25.12.1403.2

Load Wine

این برنامه یک پروژه یادگیری ماشین است که از الگوریتم **K-Nearest Neighbors (KNN)** برای طبقه‌بندی داده‌های مجموعه‌داده شراب (**Wine Dataset**) استفاده می‌کند. هدف اصلی این برنامه بررسی دقت مدل KNN با توجه به ویژگی‌های مختلف و اندازه‌های مختلف مجموعه تست است. در ادامه، هر بخش از کد به طور کامل و با جزئیات توضیح داده می‌شود:

---

۱. وارد کردن کتابخانه‌های لازم

- **numpy**: برای انجام محاسبات عددی و کار با آرایه‌ها.
- **pandas**: برای کار با داده‌ها در قالب دیتافریم.
- **sklearn**: برای پیاده‌سازی مدل‌های یادگیری ماشین، تقسیم داده‌ها، استانداردسازی و محاسبه دقت.
- **scipy.stats.mode**: برای محاسبه مُد (مقدار پرتکرار) در آرایه‌ها.
- **matplotlib.pyplot**: برای رسم نمودارها.
- **seaborn**: برای زیباسازی نمودارها.(set_style('darkgrid'))
- **time**: برای اندازه‌گیری زمان اجرای کد.

---

۲. تعریف تابع calculate_accuracy

این تابع دقت مدل KNN را محاسبه می‌کند. مراحل آن به شرح زیر است:

- **ورودی‌ها:**
  - dataset: مجموعه‌داده ورودی (دیتافریم).
  - test_size: اندازه مجموعه تست (به صورت درصدی از کل داده‌ها).
  - features_to_use: لیست اندیس‌های ویژگی‌هایی که باید در مدل استفاده شوند.
- **مراحل اجرا:**
  ۱. **جداسازی ویژگی‌ها و برچسب‌ها:**
     - X: ویژگی‌های مجموعه‌داده (بدون ستون target).
     - y: برچسب‌های مجموعه‌داده (ستون target).
  ۲. **تقسیم داده‌ها به مجموعه‌های آموزشی و تست:**
     - داده‌ها به دو بخش train و test تقسیم می‌شوند. اندازه مجموعه تست بر اساس ورودی test_size تعیین می‌شود.
  ۳. **استانداردسازی ویژگی‌ها:**
     - ویژگی‌ها با استفاده از StandardScaler استانداردسازی می‌شوند تا میانگین صفر و واریانس یک داشته باشند.
  ۴. **محاسبه تعداد همسایه‌ها (k):**
     - تعداد همسایه‌ها (k) به عنوان جذر تعداد نمونه‌های آموزشی محاسبه می‌شود.
  ۵. **آموزش مدل KNN و پیش‌بینی:**
     - برای هر ویژگی مشخص شده(features_to_use) ، مدل KNN آموزش داده می‌شود و پیش‌بینی‌ها انجام می‌شوند.
     - پیش‌بینی‌ها برای هر ویژگی در لیست predicted_labels_per_features ذخیره می‌شوند.
  ۶. **ترکیب پیش‌بینی‌ها:**
     - پیش‌بینی‌های تمام ویژگی‌ها با هم ترکیب شده و مُد (مقدار پرتکرار) به عنوان پیش‌بینی نهایی انتخاب می‌شود.
  ۷. **محاسبه دقت:**
     - دقت مدل (overall_accuracy) با مقایسه پیش‌بینی‌های نهایی با برچسب‌های واقعی (y_test) محاسبه می‌شود.
     - همچنین، دقت مدل زمانی که از تمام ویژگی‌ها استفاده می‌شود (accuracy1) نیز محاسبه می‌شود.

- **خروجی:**
  - accuracy1: دقت مدل زمانی که از تمام ویژگی‌ها استفاده می‌شود.
  - overall_accuracy: دقت مدل زمانی که از ویژگی‌های انتخابی استفاده می‌شود.

---

۳ .بارگذاری مجموعه‌داده شراب

- مجموعه‌داده شراب (load_wine) بارگذاری می‌شود. این مجموعه‌داده شامل ۱۳ ویژگی (مانند اسیدیته، الکل، فلاونوئیدها و ...) و ۳ کلاس (نوع شراب) است.
- داده‌ها به یک دیتافریم (df) تبدیل می‌شوند تا پردازش آن‌ها آسان‌تر شود.

---

۴ .تعریف تابع run_knn

این تابع مدل KNN را برای تعداد مشخصی از تکرارها اجرا می‌کند. مراحل آن به شرح زیر است:

- **ورودی‌ها:**
  - iterations: تعداد تکرارهایی که مدل باید اجرا شود.
- **مراحل اجرا:**

۱. **تکرار مدل:**
   - در هر تکرار، اندازه مجموعه تست (test_size) به صورت تصادفی بین ۱۰٪ تا ۳۰٪ انتخاب می‌شود.
   - تابع calculate_accuracy فراخوانی می‌شود و دقت مدل محاسبه می‌شود.

۲. **شمارش ویژگی‌های مهم:**
   - اگر دقت مدل در یک تکرار بیشتر از ۷۰٪ باشد، ویژگی‌هایی که در آن تکرار استفاده شده‌اند، شمارش می‌شوند.

- **خروجی:**
  - all_accuracy_results: لیستی از دقت‌های مدل در تمام تکرارها.
  - feature_counts: تعداد دفعاتی که هر ویژگی در تکرارهایی با دقت بالای ۷۰٪ استفاده شده است.

---

۵ .دریافت تعداد تکرار از کاربر

- کاربر تعداد تکرارها را وارد می‌کند. این تعداد تعیین می‌کند که مدل چند بار اجرا شود.

---

۶ .اجرا و دریافت نتایج

- تابع run_knn اجرا می‌شود و نتایج all_accuracy_results) و (feature_counts دریافت می‌شوند.

---

۷ .جمع‌آوری ویژگی‌هایی که دقتشان بالای ۷۰ درصد بوده است

- ویژگی‌هایی که در تمام تکرارها دقت بالای ۷۰٪ داشته‌اند، شناسایی می‌شوند.
- اندیس‌های این ویژگی‌ها در لیست features_indices ذخیره می‌شوند.

---

۸ .بررسی دقت‌ها در درصدهای مختلف

- دقت مدل برای اندازه‌های مختلف مجموعه تست (از ۱۱٪ تا ۲۵٪ با گام ۲٪) محاسبه می‌شود.
- دقت‌ها در لیست‌های accuracy1_list و overall_accuracy_list ذخیره می‌شوند.

---

۹ .زمان‌سنجی کلی برای هر دو روش

- زمان اجرای کل روش اول (Accuracy1) و روش دوم (Overall Accuracy) به‌صورت جداگانه اندازه‌گیری می‌شود.
- زمان‌های اجرای کلی در خروجی نمایش داده می‌شوند.

---

۱۰ .رسم نمودار

- هر دو نمودار Accuracy1) و (Overall Accuracy در یک شکل رسم می‌شوند تا مقایسه آن‌ها آسان‌تر باشد.
- نمودار شامل دو خط است:
  ۱. **خط آبی** :مربوط به) Accuracy1 دقت زمانی که از تمام ویژگی‌ها استفاده می‌شود.(
  ۲. **خط نارنجی** :مربوط به) Overall Accuracy دقت زمانی که از ویژگی‌های انتخابی استفاده می‌شود.(

---

۱۱ .نمایش زمان‌های اجرای کلی

- زمان کل اجرای هر دو روش در خروجی نمایش داده می‌شود.
- مثال:

```
Copy

Total time for Accuracy1: 0.1234 seconds
Total time for Overall Accuracy: 0.1456 seconds
```

---

۱۲ .خروجی برنامه

- **نمودار** :نمودار مقایسه دقت‌ها برای اندازه‌های مختلف مجموعه تست.
- **زمان‌های اجرای کلی** :زمان کل اجرای هر دو روش.

جمع‌بندی

این برنامه به طور سیستماتیک دقت مدل KNN را برای مجموعه‌داده شراب بررسی می‌کند. با استفاده از این برنامه، می‌توان فهمید که:

- کدام ویژگی‌ها بیشترین تأثیر را بر دقت مدل دارند.
- چگونه اندازه مجموعه تست بر دقت مدل تأثیر می‌گذارد.
- آیا استفاده از زیرمجموعه‌ای از ویژگی‌ها می‌تواند دقت مدل را بهبود بخشد یا خیر.