تشخیص دیابت با استفاده ازSVM

محمدرضا تاجیک

400521198

هدف پروژه

هدف از این پروژه، آشنایی با نحوه استفاده از ماشین بردار پشتیبانی (SVM) برای دسته‌بندی داده‌های پزشکی و تشخیص بیماری دیابت است.

توضیحات پروژه

این پروژه شامل مراحل زیر است:

1. بارگذاری دیتاستPima Indians Diabetes
2. تقسیم داده‌ها به داده‌های آموزشی و تست
3. استانداردسازی داده‌ها
4. آموزش مدل SVM با کرنلRBF
5. ارزیابی مدل و گزارش معیارهایی مانند دقت، فراخوانی وF1-Score
6. رسم نمودار ماتریس سردرگمی وROC

بارگذاری دیتاست

دیتاست Pima Indians Diabetes شامل اطلاعات پزشکی بیماران و تشخیص بیماری دیابت است.

توضیح مراحل پروژه

1. بارگذاری دیتاست:
   - ابتدا دیتاست به صورت فایل CSV بارگذاری می‌شود و به یک DataSet در pandas تبدیل می‌گردد.
2. تقسیم داده‌ها به داده‌های آموزشی و تست:
   - داده‌ها به دو مجموعه‌ی آموزشی و تست تقسیم می‌شوند. این تقسیم به گونه‌ای است که 80 درصد داده‌ها برای آموزش و 20 درصد برای تست استفاده می‌شوند.
3. استانداردسازی داده‌ها:
   - برای بهبود عملکرد مدلSVM ، ویژگی‌ها استانداردسازی می‌شوند. استانداردسازی به معنای مقیاس‌بندی ویژگی‌ها به گونه‌ای است که میانگین آن‌ها صفر و انحراف معیار آن‌ها یک شود.
4. آموزش مدل SVM با کرنل:RBF
   - یک مدل SVM با کرنل RBF (Radial Basis Function) آموزش داده می‌شود. این کرنل به دلیل انعطاف‌پذیری بالا در تفکیک داده‌ها استفاده می‌شود.
5. ارزیابی مدل:
   - مدل آموزش دیده با استفاده از داده‌های تست ارزیابی می‌شود. معیارهای دقت، فراخوانی، دقت و F1-Score برای ارزیابی عملکرد مدل محاسبه و گزارش می‌شوند.
   - دقت (Accuracy) نشان‌دهنده نسبت کل پیش‌بینی‌های صحیح به تعداد کل نمونه‌ها است.
   - دقت (Precision) نشان‌دهنده نسبت نمونه‌های صحیحاً مثبت پیش‌بینی‌شده به کل نمونه‌های پیش‌بینی‌شده به عنوان مثبت است.
   - فراخوانی (Recall) نشان‌دهنده نسبت نمونه‌های صحیحاً مثبت پیش‌بینی‌شده به کل نمونه‌های واقعاً مثبت است.
   - F1-Scoreمیانگینی از دقت و فراخوانی است و توازنی بین این دو ایجاد می‌کند.
6. رسم نمودار ماتریس سردرگمی وROC:
   - ماتریس سردرگمی برای نشان دادن تعداد نمونه‌های صحیح و ناصحیح پیش‌بینی‌شده به کار می‌رود و به صورت یک نقشه حرارتی ترسیم می‌شود.
   - نمودار ROC (Receiver Operating Characteristic) منحنی‌ای است که عملکرد مدل را با استفاده از نسبت نرخ مثبت کاذب به نرخ مثبت حقیقی نشان می‌دهد. مقدار AUC (Area Under the Curve)نیز محاسبه می‌شود که بیانگر کیفیت مدل است؛ هرچه AUC بیشتر باشد، مدل بهتر عمل کرده است.