcall is 0.837 & F1 is 0.833.
Overall F1 score is 0.879.
```

همانطور که مشاهده می‌شود، بهترین نتیجه را برای $C = ۰/۵$ داریم. پس حال بر روی داده‌های test آن را می‌سنجیم.

```
C = 0.5:
Accuracy is 0.843.
Category0 : Precision is 0.865 & Rcall is 0.86 & F1 is 0.862.
Category1 : Precision is 0.898& Rcall is 0.94 & F1 is 0.918.
Category2 : Precision is 0.7625 & Rcall is 0.813 & F1 is 0.787.
Category3 : Precision is 0.850 & Rcall is 0.76 & F1 is 0.802.
Overall F1 score is 0.842.
```

### Random Forest ۴.۱

```
Accuracy is 0.738.
Category0 : Precision is 0.788 & Rcall is 0.746 & F1 is 0.767.
Category1 : Precision is 0.818 & Rcall is 0.873 & F1 is 0.845.
Category2 : Precision is 0.720 & Rcall is 0.653 & F1 is 0.685.
Category3 : Precision is 0.629& Rcall is 0.68 & F1 is 0.653.
Overall F1 score is 0.737.
```

## ۲ توضیح Random Forest

به صورت بسیار کلی می‌توان گفت که الگوریتم جنگل تصادفی درخت تصمیم‌های مختلفی را می‌سازد و سپس در هر بار تصمیم روی تمام درخت‌ها اجرا می‌کند و بیشترین تکرار را خروجی می‌دهد. در واقع اگر بخواهیم دقیق‌تر توضیح بدهیم روش این الگوریتم به این صورت است که به هر درخت یک زیرمجموعه از داده‌ها را می‌دهد و هم‌چنین در ساختن هر شاخه‌ی از درخت به جای درخت تصمیم که تمام ویژگی‌ها را بررسی می‌کند تا بهترین ویژگی را بیابد، زیرمجموعه‌ی تصادفی‌ای از ویژگی‌ها را بررسی می‌کند. سپس برای هر task طبقه‌بندی این طبقه‌بندی را توسط این درخت‌های کوچک انجام

می‌دهد و بعد از آن در بین خروجی این درختان مختلف از نحوه‌ای روش تجمیع نظرات استفاده می‌کند که در واقع در اکثر موارد گرفتن mode بین خروجی این درخت‌های مختلف است.