

**گروه کامپیوتر** – **دانشکده مهندسی** – **دانشگاه فردوسی مشهد**

**گزارش فاز اول پروژه**

**درس: مبانی داده کاوی**

**ارائه دهندگان:**

**سبحان مرادیان دقیق - فاطمه چاجی**

**استاد:**

**دکتر بهشید بهکمال**

**پاییز 1399**

# **توضيح پروژه انجام شده**

پروژه انجام شده در [آدرس گيت هاب](https://github.com/nahbos/Data-Mining-on-DigiKala-dataset) قابل مشاهده است.

# **بیان راه‌کارهای برخورد با داده‌های غیر معتبر**

**2.1. دیتاست محصولات**

* **ويژگي عنوان فارسی محصول**

فيلدهايي با عنوان فارسي نامفهوم را حذف مي‌كنيم.

* **ويژگي های عنوان انگلیسی محصول و عنوان انگلیسی نام تجاری محصول**

با توجه به قواین تعریف شده در کد برای معتبر بودن، فیلدهای دارای مقدار نامعتبر را حذف می‌کنیم

* **ويژگي های كد لينك و title\_alt**

با توجه به قواین تعریف شده در کد برای معتبر بودن، فیلدهای دارای مقدار نامعتبر را حذف می‌کنیم.

**2.2. دیتاست تاریخچه خرید مشتریان**

* **ویژگی میزان سفارشات ناخالص**

برای مقادیر معتبر رنجی تعریف کرده ایم و مقادیر خارج از این رنج که غیر معتبر می‌باشند را حذف می‌کنیم.

* **ویژگی تعداد خریداری شده از هر محصول**

با این که تقریبا تمامی مقادیر گزارش شده برای این ویژگی کوچک می‌باشند، تعداد محدودی به طور غیر طبیعی بزرگ می‌باشند. در برخورد با این داده های غیر معتبر، آن ها را حذف می‌کنیم.

**2.3. دیتاست تاریخچه محصولات**

* **ویژگی قیمت فروش محصولات**

براي تشخيص معتبر بودن داده ها از آستانه گذاري بر روي آن ها استفاده شده است و مقادير كمتر از 800 نامعتبر شناخته مي‌شوند.

براي برخورد با اين داده هاي غير معتبر مي‌توان آن ها را حذف كرد اما اگر بخواهيم با اطمينان خاطر بيشتري حذف داده ها را انجام دهيم، اين آستانه را بجاي 800 بر روي مقداري مانند 500 قرار داده و مقادير كمتر از اين آستانه را حذف می‌کنیم.

* **ویژگی rrp\_price**

مقادیر این ویژگی مشابه با ویژگی قیمت فروش محصولات هستند پس بطور مشابه رفتار می‌کنیم.

* **ویژگی قیمت پایه محصولات**

مقادیر NULL به دلیل اختیاری بودن اعلام قیمت محصولات می‌باشند اما تعدادی مقادیر دیگر که صفر گزارش شده اند، نامعتبر می‌باشند و ما در برخورد با این مقادیر آن ها را نیز NULL تنظیم می‌کنیم.

* **ویژگی قیمت خرید محصولات**

مشابه با ویژگی بالا می‌باشد.

* **ویژگی میزان محدودیت سفارش هر محصول**

برای بررسی معتبر بودن مقادیر این ویژگی، از آستانه ای برابر با 500 استفاده شده و مقادیر بالاتر از این آستانه به دلیل غیر منطقی بودن، نامعتبر شناخته می‌شوند.

**2.4. دیتاست کیفیت محصولات**

* **ویژگی عنوان محصول**

تعداد محدودی از عناوین گزارش شده برای این ویژگی نامفهوم می‌باشند. در برخورد با این داده های غیر معتبر، آن ها را حذف می‌کنیم.

* **ویژگی نظرات درباره محصول**

تعداد محدودی از نظرات گزارش شده نامفهوم می‌باشند. در برخورد با این داده های غیر معتبر، آن ها را حذف می‌کنیم.

* **ویژگی های مزایا و معایب محصول**

مقادیر NULL به دلیل اختیاری بودن گزارش مزایا و معایب محصولات می‌باشند اما تعدادی از گزارش ها که خالی یا نامفهوم هستند، نامعتبر می‌باشند و ما در برخورد با این مقادیر آن ها را نیز NULL تنظیم می‌کنیم.

**2.5. دیتاست نظرات**

* **ویژگی نظرات**

تنها یکی از مقادیر وارد شده نامعتبر است چرا که "." گزارش شده است. این داده غیر معتبر را حذف می‌کنیم.

# **ارزیابی صحت داده ها**

**3.1. دیتاست محصولات**

**3.2. دیتاست تاریخچه خرید مشتریان**

* **ویژگی های شناسه خرید، شناسه سفارش، شناسه محصول**

مقادیر این شناسه ها باید یکتا باشند و تکراری بودن یک مقدار در یکی از ویژگی ها، نشان دهنده غیر صحیح بودن آن است.

* **ویژگی زمان تحویل محصول**

با فرض آن‌که تاریخ تاسیس سایت دیجی کالا سال 2012 بوده است، تاریخ سفارش ثبت شده پیش از این زمان غیر صحیح می‌باشد که این امر در کد زده شده نیز پوشش داده شده است.

* **ویژگی نام شهر**

برای بررسی صحت مقدار وارد شده برای این ویژگی، يك مقدار آستانه تعريف مي‌كنيم. از آن جا كه اكثر اسامي بكار رفته براي اين ويژگي به ميزان زياد تكرار شده اند، اسامي تكرار شده كمتر از آستانه مساوي 2، صحيح نمي‌باشند و به درستي وارد نشده اند.

