دوره یادگیری ماشین

تمرین ۱

مجموعه داده هزینه های پزشکی

توضیحات

مجموعه داده‌های هزینه‌های پزشکی شامل ویژگی‌های جمعیت‌شناختی و مربوط به سلامتی بیماران می‌شود، مانند سن، BMI، عادت سیگار کشیدن، تعداد فرزندان، منطقه و غیره، به همراه هزینه‌های پزشکی متناظر آن‌ها.

ویژگی ها

- **Age**: سن بیمار (سال)
- **Sex**: جنسیت بیمار (مرد یا زن)
- **BMI**: شاخص توده بدن
- **Children**: تعداد فرزندان / وابستگی‌ها تحت پوشش بیمه‌ی سلامت این بیمار
- **Smoker**: عادت سیگار کشیدن بیمار (بله یا خیر)
- **Region**: منطقه جغرافیایی بیمار (شمال شرق، جنوب شرق، جنوب غرب، شمال غرب)
- **Charges**: هزینه‌های پزشکی بیمار

هدف

هدف این مجموعه داده انجام یک تسک رگرسیون است برای پیش‌بینی هزینه‌های پزشکی که بیماران براساس ویژگی‌های جمعیتی و مربوط به سلامتی‌شان متحمل می‌شوند.

منبع

می‌توانید این مجموعه داده را از اینجا دانلود کنید.

وظایف

1. **پیش‌پردازش داده**: متغیرهای زمینه‌ای را رمزگذاری کنید، مقادیر گمشده را پردازش کنید، ویژگی‌ها را در صورت لزوم نرمال کنید / استاندارد کنید.
2. **تحلیل داده‌های اکتشافی (EDA)**: تحلیل توزیع ویژگی‌ها، همبستگی‌ها و غیره.
3. **انتخاب / مهندسی ویژگی**: ویژگی‌های مرتبط را انتخاب کنید و / یا اگر لازم است ویژگی‌های جدید ایجاد کنید.
4. **انتخاب مدل**: با الگوریتم‌های رگرسیون مختلف (مانند رگرسیون خطی، رگرسیون جنگل تصادفی، افزایش گرادیان و غیره) آزمایش کنید.

5. **ارزیابی مدل**: مدل‌ها را با استفاده از معیارهای ارزیابی مناسب (مانند میانگین خطای مطلق، میانگین مربعات خطا، R-مربع) ارزیابی کنید.
6. **تنظیم پارامترهای پیش‌پردازش**: پارامترهای بهترین مدل را تنظیم کنید.
7. **انتخاب نهایی مدل**: بر اساس نتایج ارزیابی، بهترین مدل را انتخاب کنید.
8. **تفسیر مدل**: نتایج را تفسیر کرده و عوامل تأثیرگذار بر هزینه‌های پزشکی را مشخص کنید.

مجموعه داده دیابت

توضیحات

مجموعه داده دیابت شامل ویژگی‌های مختلف مرتبط با سلامتی بیماران است، مانند سطوح گلوکز، سطوح انسولین، BMI، سن و غیره، و یک متغیر هدف دودویی که نشان دهنده این است که بیمار دیابت دارد یا خیر.

ویژگی ها

- **Pregnancies**: تعداد بارهای بارداری
- **Glucose**: غلظت گلوکز پلاسمایی پس از ۲ ساعت در یک آزمون تحمل گلوکز دهانی
- **BloodPressure**: فشار خون دیاستولیک (میلی‌متر جیوه)
- **SkinThickness**: ضخامت چربی پوست سه سطحی (میلی‌متر)
- **Insulin**: انسولین سرمی ۲ ساعته (مو یو / میلی لیتر)
- **BMI**: شاخص توده بدن
- **DiabetesPedigreeFunction**: تابع موروثی دیابت (تابعی که احتمال داشتن دیابت را بر اساس تاریخچه خانوادگی امتیاز می‌دهد)
- **Age**: سن بیمار (سال)
- **Outcome**: متغیر هدف (۰ اگر غیر دیابتی باشد، ۱ اگر دیابتی باشد)

هدف

هدف این مجموعه داده انجام وظیفه‌ی طبقه‌بندی است برای پیش‌بینی این که آیا یک بیمار دیابت دارد یا خیر بر اساس ویژگی‌های مربوط به سلامتی داده شده است.

منبع

می‌توانید این مجموعه داده را از اینجا دانلود کنید.

وظایف

1. **پیش‌پردازش داده**: مقادیر گمشده را پردازش کنید، ویژگی‌ها را در صورت لزوم نرمال کنید / استاندارد کنید.
2. **تحلیل داده‌های اکتشافی (EDA)**: تحلیل توزیع ویژگی‌ها، همبستگی‌ها و غیره.
3. **انتخاب / مهندسی ویژگی**: ویژگی‌های مرتبط را انتخاب کنید و / یا اگر لازم است ویژگی‌های جدید ایجاد کنید.
4. **انتخاب مدل**: با الگوریتم‌های طبقه‌بندی مختلف (مانند رگرسیون لجستیک، جنگل تصادفی، SVM و غیره) آزمایش کنید.
5. **ارزیابی مدل**: مدل‌ها را با استفاده از معیارهای ارزیابی مناسب (مانند دقت، دقت، بازیابی، امتیاز F1) ارزیابی کنید.
6. **تنظیم پارامترهای پیش‌پردازش**: پارامترهای بهترین مدل را تنظیم کنید.
7. **انتخاب نهایی مدل**: بر اساس نتایج ارزیابی، بهترین مدل را انتخاب کنید.
8. **تفسیر مدل**: نتایج را تفسیر کرده و اهمیت ویژگی‌ها در پیش‌بینی دیابت را مشخص کنید.