**پروژه**

**درس بزرگداده و سیاست گذاری**

پیش از پاسخ سوالات به نکات زیر توجه کنید:

* پروژه درس به صورت انفرادی است و دارای دو بخش کلی کد نویسی و تحلیل می باشد.
* پاسخ شما باید شامل کدنویسی پایتون و یک گزارش از تحلیل کدنویسی خودتان و همچنین یک گزارش Academic با رعایت اضول گزارشنویسی که شامل تحلیل کد، مصورسازی دادهها، مدلها و هرگونه تحلیل دیگر میباشد.
* مدت زمان پاسخدهی حداکثر 7 روز بوده و در صورت لزوم 1 روز تمدید خواهد شد.

نحوه بارگذاری پاسخ پروژه

* در پاسخ­دهی به سوالات موجود هیچگونه محدودیتی وجود ندارد و میتواند از پکیج مناسب استفاده نمایید. ( خارج درس کلاس نیز مشکلی نخواهد داشت)
* در صورت وجود هر گونه ابهام یا سوال از طریق گروه درس در پلتفرم تلگرام و یا از طریق ایمیل پیگیری نمایید. hamed\_meraji@ie.sharif.edu
* برای تحویل فایل پاسخ خود لطفا تمامی فایلهایی که ممکن است در تصحیح برایتان موثر باشد را در یک فایل rarقرار دهید.
* فایل ارسالی حتما شامل نام، نام خانوادگی، شماره دانشجویی و همچنین نام فایل ارسالی به صورت: HW1\_name\_StudentNumber باشد.
* نمره تشویقی هر تمرین صرفا در همان تمرین لحاظ خواهد شد.

Juko Analytics - مشاوره خرده فروشی مبتنی بر داده

بررسی اجمالی:

Juko Analytics یک استارتاپ مشاوره تحلیلی خرده فروشی است که هدف آن کمک به زنجیره­های خرده فروشی است تا استراتژی فروش و بازاریابی خود را با استفاده از بینش های مبتنی بر داده بهینه کنند.

قابلیت های کلیدی ارائه شده عبارتند از:

داشبورد تجزیه و تحلیل فروش: داشبوردهای تعاملی برای تجسم و نظارت بر داده های معاملات خرده فروشی در ابعاد مختلف مانند شعبه، ویژگی های مشتری، رشد فروش در طول زمان و غیره.

تجزیه و تحلیل پیشرفته: مدل سازی آماری و تجزیه و تحلیل what-if برای پیش بینی تقاضا، بهینه سازی قیمت گذاری، مدل سازی آمیخته بازاریابی بر اساس داده های معاملات گذشته.

تقسیم بندی و هدف گذاری: الگوهای تراکنش را تجزیه و تحلیل کنید تا بخش های مشتریان با ارزش بالا را شناسایی کرده و برای هدف قرار دهید.

مدیریت کمپین: کمپین های تبلیغاتی، طراحی کاتالوگ و اعلان های فشاری موبایل را با استفاده از داده ها برنامه ریزی کنید.

مزیت رقابتی:

این استارت‌آپ از یک تیم متخصص علم داده برای استخراج داده‌های تراکنش‌ها و ارائه تحلیل‌های عملی استفاده می‌کند که برنامه‌ریزی و تصمیم‌گیری‌های خرده‌فروشی را به‌طور قابل‌توجهی تضمین می‌کند. داشبوردهای سفارشی سازی شده پیشرفته به جای تکیه بر گزارش های ماهانه منسوخ شده، فروش ۲۴x۷ را ارائه می دهند.

طرح کشش:

این استارت آپ مشاوره، زنجیره های خرده فروشی کوچک و متوسط ​​را به عنوان مشتری در اختیار دارد. برای دسترسی به پیشنهادات تجزیه و تحلیل استاندارد و پروژه های مشاوره سفارشی، هزینه اشتراک می گیرد. از طریق یک مدل فریمیوم که تجزیه و تحلیل های اولیه را برای فروشگاه های منتخب ارائه می دهد، اعتبار ایجاد می کند.

پتانسیل رشد:

با ثبت نام مشتریان بیشتر، استارت آپ با عمودی کردن راه حل ها بر اساس دسته های عمده خرده فروشی گسترش می یابد تا بیشتر پیشنهادات خود را سفارشی کند. گسترش بین المللی همچنین یک مسیر رشد جذاب را ارائه می دهد. مدل‌های تحلیلی و ML با افزایش داده‌ها اصلاح‌تر می‌شوند.

در اینجا توضیح مختصری از هر یک از ویژگی ها یا برچسب ها در مجموعه داده آمده است:

شناسه فاکتور: یک شناسه منحصر به فرد برای هر فاکتور یا تراکنش.

شعبه: شعبه یا مکانی که معامله انجام شده است.

شهر: شهری که شعبه در آن قرار دارد.

نوع مشتری: نشان می دهد که مشتری مشتری دائمی است یا جدید.

جنسیت: جنسیت مشتری.

خط تولید: دسته یا نوع محصول خریداری شده.

قیمت واحد: قیمت یک واحد از محصول.

