نمونه سوال برنامه‌نویسی

دوره استادی هوش مصنوعی درسمن

یادگیری ماشین

موضوع: پیش‌پردازش مجموعه داده

درجه سختی سوال: آسان☐ متوسط☐ سخت☐

تمرین شماره 1:

پیش‌پردازش مجموعه داده با نام Loans.csv برای اهداف یادگیری ماشین

این مجموعه داده شامل فیلدهای زیر است:

- شناسه مشتری
- نوع وام
- مقدار وام
- پرداخت شده
- شناسه وام
- شروع وام
- پایان وام
- نرخ سود

تمرین باید شامل مراحل زیر باشد:

- **وارد کردن داده‌های فایل CSV و بررسی کلی آن** : در این مرحله فایل دیتاست را واکشی کرده و اطلاعات لازم را از دیتاست بدست آورید
- **تشخیص و پردازش مقادیر گم‌شده:** هر مقدار گم‌شده را در مجموعه داده شناسایی کرده و استراتژی‌های مناسب مانند جایگزینی یا حذف را پیاده‌سازی کنید.
- **تشخیص و پردازش مقادیر پرت:** مقادیر پرت را در متغیرهای عددی تشخیص داده و تکنیک‌های مناسبی مانند حذف یا استفاده از روش‌های پردازش مناسب اعمال کنید، از جمله استفاده از نقاط مرزی.
- **تبدیل متغیرها:** در صورت لزوم، تبدیل‌هایی روی متغیرها انجام دهید تا با فرضیات الگوریتم‌های یادگیری ماشین همخوانی داشته باشند، مانند تبدیل لگاریتمی و یا رادیکالی(SQRT , LOG).
- **مقیاس‌بندی متغیرهای عددی:** متغیرهای عددی را استاندارد یا نرمال کنید تا اطمینان حاصل شود که بر روی مقیاس مشابهی هستند که می‌تواند عملکرد برخی از الگوریتم‌ها را بهبود بخشد.
- **رمزگذاری متغیرهای دسته‌ای:** متغیرهای دسته‌ای را به فرمت عددی تبدیل کنید با استفاده از تکنیک‌هایی مانند رمزگذاری OneHot Encoding و Label Encoding .
- **ایجاد متغیرهای جدید:** هر متغیر جدیدی را که ممکن است قدرت پیش‌بینی مدل شما را افزایش دهد، مانند اصطلاحات تعاملی یا ویژگی‌های چندجمله‌ای ایجاد کنید.
- **تقسیم داده به مجموعه آموزش و آزمون:** در نهایت، مجموعه داده پیش‌پردازش شده را به مجموعه‌های آموزش و آزمون تقسیم کنید تا آموزش و ارزیابی مدل را تسهیل کنید.

- اطمینان حاصل کنید که هر مرحله به طور روشن پیاده‌سازی و کامنت‌گذاری شده است و کد شما ساختارمند و خوانا است.
- پرونده پروژه خود را به GitHub آپلود کرده و لینک آن را به عنوان پاسخ به تمرین ارسال کنید.

بخش دوم: مدل‌سازی و پیش‌بینی مقدار پرداخت شده

با توجه به داده‌های پیش‌پردازش شده در بخش قبل، در این بخش شما باید مدل‌های رگرسیونی مختلف را ایجاد کرده و مقدار وام (loan amount) پرداخت شده توسط مشتریان را پیش‌بینی کنید. این مدل‌ها باید شامل رگرسیون خطی یک متغیره، رگرسیون خطی چند متغیره و رگرسیون چند جمله‌ای باشند.

- رگرسیون خطی یک متغیره: ابتدا با استفاده از یک ویژگی از داده‌ها (با انتخاب موثرترین ویژگی)، یک مدل رگرسیون خطی ایجاد کنید، مقادیر پیش‌بینی شده را بر اساس این مدل ذخیره کنید و با مقادیر واقعی مقایسه کنید.
- رگرسیون خطی چند متغیره: با استفاده از چند ویژگی از داده‌ها (با انتخاب موثرترین ویژگی‌ها)، یک مدل رگرسیون خطی چند متغیره ایجاد کنید، مقادیر پیش‌بینی شده را بر اساس این مدل ذخیره کنید و با مقادیر واقعی مقایسه کنید.
- رگرسیون چند جمله‌ای: با استفاده از ترکیب متغیرهای موجود (درجه 2 و 3)، یک مدل رگرسیون چند جمله‌ای ایجاد کنید، مقادیر پیش‌بینی شده را بر اساس این مدل ذخیره کنید و با مقادیر واقعی مقایسه کنید.
- ارزیابی مدل: هرکدام از مدل‌های ساخته شده را با استفاده از داده‌های آزمون ارزیابی کنید.
- کامنت‌گذاری: هر گام از فرآیند مدل‌سازی و ارزیابی را به طور دقیق توضیح دهید و کدهای خود را کامنت‌گذاری کنید.

اطمینان حاصل کنید که هر مرحله به طور روشن پیاده‌سازی و کامنت‌گذاری شده است و کد شما ساختارمند و خوانا است.

در پایان، پروژه خود را در GitHub آپدیت نمایید.