داک سیستم تشخیص چاقی با یادگیری ماشین

مراحل پیش‌پردازش داده‌ها

الف. تحلیل اولیه داده‌ها

ابتدا داده‌ها از نظر ساختاری بررسی می‌شوند تا متغیرهای عددی و دسته‌ای شناسایی شوند.

برای متغیر هدف سطح چاقی که از نوع دسته‌ای است، کدگذاری عددی انجام می‌شود تا برای مدل‌های یادگیری ماشین قابل فهم باشد.

ب. تبدیل ویژگی‌های دسته‌ای

از روش one-hot encoding برای متغیرهای دسته‌ای مانند جنسیت و سابقه خانوادگی استفاده می‌شود. این کار باعث می‌شود مدل بتواند از این اطلاعات به شکل مناسبی استفاده کند.

ویژگی‌های عددی نیز استانداردسازی می‌شوند تا همه متغیرها در یک محدوده یکسان قرار بگیرند و مدل بر اساس مقیاس متغیرها گمراه نشود.

توسعه مدل‌های یادگیری ماشین

الف. مدلXGBoost

این مدل به دلیل قدرت بالا در پردازش داده‌های جدولی انتخاب شده است.

از بهینه‌سازی بیزی برای تنظیم هایپرپارامترها استفاده می‌شود که مزایای زیر را دارد:

جستجوی هوشمندانه فضای پارامترها

نیاز به اجرای کمتر مدل برای یافتن ترکیب بهینه

در نظر گرفتن روابط بین پارامترها

پارامترهای مهمی که تنظیم می‌شوند شامل عمق درخت، نرخ یادگیری، و پارامترهای کنترل پیچیدگی مدل هستند.

ب. مدل درخت تصمیم

این مدل برای مقایسه و ارائه یک روش ساده‌تر و قابل تفسیرتر انتخاب شده است.

بهینه‌سازی پارامترها در این مدل نیز با روش بیزی انجام می‌شود.

پارامترهای اصلی که تنظیم می‌شوند شامل معیار تقسیم، حداکثر عمق و حداقل نمونه در هر گره هستند.

ارزیابی مدل‌ها

الف. معیارهای ارزیابی

از معیارهای مختلفی شامل دقت، حساسیت، دقت مثبت و ویژگی‌های ماتریس درهم‌ریختگی استفاده می‌شود.

برای مدل چندکلاسه، این معیارها به صورت میانگین وزنی محاسبه می‌شوند تا توازن بین کلاس‌ها رعایت شود.

ب. تحلیل نتایج

ماتریس درهم‌ریختگی به صورت گرافیکی نمایش داده می‌شود تا خطاهای مدل در پیش‌بینی هر کلاس به وضوح دیده شود.

اهمیت ویژگی‌ها بررسی می‌شود تا مشخص شود کدام عوامل بیشترین تاثیر را در پیش‌بینی سطح چاقی دارند.

یک نکته حائز اهمیت این است که در مراحل اولیه و در حین نمایش دادن شکل های مختلف از ویژگی ها، یک ویژگی جدید به اسم BMI به دیتافریم اضافه شد و با اینکه این ویژگی ترکیب محاسباتی از چند ویژگی دیگر بود، باز هم به طور چشمگیری روی هردو الگوریتم به ویژه درخت تصمیم اثر گذاشت.

به طوری که دقت XGBoost را در حد **دو درصد** و دقت درخت تصمیم را تا **ده درصد** افزایش داد

نتایج بدون BMI (عکس سمت راست مربوط به XGBoost است):

```
Best Decision Tree Parameters:
...
Metric          Training Set    Test Set
Accuracy        0.9562      0.8771
Recall          0.9562      0.8771
Precision       0.9563      0.8870
```

```
=== Model Evaluation Results ===
Metric          Training Set    Test Set
Accuracy        0.9751      0.9622
Recall          0.9751      0.9622
Precision       0.9751      0.9641
Specificity     -               0.9977
```

نتایج با حضور BMI :

```
Best Decision Tree Parameters:
...
Metric          Training Set    Test Set
Accuracy        0.9733      0.9764
Recall          0.9733      0.9764
Precision       0.9736      0.9767
```

```
=== Model Evaluation Results ===
Metric          Training Set    Test Set
Accuracy        0.9858      0.9882
Recall          0.9858      0.9882
Precision       0.9859      0.9882
Specificity     -               0.9977
```

دلیل این امر هم منطقی است چون BMI یک معیار اساسی در تعیین وضعیت چاقیست.

مقایسه نهایی و انتخاب مدل

الف. نقاط قوت و ضعف

مدل XGBoost دقت بالاتری دارد اما تفسیر آن پیچیده‌تر است و محاسبات بیشتری هم دارد.

مدل درخت تصمیم ساده‌تر و قابل تفسیرتر و سریعتر است اما دقت کمتری دارد.

ب. معیارهای انتخاب مدل نهایی

اگر هدف دستیابی به بالاترین دقت ممکن باشد، XGBoostانتخاب بهتری است.(که واقعا هم انتخاب بهتر همین است)

اگر نیاز به تفسیرپذیری و فهم نحوه تصمیم‌گیری مدل باشد، درخت تصمیم گزینه مناسب‌تری است.