۱) فایل ToyotaCorola.csv حاوی داده‌هایی مربوط به فروش اتومبیل‌های دست دوم تویوتا در اواخر تابستان ۲۰۰۴ در کشور هلند است. این فایل شامل ۱۴۳۶ رکورد و ۳۸ متغیر از جمله قیمت، مدل، کارکرد، اسب‌بخار و مشخصات دیگر است. هدف پیش‌بینی قیمت یک اتومبیل تویوتا کارکرده بر اساس مشخصات آن است (مثال ارایه شده در بخش ۶-۳ زیرمجموعه‌ای از این مجموعه داده‌ها را نشان می‌دهد). داده‌ها را به سه بخش آموزشی (۵۰ درصد)، اعتبارسنجی (۳۰ درصد) و آزمایشی (۲۰ درصد) تقسیم کنید.

متغیر خروجی Price را در نظر گرفته و بر اساس متغیرهای زیر یک مدل رگرسیون خطی چندگانه تدوین کنید.

Age_08_04, KM, Fuel_Type, HP, Automatic, Doors, Quarterly_Tax, Mfr_Guarantee, Guarantee_Period, Airco, Automatic_airco, CD_Player, Powered_Windows, Sport_Model, and Tow_Bar.

الف) سه یا چهار متغیر مهم برای پیش‌بینی قیمت اتومبیل کدامند؟

ب) با استفاده از متریک‌هایی که در اینجا مفید می‌دانید، کارآیی مدل در پیش‌بینی قیمت اتومبیل‌ها را ارزیابی کنید.

۲) بانک یونیورسال نسبتاً یک بانک جوانی است که به سرعت از لحاظ جذب مشتری رشد کرده است. اکثر مشتریان این بانک را بدهکارانی تشکیل می‌دهند که هر یک با مقدار بدهی متفاوتی با بانک در ارتباط هستند. پایگاه مشتریان وام‌گیرنده بسیار اندک است و بانک مایل است وام‌گیرندگان خود را به سرعت افزایش دهد. در واقع بانک به دنبال روشی است که مشتریان بدهکار خود را به وام‌گیرندگان تبدیل کند (در حالی‌که همچنان بدهکار باقی بمانند).

نتایج حاصل از راه‌اندازی پویشی که سال گذشته اجرا شد نشان از نرخ تبدیل ۹ درصدی است. این موضوع باعث شد تا دپارتمان بازاریابی به دنبال راه‌اندازی پویشی هوشمندتر همراه با بازاریابی هدف باشد. هدف استفاده از KNN برای پیش‌بینی این موضوع است که آیا یک مشتری جدید پیشنهاد وام را قبول می‌کند یا خیر.

فایل UniversalBank.csv حاوی اطلاعاتی در مورد ۵۰۰۰ مشتری است. این داده‌ها شامل اطلاعات دموگرافیک مشتری (سن، درآمد و ...)، رابطه مشتری با بانک (وثیقه، حساب اوراق بهادار و ...) و پاسخ او

به پیشنهاد وام بانک می‌باشد. در میان این ۵۰۰۰ مشتری تنها ۴۸۰ نفر (۹/۶ درصد) درخواست وام را پذیرفته‌اند.

داده‌ها را به دو بخش آموزشی (۶۰ درصد) و اعتبارسنجی (۴۰ درصد) اِفراز کنید.

الف) یک مشتری با مشخصات زیر را در نظر بگیرید:

Age = 40, Experience = 10, Income = 84, Family = 2, CCAvg = 2, Education_1 = 0, Education_2 = 1, Education_3 = 0, Mortgage = 0, Securities Account = 0, CD Account = 0, Online = 1, and Credit Card = 1.

الگوریتم KNN را با مقدار یک برای K و برای تمامی متغیرهای پیشگو به جز ID و ZIPCode اجرا کنید. به خاطر داشته باشید که ابتدا متغیرهای طبقه‌ای با بیش از دو مقدار را به متغیرهای دودویی تبدیل کنید. رده موفق (پذیرش وام) را با عدد یک بیان و از مقدار بُرش پیش‌فرض ۰/۵ نیز استفاده کنید. این مشتری چگونه رده‌بندی می‌شود؟

ب) با چه مقداری از K موازنه‌ای میان بیش‌برازش و صرف‌نظر از اطلاعات متغیرها رخ می‌دهد؟

ج) ماتریس درهم‌ریختگی داده‌های اعتبارسنجی حاصل از استفاده بهترین مقدار K را محاسبه کنید.

