**« پروژه اول »**

**بوت‌کمپ هوش مصنوعی کوئرا**

**بهار۱۴۰۳**



**مهلت ارسال پاسخ:** تا ساعت ۲۳:۵۹ روز جمعه ۱۸ خرداد

**زمان ارائه‌ی گروهی:** شنبه ۱۹ خرداد و یکشنبه ۲۰ خرداد

**بخش اول:معرفی داده**

| [**جهت دریافت مجموعه‌داده‌ بخش کلیک کنید.**](https://drive.google.com/file/d/1jFZgYesPD9e5RX82TAofA1LvXuhbtlHJ/view?usp=sharing) |
| --- |

مرکز آمار ایران (SCI, https://amar.org.ir) از سال ۱۳۴۲ در مناطق روستایی و شهری به اجرای طرح آمارگیری هزینه و درآمد خانوار که قبلاً به عنوان بررسی بودجه خانوار شناخته می شد، می‌پردازد. در ابتدا، فقط شامل سوالات مخارج خانوار بودو در سال ۱۳۵۳ سوالات مربوط به درآمد خانوار به پرسشنامه اضافه شد.هدف از این آمارگیری ارائه برآورد هایی از میانگین درآمد و هزینه خانوارهای شهری و روستایی در سطوح استانی و کشوری است. بررسی‌ این داده ها محققان را قادر می‌سازد تا ترکیب درآمد و هزینه خانوار و الگوهای توزیع، الگوی مصرف خانوار، وزن هر کالا در سبد مصرف خانوار، همچنین خط فقر را محاسبه کنند و نابرابری در درآمد خانوار را مطالعه کنند.

داده های که در اختیار شما قرار گرفته است، با دو حرف R و U در اسم فایل آغاز می‌شوند که R مخفف Rural، معرف جداول روستايی و U مخفف Urban معرف جداول شهری است. در هر فایل تعدادی شیت وجود دارد که نمایانگر بخش‌های مختلف پرسشنامه‌ی درآمد و هزینه‌ی خانوار است جدول راهنمای شیت‌ها به شرح زیر می‌باشد:

| **نماد** | **شرح** |
| --- | --- |
| مشخصات پرسشنامه | <File\_Name>Data |
| قسمت یکم: خصوصیات اجتماعی اعضای خانوار | <File\_name>P1 |
| قسمت دوم: مشخصات محل سکونت | <File\_name>P2 |
| قسمت سوم: بخش ۱ هزینه‌های خوراکی و دخانی | <File\_name>P3S01 |
| قسمت سوم: بخش ۲ هزینه‌های نوشیدنی‌های طبقه‌بندی نشده | <File\_name>P3S02 |
| قسمت سوم: بخش‌های ۳ تا ۱۲ به جز ۴ شامل هزینه‌های غیر خوراکی | <File\_name>P3S[03-05-06-07-08-09-11-12] |
| قسمت سوم: بخش ۴ هزینه‌های بخش مسکن | <File\_name>P3S04 |
| قسمت سوم: بخش ۱۳ سایر هزینه‌ها و انتقالات | <File\_name>P3S13 |
| قسمت سوم: بخش ۱۴ سرمایه‌گذاری خانوار در ۱۲ ماه گذشته | <File\_name>P3S14 |
| قسمت چهارم: بخش ۱ درآمد پولی اعضای شاغل خانوار از مشاغل مزد و حقوق بگیری | <File\_name>P4S1 |
| قسمت چهارم: بخش ۲ درآمد پولی اعضای شاغل خانوار از مشاغل آزاد | <File\_name>P4S2 |
| قسمت چهارم: بخش ۳ درآمدهای متفرقه خانوار | <File\_name>P4S3 |
| قسمت چهارم: ستون ۹ بخش ۳ شامل وام و یارانه | <File\_name>P4S4 |

به جای <File\_name> نام فایل مورد نظر را قرار دهید به عنوان مثال اگر با داده‌های فایل U1401 کار میکنید شیت‌های مربوطه به صورت U1401P\*S هستند.

در ادامه به معرفی ستون‌های موجود در هر شیت و مقادیر وارد شده در آنها می‌پردازیم.

| **قسمت مشخصات پرسشنامه: R1401Data و U1401Data** | | |
| --- | --- | --- |
| **نام ستون** | **شرح** | **مقادیر** |
| Address | آدرس خانوار | - |
| Fasl | فصل آمارگیری | - |
| Weight | وزن خانوار | - |
| Khanevartype | نوع خانوار | ۱: ساکن معمولی- ۲: ساکن به صورت گروهی |
| Takmil | وضعیت تکمیل پرسشنامه برای خانوار اصلی | ۱: تکمیل شده- ۲: عدم تکمیل |
| TakmilDescA | عدم تکمیل- تکراری نبودن محل اقامت خانوار اصلی |  |
| TakmilDescB | عدم تکمیل- جدول ۲ | ۱: عدم تمایل به تکمیل- ۲: کهولت سن، ناتوانی، بیسوادی- ۳: عدم دسترسی به مکان- ۴: تبدیل آبادی به شهر |
| TakmilDescC | عدم تکمیل- جدول ۳ | ۱: غیبت خانوار- ۲: خالی از سکنه بودن مسکن- ۳: اقامتگاه غیر رایج- ۴: تخریب شده یا در دست ساخت- ۵: موسسه- ۶: آدرس پیدا نشد- ۷:سایر |
| Jaygozin | وضعیت تکمیل پرسشنامه برای خانوار جایگزین | ۱: تکمیل  ۲: عدم تکمیل |
| jaygozinDescA | عدم تکمیل- جدول ۴ | ۱: عدم تمایل به تکمیل- ۲: کهولت سن، ناتوانی، بیسوادی- ۳: عدم دسترسی به مکان |
| jaygozinDescB | عدم تکمیل- جدول ۵ | ۱: غیبت خانوار- ۲: خالی از سکنه بودن مسکن- ۳: اقامتگاه غیر رایج- ۴: تخریب شده یا در دست ساخت- ۵: موسسه- ۶: آدرس پیدا نشد- ۷:سایر |
| BlkAbdJaygozin | بلوک یا آبادی خانوار جایگزین | - |
| RadifJaygozin | شماره ردیف خانوار جایگزین | - |
| province | استان | - |
| town | شهر | - |

| **قسمت يكم: خصوصیات اجتماعي اعضای خانوار U1401P1 وR1401P1** | | |
| --- | --- | --- |
| **نام ستون** | **شرح** | **مقادیر** |
| Address | آدرس خانوار | - |
| member | شماره ردیف اعضای خانوار | - |
| Relation | بستگی با سرپرست | Head- Spouse-Child-SonDaughter\_inLaw-GrandSonDaughter-parent-sibling-OtherRelative-NonRelative |
| Gender | جنس | Male- Female |
| Age | سن | - |
| Literacy | وضع سواد | Literate-illiterate |
| Studying | آیا در حال حاضر تحصیل می کند؟ | Yes-No |
| Degree | پایه یا مدرک | Elemantry  Secondary  HighSchool  Diploma  College  Bachelor  Master  PhD  Other |
| Occupationalst | وضع فعالیت | Employed- unemployed-IncomeWOJob-student-housewife-other |
| Martialst | وضعیت تاهل | Single- married-Widowed-Divorced |

| **قسمت دوم مشخصات محل سكونت U1401P2 و R1401P2** | | |
| --- | --- | --- |
| **نام ستون** | **شرح** | **مقادیر** |
| Address | آدرس خانوار | - |
| tenure | نحوه تصرف محل سکونت | OwnedEstateLand- OwnedEstate- Rent-Mortgage-Service-Free-Other |
| room | تعداد اتاق در اختیار خانوار | - |
| space | سطح زیربنا | - |
| construction | نوع اسکلت بنای محل سکونت | ۱:فلزی- ۲:بتون آرمه- ۳:سایر |
| Material | مصالح عمده | MetalBlock-BrickWood-Cement-Brick-Wood-WoodKesht-KeshtGel-Other |
| vehicle | اتومبیل سواری شخصی | True-False |
| motorcycle | موتورسیکلت | True-False |
| bicycle | دوچرخه | True-False |
| Radio | رادیو | True-False |
| Radiotape | رادیو ضبط، ضبط و پخش صوت | True-False |
| TVbw | تلویزیون (سیاه وسفید) | True-False |
| TV | تلویزیون رنگی | True-False |
| VHS-VCD-DVD | انواع ویدئو و VCD و DVD | True-False |
| computer | رایانه | True-False |
| Cellphone | تلفن همراه (غیر شغلی ) | True-False |
| Freezer | فریزر | True-False |
| refrigerator | یخچال | True-False |
| fridge | یخچال فریزر | True-False |
| stove | اجاق گاز | True-False |
| vacuum | جاروبرقی | True-False |
| washingMachine | ماشین لباسشویی | True-False |
| sewingMachine | چرخ خیاطی | True-False |
| fan | پنکه | True-False |
| evapcoolingportable | کولر آبی متحرک | True-False |
| splitportable | کولر گازی متحرک | True-False |
| dishwasher | ماشین ظرفشویی | True-False |
| microwave | مایکروویو | True-False |
| none | هیچکدام | True-False |
| pipewater | آب لوله کشی | True-False |
| electricity | برق | True-False |
| pipegas | گازلوله کشی | True-False |
| telephone | تلفن | True-False |
| internet | اینترنت | True-False |
| bathroom | حمام | True-False |
| kitchen | آشپزخانه | True-False |
| evapcooling | کولرآبی ثابت | True-False |
| centralcooling | برودت مرکزی | True-False |
| centralheating | حرارت مرکزی | True-False |
| package | پکیج | True-False |
| split | کولر گازی ثابت | True-False |
| wastewater | شبکه فاضلاب شهری | True-False |
| cookingFuel | نوع سوخت عمده مصرفی خانوار (پخت و پز) | Oil-Gasoline-LiquidGas-NaturalGas-Electricity-Wood-AnimalOil-Coke-Other-None |
| HeatingFuel | نوع سوخت عمده مصرفی خانوار (گرما) | Oil-Gasoline-LiquidGas-NaturalGas-Electricity-Wood-AnimalOil-Coke-Other-None |
| waterheatingfuel | نوع سوخت عمده مصرفی خانوار (تهیه آب گرم ) | Oil-Gasoline-LiquidGas-NaturalGas-Electricity-Wood-AnimalOil-Coke-Other-None |

| **قسمت سوم بخش ۱ هزينه هاي خوراكي ودخانی وR1401P3S01 U1401P3S01 و**  **بخش ۲ هزينه‌های نوشیدنی‌های طبقه بندی نشده R1401P3S02 و U1401P3S02** | |
| --- | --- |
| **نام ستون** | **شرح** |
| Address | آدرس خانوار |
| Code | کدکالا |
| Purchased | طریق تهیه |
| Gram | مقدار گرم |
| KiloGram | مقدار کیلو |
| Price | قیمت واحد (ریال ) |
| Value | ارزش (ریال ) |

| **قسمت سوم: بخش‌های ۳و ۵و ۶و ۷و ۸و ۹و ۱۱و ۱۲ - R1401P3S03 تا R1401P3S12 و U1401P3S03 تا U1401P3S12** | |
| --- | --- |
| **نام ستون** | **شرح** |
| Address | آدرس خانوار |
| Code | کدکالا |
| Purchased | طریق تهیه |
| Value | ارزش (ریال) |

| **قسمت سوم: بخش ۴ هزينه‌های بخش مسكن R1401P3S04 و U1401P3S04** | | |
| --- | --- | --- |
| **نام ستون** | **شرح** | **مقادیر** |
| Address | آدرس خانوار | - |
| Code | کدکالا | - |
| Mortgage | مبلغ رهن | - |
| Purchased | طریق تهیه | ۱:خرید- ۳و ۴و ۵:در برابر خدمت- ۸:رایگان |
| Value | ارزش (ریال) |  |

| **قسمت سوم: بخش ۱۳ هزينه ها R1401P3S13 و U1401P3S13** | |
| --- | --- |
| **نام ستون** | **شرح** |
| Address | آدرس خانوار |
| Code | کدکالا |
| Value | خرید یا هزینه |

| **قسمت سوم: بخش ۱۴ سرمايه گذاري خانوار در۱۲ ماه گذشته R1401P3S14 و U1401P3S14** | |
| --- | --- |
| **نام ستون** | **شرح** |
| Address | آدرس خانوار |
| Code | کدکالا |
| Purchased | طریق تهیه |
| Value | خرید یا هزینه |
| Sales | فروش دست دوم |

| **قسمت چهارم: بخش ۱ درآمد پولي اعضاي شاغل خانوار از مشاغل مزد و حقوق بگیری R1401P4S1 و U1401P4S1** | | |
| --- | --- | --- |
| **نام ستون** | **شرح** | **مقادیر** |
| Address | آدرس | - |
| Member | شماره ردیف عضو شاغل | - |
| Employed\_w | آیا در حال حاضر شاغل است؟ | ۱: بلی- ۲:خیر |
| ISCO\_w | کد شغل | - |
| ISIC\_w | کد فعالیت اصلی محل کار | - |
| status\_w | وضع شغلی | ۱: بخش عمومی- ۲: بخش تعاونی- ۳: بخش خصوصی |
| hours\_w | میزان ساعتهای کار در روز | - |
| days\_w | تعداد روزهای کار در هفته | - |
| income\_w\_m | مجموع درآمدهای ناخالص مستمر و غیرمستمر قبل از کسورات (ماه گذشته) | - |
| income\_w\_y | مجموع درآمدهای ناخالص مستمر و غیرمستمر قبل از کسورات 12(ماه گذشته) | - |
| wage\_w\_m | مزد و حقوق و مزایای مستمر (ماه گذشته) | - |
| wage\_w\_y | مزد و حقوق و مزایای مستمر 12(ماه گذشته) | - |
| perk\_w\_m | مزایای غیر مستمر (ماه گذشته) | - |
| perk\_w\_y | مزایای غیر مستمر 12(ماه گذشته) | - |
| netincome\_w\_m | مجموع درآمد خالص (ماه گذشته) | - |
| netincome\_w\_y | مجموع درآمد خالص 12(ماه گذشته) | - |

| **قسمت چهارم: بخش ۲ درآمد پولی اعضای شاغل خانوار از مشاغل آزاد U1401P4S2و R1401P4S2** | | |
| --- | --- | --- |
| **نام ستون** | **شرح** | **مقادیر** |
| Address | آدرس | - |
| Member | شماره ردیف عضو شاغل | - |
| Employed\_s | آیا در حال حاضر شاغل است ؟ | ۱: بلی- ۲: خیر |
| ISCO\_s | کد شغل | - |
| ISIC\_s | کد فعالیت اصلی محل کار | - |
| Status\_s | وضع شغلی | ۱:کارفرما- ۲: کارکن مستقل- ۳: کارکن فامیلی |
| agriculture | کشاورزی 1 غیر کشاورزی 2 | ۱:کشاورزی- ۲:غیر کشاورزی |
| Hours\_s | میزان ساعت‌های کار در روز | - |
| Days\_s | تعداد روزهای کار در هفته | - |
| cost\_employment | مزد و حقوق و مزایا | - |
| cost\_raw | بذر آب کود | - |
| cost\_machinery | تهیه ابزار کار بی دوام | - |
| cost\_others | کار مزد شغلی | - |
| cost\_tax | مالیات شغلی | - |
| sale | فروش (دریافتی ناخالص) | - |
| income\_s\_y | درآمد خالص | - |

| **قسمت چهارم: بخش ۳ درآمدهای متفرقه خانوار U1401P4S3 و R1401P4S3** | |
| --- | --- |
| **نام ستون** | **شرح** |
| Address | آدرس |
| Member | شماره ردیف اعضای خانوار |
| Income\_pension | حقوق بازنشستگی، حقوق وظیفه و آماده به خدمت، و ... |
| Income\_rent | درآمد حاصل از اجاره محل کسب، باغ، زمین و ... |
| income\_interest | درآمد حاصل از حساب پس انداز سپرده ثابت، سهام، و ... |
| income\_aid | کمک هزینه تحصیلی و ... |
| income\_resale | درآمد حاصل از محل فروش مصنوعات ... |
| income\_transfer | دریافتی های انتقالی از خانوارهای دیگر |

