به نام خداوند جان و خرد

پروژه شماره ۲: طبقه‌بندی
نام درس: داده کاوی
استاد درس: دکتر رضا رمضانی
حل تمرین‌ها: محمدرضا توکلیان، نگین آبادانی، علی عابدزاده، سپیده صالح
مهلت تحویل: ۱۴۰۰/۰۳/۳۱
سامانه تحویل: http://lms.ui.ac.ir

در این پروژه شما باید با استفاده کتابخانه‌های پایتون مطالبی که در فصل‌های ۲ ، ۳ و ۶ آموخته‌اید را پیاده‌سازی کنید.

مجموعه داده‌ای که همراه این فایل ضمیمه شده است، شامل چندین ستون عددی و غیرعددی (catergorial) است. این مجموعه داده در هر سطر شامل اطلاعات یک معامله است که در ستون Stage مشخص شده است که شخصی که با مشتری این معامله را برقرار کرده است در نهایت برنده شده است یا خیر. ستون Stage شامل سه مقدار Won, Lost و In progress می‌باشد. هدف شما طبقه‌بندی این ستون به عنوان برچسب کلاس برای پیش‌بینی مقادیر Won و Lost است. (مقادیر In Progress باید توسط شما به نحوی سازماندهی شوند، و شما باید تصمیم بگیرید با سطرهایی که مقدار In progress دارند چکار کنید)

در ادامه به معرفی ستون‌های دیگر پرداخته می‌شود:

نام مشتری	Customer
نام کارمندی که مسئول فروش است	Agent
ایمیل کارمندی که مسئول فروش است	SalesAgentEmailID
ایمیل کارمندی که از سمت مشتری در ارتباط است	ContactEmailID
نام محصولی که قرار است فروخته شود	Product
مقداری که تا به لحظه برای معامله تعیین شده است	Close_Value
تاریخی که معامله آغاز شده است	Created Date
تاریخی که معامله پایان یافته است	Close_Date

همچنین فایلی به نام Interactions در کنار این فایل آورده شده است که شما باید از محتویات آن استفاده کنید. این فایل حاوی مکالمات بین فروشنده و خریدار، نوع مکالمه و تاریخ مکالمه است. شما با استفاده از خلاقیت خود می‌توانید از اطلاعات این فایل نیز استفاده کنید تا دقت طبقه بندی را افزایش دهید. استفاده از این فایل الزامی است. می‌توانید از روش‌های تحلیل احساسات برای بررسی متون مکالمات استفاده نمایید.

مراحل زیر را برای انجام و تحویل پروژه انجام دهید.

مرحله ۱: دیتاست داده شده را پیش پردازش کنید. مقادیر NA را مقدار دهی کنید و EDA (تحلیل داده اکتشافی) را به خوبی انجام دهید. این ستون‌ها براساس ماهیت خود میتواند تولید کننده ویژگی‌های بیشتری باشند که ممکن است دقت مدل شما را بالاتر ببرند. توصیه می‌شود در این مرحله همبستگی و ارتباط بین تمام ویژگی هایی که میتوانید استخراج کنید را بررسی بنمایید.

مرحله ۲: شش الگوریتم Naïve Bayes, Decision Tree , Random Forest, KNN, SVM و MLP (multi-layer perceptron) را برای این دیتاست استفاده کنید و نتایج را با یکدیگر بررسی کنید. یک روش را به عنوان بهترین روشی که به آن دست پیدا کرده‌اید به عنوان مدل خودتان انتخاب کنید.

مرحله ۳: مقدار Accuracy و F1 را برای داده‌ی تست خود حساب کنید. به نظر شما کدام متریک ارزیابی برای داده‌ی داده شده بهتر است؟

مرحله ۴: ماتریس Confusion و نمودار ROC را رسم کنید و نتایج را تحلیل کنید.

مرحله ۵: در انتها شما باید تابع inference در فایل inference.py که فایل پروژه آورده شده است را بنویسید. ورودی این تابع یک dataframe است که شامل ستون‌های دیتاست اصلی به جز ستون Stage است. خروجی این تابع باید یک لیست باینری باشد که در آن صفر به معنای Lost و یک به معنای Won است.

مرحله ۶: مرحله آخر، در سه روز پایانی تحویل انجام می‌شود که نمره اضافه محسوب می‌شود. فایل‌های پروژه شامل کد و مدل را داخل Git گذاشته و لینک مخزن گیت را به آدرسی که به شما داده خواهد شد ارسال می‌کنید. سرور به صورت خودکار مخزن شما را Clone گرفته و از فایل inference.py شما برای محاسبه دقت مدلتان استفاده می‌کند. شما میتوانید با مراجعه به لینکی که داده می‌شود دقت مدل تیم خود را در مقایسه با بقیه تیم‌ها مشاهده کنید.

نکات بسیار مهم:

- حتما پیش‌پردازش را به صورت درست انجام دهید.
- داده‌ها را خوب بررسی کنید و feature engineering را به درستی انجام دهید، اطلاعات زیادی در دل ستون‌ها میتواند نهفته باشد.
- کپی نکنید! از قبل تمام کدهای نوشته شده در اینترنت جمع‌اوری شده است چون تعداد آن ها محدود است کپی کردن شما مشخص می‌شود.
- در این پروژه به جز کتابخانه‌های خود پایتون فقط می‌توانید از سه کتابخانه numpy, pandas و scikit-learn استفاده کنید.