**3.3. دیتاست تاریخچه محصولات**

* **ویژگی های شناسه، شناسه متغیر محصولات، شناسه محصول و شناسه فروشنده کلی محصول**

مقادیر این شناسه ها باید یکتا باشند و تکراری بودن یک مقدار در یکی از ویژگی ها، نشان دهنده غیر صحیح بودن آن است.

* **ویژگی میزان محدودیت سفارش هر محصول**

با توجه به قیمت فروش هر محصول، در صورتی که قیمت فروش بالا باشد محدودیت سفارش محصول نمی‌تواند بالا باشد. برای مثال محصول با قیمت 10 میلیون، محدودیت سفارش بالای 30 نخواهد داشت چرا که منطقا به این میزان در انبار فروشگاه نیز موجود نخواهد بود. بنابراین با توجه به قیمت محصول، می‌توانیم برای میزان محدودیت سفارش، آستانه تعیین کنیم.

* **ویژگی هاي selling\_price و rrp\_price**

با مقایسه اين دو ويژگي و با توجه به تعریف rrp\_price[[1]](#footnote-1)، می‌توان نتیجه گرفت همواره رابطه زیر بین مقادیر وارد شده برای این ویژگی برقرار است:

selling\_price >= rrp\_price

* **ویژگی هاي قیمت پایه و قیمت خرید محصولات**

با توجه به تعريف دو ویژگی buy\_priceو base\_price می‌توان نتیجه گرفت همواره رابطه زیر بین مقادیر وارد شده برای این ویژگی برقرار است:

buy\_price <= base\_price

* **ویژگی های تاریخ افزوده شدن محصول و تاریخ ایجاد محصول در سایت**

با توجه به تعريف این دو ویژگی می‌توان نتیجه گرفت همواره رابطه زیر بین مقادیر وارد شده برای این ویژگی برقرار است:

start\_at >= created\_at

**3.4. دیتاست کیفیت محصولات**

* **ویژگی شناسه محصول**

مقادیر این شناسه باید یکتا باشند و تکراری بودن یک مقدار در اين ویژگی، نشان دهنده غیر صحیح بودن آن است.

* **ویژگی عنوان انگليسي محصول**

مقادير مربوط به اين ويژگي همگي دو كاراكتري هستند اما الزاما هر دو كاراكتري براي اين ويژگي مقداري صحيح نمي‌باشند. براي بررسي صحيح بودن مقادير اين ويژگي، بايد از ديتاست ديگري از ديجي كالا كه شامل عناوين انگليسي محصولات موجود مي‌باشد، كمك گرفت.

* **میزان دوست داشتن و نداشتن یک محصول**

از آنجا كه ممكن است ديتاست به طريقي مثلا توسط يك بدافزار دستكاري شده باشد، براي ميزان دوست داشتن و نداشتن محصول يك آستانه تعريف مي‌كنيم. بدين صورت كه هر عددي بالاتر از 3 برابر ميزان فروش محصول، مقداري صحيح براي ميزان دوست داشتن و نداشتن نمي‌باشد.

**3.5. دیتاست نظرات**

* **ویژگی شناسه محصول**

مقادیر این شناسه باید یکتا باشند و تکراری بودن یک مقدار در اين ویژگی، نشان دهنده غیر صحیح بودن آن است.

* **ویژگی زمان ثبت نظر**

با فرض آن‌که تاریخ تاسیس سایت دیجی کالا سال 2012 بوده است، تاریخ نظر ثبت شده پیش از این زمان غیر صحیح می‌باشد.

# **بیان روش مناسب** **ارزیابی کامل بودن داده ها**

**4.1. دیتاست محصولات**

تعداد زیادی داده از دست رفته در ستون "ویژگی های محصول" وجود ندارد که نمی توانیم کاری برای آنها انجام دهیم. علاوه بر این، این ویژگی یک ویژگی لازم برای دیتاست ماست، بنابراین نمی توانیم از این ویژگی چشم پوشی کنیم. بنابراین با استفاده از استراتژی سوم میزان کامل بودن را اندازه گیری کنیم یعنی:

A: attribute fields containing values | B: records × attributes

**4.2. دیتاست تاریخچه خرید مشتریان**

اين ديتاست كامل مي‌باشد!

**4.3. دیتاست تاریخچه محصولات**

ما دریافتیم که 5 ویژگی در این پایگاه داده تقریباً به طور کامل NULL هستند و به نظر این 5 ویژگی مهم نیستند.

بنابراین از این ویژگی ها چشم پوشی می‌کنیم و استراتژی دوم را دنبال کنیم یعنی:

A: number of data required for the particular context in the data file | B: number of data in the specified particular context of intended use

**4.4. دیتاست کیفیت محصولات**

ما فکر می کنیم ويژگي هاي نظرات درباره محصول و عنوان انگلیسی محصول برای ما مهم هستند اما ويژگي هاي مزایا و معایب محصولات ويژگي هايي ضروري نيستند و می توانند NULL باشد، بنابراین ما از این 2 ستون چشم پوشی می کنیم اما مقادیر از دست رفته ويژگي هاي نظرات درباره محصول و عنوان انگلیسی محصول را حفظ می کنیم. در نهايت برای سنجش میزان کامل بودن از استراتژی دوم استفاده مي‌كنيم.

**4.5. دیتاست نظرات**

مهم‌ترین ویژگی در این پایگاه داده، ویژگی نظر است که دارای 105 داده NULL می‌باشد. ما نمی توانیم این ویژگی مهم را نادیده بگیریم بنابراین از استراتژي اول استفاده می کنیم.

1. recommended retail price: the price that the company that makes a product says it should be sold for [↑](#footnote-ref-1)