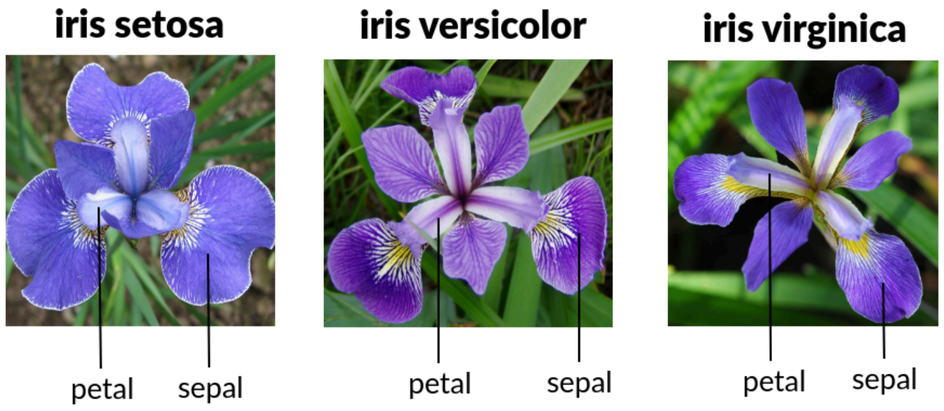
تمرین ۱۴ / ۳۵۰۰ امتیاز / DT

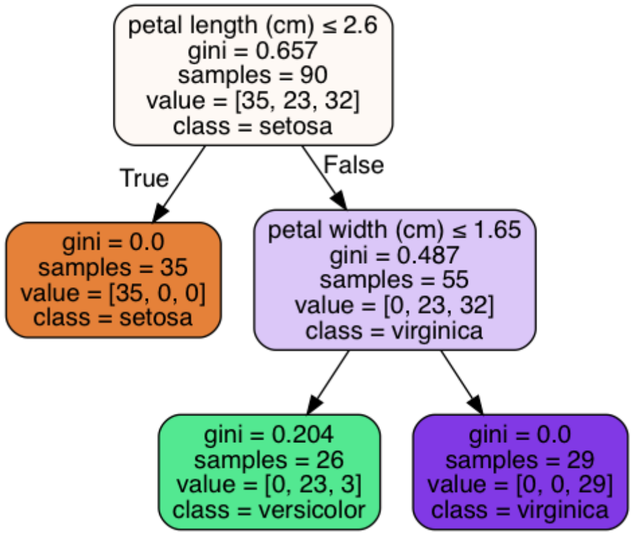


بخش یک

1. دیتاست Iris شامل چه ویژگی‌هایی است و هدف (Target) در این داده چیست؟
2. داده را از فایل iris.csv بارگذاری کنید. چند ردیف و ستون دارد؟
3. آیا دادهٔ گمشده یا مقدار غیرعادی در دیتاست وجود دارد؟ چگونه بررسی می‌کنید؟
4. ستون هدف (species) چند کلاس دارد و توزیع آن‌ها چگونه است؟
5. داده را به دو بخش X (ویژگی‌ها) و y (هدف) تقسیم کنید. چه تعداد ویژگی در X وجود دارد؟
6. داده را با train_test_split به آموزش و آزمون تقسیم کنید. چرا از stratify=y استفاده می‌کنیم؟
7. مدل پایه‌ی Decision Tree را بسازید و آن را روی دادهٔ آموزشی یاد بگیرید. دقت اولیه‌ی مدل چقدر است؟
8. از چه معیاری برای تقسیم (criterion) استفاده می‌کنید؟ تفاوت بین gini و entropy چیست؟
9. درخت تصمیم را با پارامترهای max_depth مختلف آزمایش کنید (مثلاً ۲، ۳، ۵، None). دقت در هر حالت چگونه تغییر می‌کند؟
10. با استفاده از GridSearchCV بهترین ترکیب پارامترها (criterion، max_depth، min_samples_split) را پیدا کنید.

مدرسه هوش مصنوعی فیلاگر

بخش دو



1. بهترین پارامترهای پیدا شده توسط GridSearchCV چه هستند؟
2. دقت مدل بهینه‌شده روی دادهٔ تست چقدر است؟
3. گزارش طبقه‌بندی (classification report) را چاپ کنید. کدام گونه‌ها بیشتر اشتباه پیش‌بینی شده‌اند؟
4. ماتریس خطا (confusion matrix) را رسم و تحلیل کنید. هر سطر و ستون آن چه معنایی دارد؟
5. درخت تصمیم نهایی را با استفاده از plot_tree رسم کنید. کدام ویژگی در ریشهٔ درخت قرار دارد؟
6. از نظر منطقی، آیا این ویژگی برای تفکیک گونه‌ها مناسب است؟ چرا؟
7. اگر درخت شما بیش‌برازش (Overfitting) نشان دهد، چه تنظیماتی می‌توانید انجام دهید تا درخت را ساده‌تر کنید؟
8. مفهوم Pruning (هرس کردن) در درخت تصمیم چیست و چه تأثیری بر عملکرد مدل دارد؟
9. تفاوت بین درخت تصمیم منفرد و Random Forest را توضیح دهید. چرا RF پایدارتر است؟
10. در پایان، عملکرد و مزایای درخت تصمیم را نسبت به KNN روی همین داده مقایسه کنید. در چه شرایطی یکی بهتر از دیگری است؟

مدرسه هوش مصنوعی فیلاگر