**به نام خدا**

# پروژه دوم درس یادگیری ماشین

## دسته‌بند و خوشه‌بند برای داده‌های برگ گیاهان

**استاد درس: دکتر محمد کیانی**

**رضا اعلائی**

**نیم‌سال دوم سال تحصیلی ۰۳ – ۱۴۰۲**

**دانشگاه اصفهان**

**دانشکده مهندسی کامپیوتر**

# فهرست

# مقدمه و اطلاعات مجموعه داده‌ای

در این پروژه، قصد ما ارائه‌ی یک خوشه‌بند و یک دسته بند برای خوشه‌بندی و دسته‌بندی اطلاعات دیتاست leaves است.

ستون اول در فایل leaves.csv به‌عنوان برچسب در عملیات دسته‌بندی است.
همچنین این ستون برای دقت سنجی در عملیات خوشه‌بندی نیز مورد استفاده قرار گرفته است.

# دسته‌بندی

## عملیات‌های انجام شده بر روی داده‌ها

### مدیریت مقادیر نامعلوم

برای اطمینان از اینکه مقداری نامعلوم در مجموعه‌ی داده‌ای وجود نداشته باشد این عملیات انجام شده است. استراتژی جایگزینی توسط simple imputer در این مدل، جایگزین کردن مقدار پرتکرار به جای مقدار نامعلوم است.

### مقیاس‌بندی ویژگی‌ها

برای هرچه بهتر شدن نتایج پیش‌بینی مدل، داده‌ها مقیاس‌بندی شده‌اند. Scaler استفاده شده در این مدل، standard scaler می‌باشد.

## انتخاب ویژگی‌ها و مهندسی آن‌ها

برای این کار از LDA با سیزده مولفه استفاده شده است.

همچنین در نهایت، ده ویژگی برتر برای عملیات دسته‌بندی انتخاب می‌شوند.

## انتخاب نوع مدل و شرح آن

در ابتدا، این پروژه به تمامی الگوریتم‌های معرفی شده در دوره اعم از: knn، svm، درخت تصمیم، جنگل تصادفی، adaboost، naive bayes، پرسپترون چند لایه و logistic regression داده شد.

برای هر کدام، درصد دقت و معیارهای دیگر مشاهده شد. با آزمون و خطا سعی بر این شد که نتیجه هر مدل را بهبود ببخشیم. مدلی که بیشترین بهبود در نتایج را داشت مدل MLP بود، همچنین این مدل پس از بهبود بالاترین دقت را نیز بدست آورد.

در نهایت این مدل به‌عنوان مدل بهتر برای انجام پروژه در نظر گرفته شد. همچنین مشخصات نهایی مدل نیز به شرح زیر است:

```
mlp_config = {
    'hidden_layer_sizes': (150,),
    'activation': 'tanh',
    'solver': 'adam',
    'alpha': 0.0001,
    'batch_size': 50,
    'learning_rate': 'invscaling',
    'learning_rate_init': 0.01,
    'max_iter': 400,
    'random_state': 42,
    'tol': 1e-4
}
```

## ارزیابی مدل

معیارهای مورد استفاده در این مدل جهت ارزیابی عبارتند از: دقت، precision، recall، f-score.

همچنین دقت cross-validation نیز اینجا برای سنجش در نظر گرفته شده.

نتایج نهایی مدل به این صورت است:

Accuracy: 0.8824

Classification Report:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 1 | 1.00 | 1.00 | 1.00 | 4 |
| 2 | 0.80 | 1.00 | 0.89 | 4 |
| 3 | 1.00 | 1.00 | 1.00 | 2 |
| 4 | 1.00 | 1.00 | 1.00 | 1 |
| 5 | 0.60 | 0.75 | 0.67 | 4 |
| 6 | 1.00 | 1.00 | 1.00 | 2 |
| 7 | 1.00 | 0.33 | 0.50 | 3 |
| 8 | 1.00 | 0.86 | 0.92 | 7 |
| 9 | 0.67 | 0.67 | 0.67 | 3 |
| 10 | 1.00 | 1.00 | 1.00 | 3 |
| 11 | 1.00 | 1.00 | 1.00 | 8 |
| 12 | 0.50 | 1.00 | 0.67 | 2 |
| 13 | 0.88 | 1.00 | 0.93 | 7 |
| 14 | 1.00 | 0.33 | 0.50 | 3 |
| 15 | 1.00 | 1.00 | 1.00 | 1 |
| 22 | 0.80 | 1.00 | 0.89 | 4 |
| 23 | 1.00 | 1.00 | 1.00 | 1 |
| 24 | 0.83 | 1.00 | 0.91 | 5 |
| 25 | 1.00 | 1.00 | 1.00 | 4 |
| 26 | 1.00 | 0.50 | 0.67 | 2 |
| 27 | 1.00 | 0.50 | 0.67 | 2 |
| 28 | 0.33 | 1.00 | 0.50 | 1 |
| 29 | 1.00 | 1.00 | 1.00 | 3 |
| 30 | 1.00 | 0.67 | 0.80 | 3 |
| 31 | 1.00 | 1.00 | 1.00 | 4 |

```
          32       0.50      1.00      0.67         1
          33       1.00      1.00      1.00         6
          34       1.00      1.00      1.00         4
          35       1.00      0.50      0.67         4
          36       1.00      1.00      1.00         4

    accuracy                           0.88       102
   macro avg       0.90      0.87      0.85       102
weighted avg       0.92      0.88      0.88       102

Cross-Validation Accuracy: 0.8231 ± 0.0332
```

# خوشه‌بندی

## عملیات‌های انجام شده بر روی داده‌ها

### مقیاس‌بندی ویژگی‌ها

برای هرچه بهتر شدن نتایج پیش‌بینی مدل، داده‌ها مقیاس‌بندی شده‌اند. Scaler استفاده شده در این مدل، standard scaler می‌باشد.

## انتخاب ویژگی‌ها و مهندسی آن‌ها

در اینجا، ۵ ویژگی برتر برای عملیات خوشه‌بندی انتخاب می‌شوند.

## انتخاب نوع مدل و شرح آن

مانند قسمت قبل، در اینجا نیز داده‌ها به چند خوشه‌بند داده شدند و در نهایت نتایج خوشه بند انتخابی بهبود بخشیده شد. خوشه‌بندها عبارتند از: k-means، FCM، agglomorative و guassian mixture.

درنهایت با توجه به معیارهای ارزیابی ابتدایی، خوشه بند agglomorative انتخاب شد.

## ارزیابی مدل

معیارهای ارزیابی در اینجا، معیار Silhouette در کنار نمودار مربوط به این خوشه‌بند معیارهای اصلی هستند که در ادامه آمده است. همچنین معیارهای دیگری نیز برای دقت سنجی این خوشه‌بند مورد استفاده قرار گرفته‌اند که در ادامه آمده‌اند.

```
Silhouette Score: 0.6311
Adjusted Rand Index: 0.0305
Adjusted Mutual Information: 0.2663
Cluster Purity: 0.2948
```

نمودار مربوط به خوشه‌بند: