تکنولوژی های استفاده شده

پایتون، fast api، داکر،swagger

پیاده سازی متد های یادگیری ماشین با استفاده پایتون صورت گرفته است.

پیاده سازی restful api با استفاده از fast api.

کانتینر کردن سرویس ها برای راحتی اجرا با استفاده از داکر.

داکیومنت کردن متد ها و سرویس ها بوسیله swagger.

نحوه اجرا

ابتدا باید داکر را روی سیستم خود نصب کنیم و بعد از آن در پوشه اصلی پروژه کامند docker -compose up -- build ران کرده و سرویس شما روی پورت ۹۰۰۰ بالا می آید. برای تغییر آن می توانید در فایل docker-compose.yaml پورت مورد نظر خود را وارد کنید.

سرویس اول

هدف این سرویس پر کردن داده های ناموجود در یک سری زمانی می باشد. دیتایی که در این سرویس میگیریم به فرمت string و بدین صورت است

%Y/%m/%d

کانفیگ ها به صورت type, time, interpolation است که باید ان ها را به صورت استرینگ وارد کنیم. بعد از آن با استفاده از متد interpolation انتخاب شده و نوع تقویم متد linear interpolation تایم سری های

مفقود شده را پیدا کرده و با استفاده از روش خطی برای آنها مقدار تعیین میکند.

سرویس دوم

هدف این سرویس همانند سرویس قبلی می باشد با این تفاوت که در ابتدا تاریخ میلادی به شمسی و سپس درون یابی صورت میگیرد و فرمت ورودی آن به مانند سرویس اول میباشد.

سرویس سوم

هدف این سرویس کشف داده پرت می باشد. برای سری های زمانی از الگوریتم های isolation forest و lof استفاده شده است. الگوریتم جنگل ایزوله (Isolation Forest) یا جنگل جداسازی، یک الگوریتم «یادگیری بدون نظارت» (Unsupervised Learning Algorithm) برای تشخیص ناهنجاری (Anomaly) است که برای جداسازی نقاط پرت (Outlier) به کار می‌رود. البته در اغلب روش‌های شناسایی نقاط پرت، بقیه نقاط که رفتار عادی دارند مورد ارزیابی قرار گرفته و براساس رفتار آن‌ها، نقاط پرت مشخص می‌شوند در حالیکه در الگوریتم جنگل ایزوله از ابتدا اینگونه نقاط مورد بررسی قرار می‌گیرند.

الگوریتم عامل دورافتاده محلی بر مبنای مفهوم چگالی محلی بنا شده و در آن محلی بودن بر اساس k نزدیک‌ترین همسایگی تعیین می‌شود که فاصله آن‌ها برای تخمین چگالی مورد استفاده قرار می‌گیرد. با مقایسه چگالی محلی یک شی با چگالی‌های همسایه‌های آن می‌توان نواحی دارای چگالی مشابه و نقاطی که اساسا چگالی کمتری نسبت به همسایه‌های

خود دارند را تعیین کرد. این موارد به عنوان دورافتادگی (داده پرت) در نظر گرفته می‌شوند. چگالی محلی به وسیله فاصله معمولی که یک نقطه داده توسط همسایه‌های خود «دسترسی‌پذیر» است تخمین زده می‌شود. برای باقی داده های غیر سری زمانی از الگوریتم های dbscan و isolation forest استفاده می شود.

نام کامل این الگوریتم، «خوشه‌بندی فضایی مبتنی بر چگالی در کاربردهای دارای نویز» (Density Based Spatial Clustering of Applications with Noise) است که به اختصار به آن DBSCAN گفته می‌شود. در الگوریتم DBSCAN نیازی به این نیست که تعداد خوشه‌ها از ابتدا تعیین شود. این الگوریتم می‌تواند خوشه‌های دارای اشکال پیچیده را کشف کند. همچنین، می‌تواند نقاط داده‌ای که بخشی از هیچ خوشه‌ای نیستند (نقاط دورافتاده یا ناهنجار) را شناسایی کند. این قابلیت برای تشخیص ناهنجاری بسیار مفید است. DBSCAN با شناسایی نقاطی که در نواحی شلوغ (چگال) از «فضای ویژگی» (Feature Space) قرار دارند کار می‌کند. منظور از نواحی چگال، قسمت‌هایی است که نقاط داده بسیار به یکدیگر نزدیک هستند (نواحی چگال در فضای ویژگی).

سرویس چهارم

هدف این سرویس مدیریت داده های نامتوازن می باشد. در این سرویس از الگوریتم های under sampling و over sampling و smote استفاده شده است.

داده های نامتوازن یکی از مشکلات موجود در طبقه بندی داده ها می باشد. داده های نامتوازن داده هایی هستند که نسبت کلاس ها در آن بسیار متفاوت با هم هستند.اگر ۹۰٪ داده ها مربوط به یک کلاس و ۱۰٪ داده های مربوط به کلاس دیگر(کلاس غالب یا اکثریت) باشد آن گاه داده ها نامتوازن هستند.در یادگیری ماشین نمونه گیری Undersampling و نمونه گیری Oversampling دو روش هستند که با در برخورد با داده های نامتوازن به کار می روند.می توانید از کلاس اکثریت کم نمونه گیری کنید یا روی کلاس اقلیت را بیش نمونه گیری انجام دهید یا از ترکیب هر دو روش استفاده کنید دلیل اینکه ما طبقه بندی نامتوازن را به عنوان یک مشکل شناسایی می کنیم ، این است که می تواند بر عملکرد الگوریتم های یادگیری ماشین تأثیر بگذارد. در نمونه گیری Oversampling سعی می شود از کلاس اقلیت نمونه های بیشتر ایجاد شود تا نسبت کلاس ها به هم نزدیک شود. همچنین در Undersampling سعی می شود از کلاس حداکثر نمونه گیری کنیم. در واقع در این روش ما از همه نمونه ها در کلاس بیشتر استفاده نمی کنیم تا نسبت کلاس ها به یکدیگر نزدیک شود.

ابزار SMOTE، پیاده‌سازی الگوریتم شناخته شده «تکنیک بیش‌نمونه‌گیری اقلیت مصنوعی» (Synthetic Minority

Oversampling Technique) است که در بسته نرم‌افزاری imbalanced-learn قرار دارد. این الگوریتم برای کلاس اقلیت، نمونه‌های جدیدی در همسایگی نمونه‌های موجود در این کلاس تولید می‌کند.