تعداد: تعداد واحدهای محصول خریداری شده.

مالیات ۵%: مبلغ مالیات (۵% از کل هزینه) اعمال شده برای معامله.

مجموع: کل هزینه معامله با احتساب مالیات.

تاریخ: تاریخی که معامله انجام شده است.

زمان: زمانی از روز که معامله انجام شده است.

پرداخت: روش پرداخت مورد استفاده (به عنوان مثال، کارت اعتباری، پول نقد).

COGS (هزینه کالاهای فروخته شده): هزینه های مستقیم مربوط به تولید یا خرید محصولات فروخته شده.

درصد حاشیه ناخالص: درصد حاشیه سود برای معامله.

درآمد ناخالص: کل سود حاصل از معامله.

رتبه بندی: رتبه بندی رضایت مشتری یا بازخورد در مورد معامله.

* **ساختار داده (15 نمره)**
1. انواع هر یک از داده­های موجود را بیان کنید و برای هر دلیل خود را ذکر کنید. (10)
2. آیا مجموعه داده­ای که دراختیار دارید، نیاز به پیش پردازش دارد؟چرا؟(5)
* مصورسازی (180 نمره)
1. برای فاکتورهایی که ممکن است بتوانید توزیعی را بیان کنید یا ترسیم کنید، ترسیم کرده و به صورت کامل تشریح کنید. (20)
2. نمودار مناسبی ایجاد کنید که درصد تراکنش‌ها را بر اساس نوع پرداخت (نقد، کارت اعتباری، Ewallet) نشان دهد. (5)
3. یک نمودار مناسب ترسیم کنید که کل مقادیر فروخته شده برای هر خط محصول را نشان می دهد، آن را تحلیل کنید. (5)
4. یک هیستوگرام از قیمت های واحد برای تجزیه و تحلیل توزیع قیمت ایجاد کنید و نسبت به کدام فاکتور دیگر می­توان قیمت را دسته بندی کنید. (10)
5. میانگین کل مبلغ فروش در هر ماه را در دوره ۳ ماهه محاسبه و ترسیم کنید. (5)
6. فروش بر اساس خط محصول را در طول زمان با استفاده از نموداری مناسب مصور کنید. (5)
7. نمودار مناسب گروه بندی شده ایجاد کنید تا میانگین کل فروش را بر اساس شهر و نوع مشتری مقایسه کند و تحلیل خود را بیان کنید. (10)
8. نمودار پراکندگی را با رتبه بندی در مقابل مقدار کل فروش ترسیم کنید تا همبستگی ها را بررسی کنید. (15)
9. نمودارهای جعبه ای ایجاد کنید تا به صورت بصری بررسی کنید که آیا تفاوت قیمت واحد بر اساس خط محصول وجود دارد یا خیر. (10)
10. گروه بندی بر اساس شاخه و ایجاد نمودارهای دایره ای برای مقایسه درصد بازدیدهای مشتریان جدید در مقابل اعضای عضو. (10)
11. برای مصورسازی مقادیر فروش بر اساس خط محصول در سراسر شاخه، یک نمودار مناسب ایجاد کنید. (5)
12. یک heatmap بصری ایجاد کنید که تغییرات مقدار کل فروش را در دوره ۳ ماهه نشان می دهد. (10)
13. کل فروش را در طول زمان، به تفکیک شاخه و با رنگی که نشان دهنده نوع پرداخت است، ترسیم کنید. (10)
14. برای مطالعه فصلی بودن روزانه و ماهانه تعداد معاملات، نمودار مناسبی تهیه کنید. (5)
15. از نمودارهای رگرسیون برای مطالعه همبستگی رتبه با مقدار کل فروش و درصد حاشیه ناخالص استفاده کنید. (10)
16. نمودار متناسبی ایجاد کنید تا میانگین فروش را بر اساس ساعت روز و روز هفته تجسم کنید. (15)
17. ترسیم نموداری دلخواه و متناسب برای به تصویر کشیدن روابط بین ویژگی های محصول مانند قیمت، حاشیه، کمیت و غیره. (5 + 10)
18. از قابلیت های ترسیم pandas برای ایجاد برنامه نویسی مضرب کوچک برای نمایش زیرمجموعه های داده استفاده کنید. (5)
19. با توجه به 18 بخش گذشته تحلیل شما و برداشت­های شما در رابطه با این مجموعه داده چیست؟ آیا نیاز به جمع­آوری داده­های دیگری هست یا خیر؟ در صورت بله بودن پاسخ علت خود را بیان کنید؟ ( 15 )
20. با توجه به داده­ها آیا روزهای تعطیل قابل شناسایی است و آیا میزان خرید و فروش در روزهای تعطیل توزیع خاصی دارد؟ (\*\*)
* مدل سازی (300 نمره)
1. برای مجموعه داده شده مدل KNN پیاده سازی کنید و آیا می­توان K بهینه را بدست آورد؟ (40)
2. برای مجموعه گفته شده یک مدل درخت تصمیم، جنگل تصادفی و مدل رگرسیون خطی (در صورت امکان) پیاده سازی کنید. (120)
3. کدام یک از مدل­های فوق عملکرد بهتری را دارد؟ کدام معیار را بررسی کردید؟ تمامی معیارهای ممکن برای ارزیابی هر یک از مدل­ها را بیان کنید؟ (50)
4. در رابطه با روش­هایی همانند رگرسیون لسو و ریج و پیاده سازی روی این مجموعه داده تحقیق کنید؟ (20)
5. در رابطه با مفاهیم feature selection در این مطالعه موردی تحقیق کنید و روش­هایی را بیان کنید. (20)
6. در این مجموعه داده روش Apriori Method قابل بررسی است؟اگر بله چرا؟ اگر خیر به چه داده­های نیاز دارید؟ (30)
7. اگر بخواهیم از روش­های خوشه­بندی استفاده کنیم کدام یک از روش­های خوشه بندی را استفاده می­کنید؟چرا؟ (20) (\*\*پیاده سازی\*\*)
* مفهوم، تجزیه و تحلیل (125 نمره)

