
بسم الله الرحمن الرحيم

چکیده

مجموعه داده شامل بیش از ۲۸۴۸۰۷ داده از داده‌های بانک‌های اروپا شامل بیش از ۲۵ کلاس و همچنین زمان و مقدار آن‌هاست که در اینجا با ۱ و ۰ شماره گذاری شده اند. ما در اینجا قصد داریم ابتدا با پیش پردازش و سپس با استفاده از الگوریتم kmeans برای خوشه بندی و همچنین الگوریتم های SVC, LogisticRegression, KNeighborsClassifier به پاسخی برای دسته بندی و همچنین خوشه بندی مجموعه داده برسیم.

واژه‌های کلیدی:

classification ، هوش مصنوعی، clustering

فهرست مطالب

فهرست اشکال

شکل صفحه

فصل اول

مقدمه

مجموعه داده شامل بیش از ۲۸۴۸۰۷ داده از داده‌های بانک‌های اروپا شامل بیش از ۲۵ کلاس و همچنین زمان و مقدار آن‌هاست که در اینجا با ۱ و ۰ شماره گذاری شده اند.

ما در اینجا قصد داریم ابتدا با پیش پردازش و سپس با استفاده از الگوریتم kmeans برای خوشه بندی و همچنین الگوریتم های SVC, LogisticRegression, KNeighborsClassifier به پاسخی برای دسته بندی و همچنین خوشه بندی مجموعه داده برسیم.

فصل دوم

۲-۱ خوشه بندی و طبقه بندی

۲-۱-۱ خوشه بندی

خوشه‌بندی یا آنالیز خوشه(Clustering)در آمار و یادگیری ماشینی، یکی از شاخه های یادگیری بدون ناظر می‌باشد که آن ورودی هست و خروجی ای وجود ندارد و مدل خودش الگوی نهفته داده را پیدا کرده و سپس نمونه‌ها را به دسته‌هایی که اعضای آن مشابه یکدیگر می‌باشند تقسیم می‌کند که به این دسته ها خوشه گفته میشود. بنابراین خوشه مجموعه ای از اشیاء می‌باشد که در آن اشیاء با یکدیگر مشابه بوده و با اشیاء موجود در خوشه‌های دیگر غیر مشابه می‌باشند.

۲-۱-۲ طبقه بندی

طبقه بندی (classification) علمی است که بر اساس داده‌های قبلی که دارای برچسب هستند، مدلی برای پیش بینی برچسب داده‌های جدید می‌سازد.

فصل سوم

پیاده سازی

۳-۱ مقدمه

کد به زبان Python نوشته شده است، بخش‌های کامل کد رو می‌توانید در گیتهاب مشاهده کنید.
کد به شکل کامل با استفاده از مفاهیم OOP نوشته شده است. لیست کلاس ها شامل موارد زیر می‌باشد:

۳-۲ پیش‌نیازها

برای استفاده از کد ها نیاز به استفاده از زبان برنامه نویسی پایتون و کتابخانه‌های numpy, pandas, matplotlib, seaborn و scikit-learn دارید. که با دستور pip install numpy pandas matplotlib seaborn scikit-learn میتوانید آن‌ها را نصب کنید.

۳-۳ ساختار کلی کد

کد به شکل □□□ نوشته شده اما در آن متاسفانه زیبایی برنامه نویسی رعایت نشده که ان شاالله در آینده تغییر خواهد یافت. کد از یک کلاس به نام CreditCardFraudClassifier تشکیل شده است که وظیفه ی آن ایمپلیمنت کردن مباحث مورد نیاز برای کد است.
توابع

۳-۳-۱ info

این تابع اطلاعات کلی از کد را نمایش می‌دهد.

۳-۳-۲ describe

مولفه های کد مانند کمینه مقدار و بیشینه مقدار و فیچر ها را نمایش می‌هد.

۳-۳-۳ process

کار های پیش پردازش دیتا بر عهده ی این تابع می‌باشد که شامل
بالانس کردن دیتا

نورمال کردن مقادیر آن

حذف مقادیر بی تاثیر مانند زمان

جدا کردن داده به نسبت بیست به هشتاد برای رسیدن به پاسخ نهایی و تست آن

۳-۳-۴ visualize_kmeans_clusters

این تابع با استفاده از روش خوشه بندی kmeans مجموعه داده را در دو مجموعه خوشه بندی کرده و نتیجه ی نهایی را نمایش می‌دهد.

۳-۳-۵ test_all_models

این تابع با دریافت لیستی از مدل ها تمام آن‌ها را تست کرده و نتایج را تحلیل و گزارش می‌کند.

فصل چهارم

نتیجه گیری

۴-۱ تاثیر پیش پردازش

در روند توسعه ی نرم افزار به این نتیجه رسیدیم که در صورت باقی ماندن دیتا های نادرست باعث به خطا افتادن مدل هوش مصوعی خود میشویم و اگر دیتا را بالانس نکنیم مدل خوشه بندی به سمت مدلی که دیتای بیشتری دارد میل می‌کند.

۴-۲ سرعت مدل

در این کد متوجه می‌شویم مدل رندوم فارست برای این دیتاست بهترین دقت را دارد ولی به دلیل کند بودن آن نمیتوان به آن تکیه کرد. لیست باقی مدل ها و سرعت و زمان و دقت آنها:

RandomForest: 0.99964 3:12 LogisticRegression: 0.99912 0:10 KNeighborsClassifier: 0.99953 0:5 SVC 0.99932 2:45

منابع و مراجع