د) مشتری جدید با مشخصات زیر را با مقدار بهینه K رده‌بندی کنید

Age = 40, Experience = 10, Income = 84, Family = 2 CCAvg = 2, Education_1 = 0, Education_2 = 1, Education_3 = 0, Mortgage = 0, Securities Account = 0, CD Account = 0, Online = 1 and Credit Card = 1.

ه) این دفعه داده‌ها را به نسبت (۵۰٪، ۳۰٪ و ۲۰٪) به سه بخش آموزشی، اعتبارسنجی و آزمایشی اِفراز کنید. روش KNN را با مقدار K منتخب در بالا اجرا کنید. ماتریس درهم‌ریختگی داده‌های آزمایشی را با ماتریس‌های مربوط به داده‌های آموزشی و اعتبارسنجی مقایسه کنید. در مورد تفاوت آنها بحث کنید.

۳) فایل eBayAuctions.csv شامل اطلاعاتی در مورد ۱۹۷۲ حراجی شرکت eBay است که بین ماه‌های مِی تا ژوئن ۲۰۰۴ رخ داده است. هدف ما ساخت مدلی از این داده‌هاست که می‌تواند حراجی‌ها را به عنوان رقابتی یا غیررقابتی رده‌بندی کند. حراجی رقابتی به یک حراجی اطلاق می‌شود که برای هر آیتم آن حداقل دو پیشنهاد وجود داشته باشد. داده‌ها شامل متغیرهایی است که آیتم‌ها و فروشنده و قوانین فروش را توصیف می‌کنند. در ضمن قیمت بسته شدن حراجی نیز در دسترس است. مایلیم پیش‌بینی کنیم که آیا حراج رقابتی خواهد بود یا خیر.

به منظور پیش‌پردازش داده‌ها، متغیر Duration را به یک متغیر طبقه‌ای تبدیل کنید. مجموعه داده‌ها را نیز به دو بخش آموزشی (۶۰ درصد) و اعتبارسنجی (۴۰ درصد) اِفراز کنید.

الف) با استفاده از تمامی متغیرها یک درخت رده‌بندی برازش کنید. به منظور اجتناب از بیش‌برازش، حداقل تعداد رکوردهای موجود در گره پایانی را با مقدار ۵۰ و حداکثر عمق درخت را با مقدار ۷ تنظیم کنید. نتایج را به صورت مجموعه‌ای از قواعد بنویسید (توجه کنید چنانچه به واسطه محدویت‌های نرم‌افزاری و شفافیت ارائه تعداد متغیرها را کاهش دهید، ممکن است متغیر مناسبی انتخاب شود).

ب) آیا این مدل برای پیش‌بینی خروجی یک حراجی جدید کاربردی است؟

ج) در مورد اطلاعات جالب و غیرجالبی که این قواعد برای ما مهیا می‌کنند بحث کنید.

د) درخت رده‌بندی دیگری برازش کنید (با همان مقدار ۵۰ برای حداقل تعداد رکوردهای گره‌های پایانی و مقدار ۷ برای حداکثر عمق درخت) اما این بار با استفاده از متغیرهایی که می‌توانند برای پیش‌بینی خروجی یک حراجی جدید استفاده شوند. نتایج درخت را با کمک قواعد بیان کنید. در مورد کوچک‌ترین مجموعه از قواعد لازم برای رده‌بندی بحث کنید.

ه) نتایج درخت را بر روی یک نمودار پراکنشی رسم کنید. برای انجام این‌کار به سراغ دو متغیر با اهمیت بالا می‌رویم. هر یک از حراجی‌ها را با یک نقطه نمایش می‌دهیم و مختصات آن بر اساس مقادیر دو متغیر فوق تعیین می‌شوند. با کمک رنگ‌ها یا نمادهای مختلف، حراجی‌های رقابتی و غیررقابتی را از یکدیگر تفکیک کنید. با رسم خطوط (دستی یا با کمک پایتون) این تفکیک را نشان دهید. با توجه به معانی این دو متغیر، آیا این تفکیک منطقی است؟ آیا برای تفکیک این دو رده، این کار مناسب است؟