در هریک از سوالات زیر صرفا تجزیه و تحلیل تشریحی خود را بیان کنید و نیازی به کد نویسی نیست.

* + یادگیری عمیق: ( هر سوال 5 نمره)
1. یک مدل RNN برای پیش بینی فروش مورد انتظار برای ماه آینده بر اساس داده های سری زمانی فروش قبلی آموزش دهید.
2. یک مدل طبقه بندی تصویر CNN بسازید تا به طور خودکار محصولات را از روی تصاویر آنها دسته بندی کند.
3. از تکنیک‌های تشخیص ناهنجاری مانند رمزگذارهای خودکار برای شناسایی تراکنش‌های غیرعادی یا تقلبی استفاده کنید.
4. از یادگیری تقویتی برای بهینه سازی پویا سیاست قیمت گذاری برای به حداکثر رساندن درآمد استفاده کنید.
5. یک مدل شبکه عصبی نموداری را برای پیش‌بینی احتمال خرید بسته‌های مختلف محصول آموزش دهید.
	* متن کاوی: (هر سوال 5 نمره)
6. بازخورد و نظرات مکتوب مشتری را برای شناسایی شکایات و پیشنهادات رایج تجزیه و تحلیل کنید.
7. از تجزیه و تحلیل احساسات در ایمیل های مشتری در مورد مسائل و تجربه استفاده کنید.
8. یک مدل موضوعی با استفاده از LDA در مورد درخواست های مشتری بسازید تا بر اساس انواع پرس و جو دسته بندی کنید.
9. پیاده سازی تجزیه و تحلیل مبتنی بر جستجو برای استخراج اطلاعات کلیدی از فیلدهای متنی رایگان فاکتورهای خرده فروشی.
10. از پردازش زبان طبیعی برای چت ربات هوشمند برای کمک به سوالات مشتریان استفاده کنید.
	* تجزیه و تحلیل ( هر سوال 5 نمره )
11. سودآورترین بخش های مشتری را از الگوهای معاملات تاریخی شناسایی کنید.
12. ارزیابی عملکرد در شعب خرده فروشی و پیشنهاد بهبود برای شعب فقیرتر.
13. بهینه سازی نیازهای کارکنان در ساعات اوج فروشگاه بر اساس تجزیه و تحلیل زمان معامله.
14. تنظیمات قیمت را برای محصولات مختلف بر اساس کشش تقاضا توصیه کنید.
15. پیشنهاد مکان برای افتتاح فروشگاه های جدید با استفاده از تجزیه و تحلیل داده های جغرافیایی و جمعیتی.
16. ارزیابی عملکرد کمپین های بازاریابی گذشته با تجزیه و تحلیل افزایش فروش ایجاد شده.
17. ترافیک وب سایت و نیازهای زیرساخت ابری را بر اساس رشد پیش بینی شده فروش پیش بینی کنید.
18. مدل های قیمت گذاری را ایجاد کنید تا به طور موثر پیشنهادات مشاوره و اشتراک را قیمت گذاری کنید.
19. فرصت های فروش متقابل را بر اساس الگوهای رایج در بسته های محصول خریداری شده برجسته کنید.
20. معیارهای عملکرد مانند حاشیه درصد در مقابل زنجیره‌های خرده‌فروشی رقیب را معیار قرار دهید.
21. تجزیه و تحلیل هزینه های حمل و نقل بر اساس مکان برای شناسایی فرصت ها برای بهینه سازی لجستیک.
22. ردیابی تأثیر عوامل خارجی مانند تعطیلات و آب و هوا بر رفتار خرید.
23. برنامه استخدام را با مهارت های کلیدی مورد نیاز بر اساس استراتژی های رشد اطلاع دهید.
24. ارزیابی امکان سنجی گسترش بین المللی با استفاده از داده های خرده فروشی جهانی.
25. صرفه جویی در هزینه های بالقوه از معرفی اتوماسیون و هوش مصنوعی را کمی کنید.

(\*\*): سوالات تشویقی و شامل نمره مازاد

اگر در سوال دو عدد جمع شده نمره اول، نمره خود سوال و نمره بعدی نمره تشویقی است.