| **قسمت چهارم (ستون 9 بخش )3 U1401P4S4 و R1401P4S4** | |
| --- | --- |
| **نام ستون** | **شرح** |
| Address | آدرس |
| Member | شماره ردیف اعضای خانوار |
| subsidy\_number | تعداد افرادی که یارانه دریافت نموده‌اند |
| subsidy\_month | تعداد ماه دریافت یارانه |
| subsidy | کل مبلغ دریافتی |

**بخش دوم: تحلیل‌های آماری**

در این بخش، به کمک دانش آماری می‌خواهیم به تعدادی از سوال‌ها پاسخ دهیم؛ برخی از سوالات برای درک و یافتن شهود از داده‌ها پرسیده شده است، برخی دیگر از سمت یک شخص خاص، و در انتها تعدادی فرضیه مطرح شده است که شما باید آن‌ها را اعتبارسنجی کنید.

**آمار توصیفی**

1. توزیع اطلاعات خانوار (میزان تحصیلات، سن، بستگی با سرپرست، وضع فعاليت) را رسم نمایید.
2. ترند هزینه خانوار برای در هر سال برای غذاهاي آماده، هتل و رستوران رسم نمایید.
3. ماتریس هم‌بستگی را برای هزینه‌های پوشاك و كفش خانوار، هزينه‌هاي خوراكي خانوار، هزينه‌های مسكن، آب، فاضلاب، سوخت و روشنايي خانوار و هزينه‌هاي بهداشتي و درماني خانوار رسم نمایید.

**آزمون فرض**

* به نظر شما درآمد خانوار های شهری و روستایی در استان چهارمحال و بختیاری با هم برابر است؟

**بخش سوم:یادگیری ماشین**

**مسئله‌ی ۱: خوشه‌بندی**

در سوال اول فرض کنید که شما دهک‌بندی فعلی خانوار ها را جالب نمی‌دانید در نتیجه می‌خواهید بر اساس تمام هزینه‌های خانوار و درآمد خانوار از سال‌های 1398 , 1401 خانوار ها را خوشه بندی کنید. توجه کنید که ستون هزینه و درآمد هر خانواده را خودتان با روش مناسبی که پیشنهاد می‌دهید باید محاسبه کنید.

**بخش ۱)**

حال الگوریتم خوشه‌بندی K-means را تنها بر حسب با **۱۰ خوشه** برای این مجموعه‌داده اجرا کنید. سپس بر روی اسکتر پلات رسم‌شده مشخص کنید کدام نقاط مربوط به کدام خوشه هستند و مرکز هر خوشه را نیز رسم کنید. به انتخاب رنگ، مارکر، نام‌گذاری محور‌ها و به‌طور کلی قابل درک بودن تصویر دقت داشته باشید.

**بخش ۲)**

پس از آن الگوریتم K-means را برای k هایی از ۱ تا ۲۰ اجرا کرده و با محاسبه‌ی *مجموع مجذورات درون خوشه‌ای (Within-Cluster Sum of Square)*، مقدار مناسبی برای هایپرپارامتر k انتخاب کنید. توجه کنید که بخش زیادی از نمره‌ی این بخش مربوط به نحوه‌ی انتخاب مقدار k است و چنان‌چه روش‌های تدریس‌شده و معمول پاسخگوی حل مسئله نبود،انتظار می‌رود با جستجو و مطالعه‌ی بیشتر، روشی مناسب برای رفع چالش‌های احتمالی پیشنهاد دهید.

**بخش ۳)**

در آخرین گام از این سوال از شما می‌خواهیم با استفاده از روش DBScan داده‌ها را خوشه‌بندی کنید و هایپرپارامتر‌ها را به‌نحوی تغییر دهید که ۱۰ کلاستر بامعنا در خروجی تولید شود. اسکتر پلات داده‌ها و نحوه‌ی خوشه‌بندی آن‌ها را رسم کنید. نحوه‌ی اثرگذاری هر یک از هایپرپارامتر‌ها بر خروجی را توضیح دهید.

**مسئله‌ی ۲: پیش‌بینی**

پیش‌بینی میزان هزینه‌ای که خانوار در هر حوزه‌ای انجام می‌دهد، می‌تواند به شما کمک کند تا تصمیم بگیرید در آن حوزه سرمایه گذاری کنید یا خیر.

در این بخش از شما می‌خواهیم مدلی آموزش دهید که با توجه به اطلاعات دریافتی از خانوار پیش‌بینی کند میزان هزینه خانوار در حوزه حمل و نقل در یک ماه چقدر است؟

شما مجاز هستید از هر کدام از الگوریتم‌های یادگیری ماشین که تاکنون در کلاس‌های بوت‌کمپ آموخته‌اید برای مدل‌سازی استفاده کنید.

| **توجه:** استفاده از الگوریتمی غیر از الگوریتم‌های اصلی‌‌ای که در کلاس‌ها آموزش داده شده‌اند در بخش اصلی مجاز نیست. در صورت علاقه و تسلط می‌توانید از آن‌ها برای بخش امتیازی استفاده کنید. البته توجه داشته باشید که نیاز است تمام اعضای گروه نسبت به نحوه‌ی کار آن الگوریتم دانش کافی داشته باشند. |
| --- |

در صورت نیاز می‌توانید هر ویژگی دلخواهی را به مجموعه‌داده اضافه کنید یا آن‌ها را مهندسی کنید. البته دقت کنید که ویژگی‌های ورودی مدل منجر به نشت متغیر هدف نشود.

به‌منظور ارزیابی مدل نهایی خود از داده‌های مربوط به تاریخ زمستان ۱۴۰۱ به‌عنوان مجموعه‌ی آزمون استفاده کنید.

با مقایسه‌ی پیش‌بینی مدل خود با مقادیر حقیقی برای داده های تست نمودارهای R\_2، loss را رسم نمایید . نیاز است در زمان ارائه تحلیل مناسبی از نتایج به‌دست‌آمده ارائه دهید.

| **توجه:** در آزمایش‌های خود و انتخاب مدل و هایپرپارامترهای آن نباید از داده‌های آزمون (Test) استفاده کنید، بلکه این کار باید با داده‌های اعتبارسنجی (Validation) انجام گیرد. تنها پس از دست‌یابی به مدل نهایی خود از مجموعه‌ی آزمون بهره ببرید. |
| --- |

**نکته‌های کلی**

* کدهای خود را خوانا و تمیز بنویسید.
* مهم‌ترین بخش این پروژه، تحلیل و تفسیر شما از شرایط مسئله و نتایج آن است. باید بتوانید برای هر کدام از انتخاب‌های خود در طول مسیر، دلیلی موجه و علمی داشته باشید. ارائه‌ی شما نیز باید بر همین محور باشد، یعنی روند حل مسئله، نتایج و تحلیل و تفسیر را ارائه دهید، نه توضیح کد.
* به نکات ذکر شده در ارتباط با نحوه‌ی ارسال فایل در [صفحه‌ی پروژه در کلاس](https://quera.org/course/assignments/68517/problems) توجه فرمایید.

**بخش امتیازی (بیشینه: ۲۵ نمره)**

* مستندسازی غنی و مناسب در نت‌بوک‌ها (۲ نمره)
* استفاده از گیت و مشارکت فعال در آن (۲ نمره)
* استخراج و اضافه کردن ویژگی‌های مناسب و بامعنا (۲ نمره)
* استفاده از مدل‌های حرفه‌ای‌تر و دست‌یابی به نتایج بهتر با تسلط کامل اعضای گروه به الگوریتم (۷ نمره)
* طرح مسئله‌ای جدید با توجه به داده‌های موجود و مرتبط (با تایید منتور) و دست‌یابی به نتایج قابل قبول و تفسیرپذیر (۱۰ نمره)
* ارائه‌ای جذاب با بهره‌گیری از خط داستانی و استفاده از ابزارهای مناسب ارائه همچون اسلاید (۲ نمره)

موفق باشید 🤩