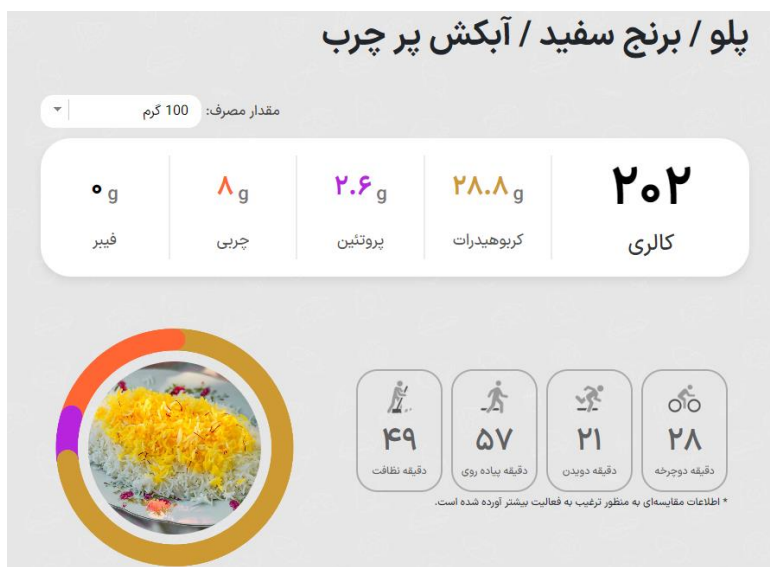


بنام خداوند بخشایگر

موضوع : گزارش پروژه ی درس داده کاوی

استاد مربوطه : دکتر شعاران

*سعید عمادی



خلاصه :

در این پروژه سعی شده است که با جمع آوری داده های موجود درمورد میزان کالری مواد خوراکی و نوشیدنی ها و همچنین میزان مواد موجود در آن ها (پروتئین، چربی، فیبر و کربوهیدرات) و اجرای مدل های آماری روی آن ها و پیدا کردن رابطه با میزان کالری با خوراکی ها و نوشیدنی ها و بلکه فعالیت های جسمی متداول ، پیشبینی درمورد میزان فعالیت و یا تناسب میزان کالری با مواد موجود انجام بدهیم.

مقدمه :

شکل 1 : این شکل نشان دهنده ی اطلاعات مورد استفاده درمورد یک مثال می باشد.

در جوامع امروزی، چاقی و افزایش وزن به معضلات بزرگی تبدیل شده است .این امر به عوامل متعددی از جمله رژیم غذایی ناسالم، کمبود فعالیت بدنی و شیوه زندگی بی تحرک مرتبط است. نقش کالری غذاهاد و خوراکی ها در این میان بسیار کلیدی است.هر ماده غذایی حاوی کالری است که واحد سنجش انرژی است . بدن ما برای عملکرد به این انرژی نیاز دارد و آن را از طریق کالری موجود در غذا و نوشیدنی ها به دست می آورد.مشکل زمانی شروع می شود که کالری دریافتی ما از کالری سوزانده شده بیشتر باشد .این امر منجر به ذخیره کالری اضافی به صورت چربی در بدن می شود . با گذشت زمان، این چربی ها می توانند منجر به افزایش وزن و چاقی شوند.

چاقی و اضافه وزن می توانند خطر ابتلا به بیماری های متعددی از جمله موارد اشاره شده را افزایش دهد.

- ❖ بیماری های قلبی: چاقی یکی از اصلی ترین عوامل خطر بیماری های قلبی است .
- ❖ دیابت نوع 2: چاقی خطر ابتلا به دیابت نوع 2 را به طور قابل توجهی افزایش می دهد .
- ❖ برخی انواع سرطان: چاقی با افزایش خطر ابتلا به برخی انواع سرطان از جمله سرطان روده بزرگ، سینه و اندام رحم مرتبط است .
- ❖ فشار خون بالا: چاقی یکی از عوامل اصلی فشار خون بالا است .

- ❖ **آپنه خواب :** چاقی خطر ابتلا به آپنه خواب، یک اختلال تنفسی جدی را افزایش می دهد .
- ❖ **آرتروز :** اضافه وزن فشار روی مفاصل را افزایش می دهد و می تواند منجر به آرتروز شود.

کاهش مصرف کالری روزانه، یکی از راه های اصلی برای **کاهش وزن و حفظ وزن سالم** است . با مصرف کالری کمتر از آنچه می سوزانید، بدن شما مجبور می شود از ذخایر چربی خود برای تامین انرژی استفاده کند، که منجر به کاهش وزن می شود. اما فواید کاهش کالری روزانه فراتر از لاغر شدن است . این کار می تواند تاثیر مثبتی بر **سلامتی کلی** شما نیز داشته باشد.

در این پروژه سعی شده است که با جمع آوری داده های موجود درمورد میزان کالری مواد خوراکی و نوشیدنی ها و همچنین میزان مواد موجود در آن ها (پروتئین، چربی، فیبر و کربوهیدرات) و اجرای مدل های آماری روی آن ها و پیدا کردن رابطه با میزان کالری با خوراکی ها و نوشیدنی ها و بلکه فعالیت های جسمی متداول ، پیشبینی درمورد میزان فعالیت و یا تناسب میزان کالری با مواد موجود انجام بدهیم تا اطلاعات لازم درمورد میزان فعالیت برای جلوگیری از چاق شدن و یا با میزان فعالیت برای مصرف مقدار کالری را به دست آوریم (برای نمونه شکل 2).

```
[27]: # random forest is good > predict as
carbo = 8.
protein = 16.
fat = 16.
fiber = 1.4
print(f"calory for {carbo}g carbo {protein} protein {fat}g fat and {fiber}g fiber ::\n
{randomReg.predict([[carbo,protein,fat,fiber]])[0]}")

calory for 8.0g carbo 16.0 protein 16.0g fat and 1.4g fiber ::      277.69

[42]: # predict calory
ridingBike = 0 # activity1
run = 0 # activity2
walking = 10 # activity3
cleaningUp = 30 # activity4
calory = randomReg.predict([[ridingBike, run, walking, cleaningUp]])[0]
print(calory)

59.49217212582815
```

شکل 2 : پیش بینی درمورد کالری موجود در میزان کربوهیدرات، پروتئین، چربی و فیبر (بالا) و میزان کالری سوزانده شده در اثر انجام فعالیت های دوچرخه سواری، دویدن، پیاده روی و نظافت خانه (پایین).

➤ لازم به ذکر است که میزان کالری درج شده بر حسب کیلو (kCal) بوده، واحد های کربوهیدرات، پروتئین، چربی و فیبر بر حسب گرم (g) و میزان فعالیت ها بر حسب دقیقه (minute) می باشد.

خط زمانی داستان پروژه :

1. مجموعه داده (Dataset)

برای این پروژه ما سعی کرده ایم تا مفاهیم جمع آوری اطلاعات و نمونه داده هارا به جا بیاوریم به همین جهت تلاش کرده ایم تا اطلاعات مورد نظر درمورد خوراکی ها و نوشیدنی ها از سایت مانکن¹ با استفاده از تکنیک های Web Scraping استخراج بکنیم.

کد های مربوط به Web Scraping در فایل spider-man.py قرار دارد.

¹ <https://mankan.me/>

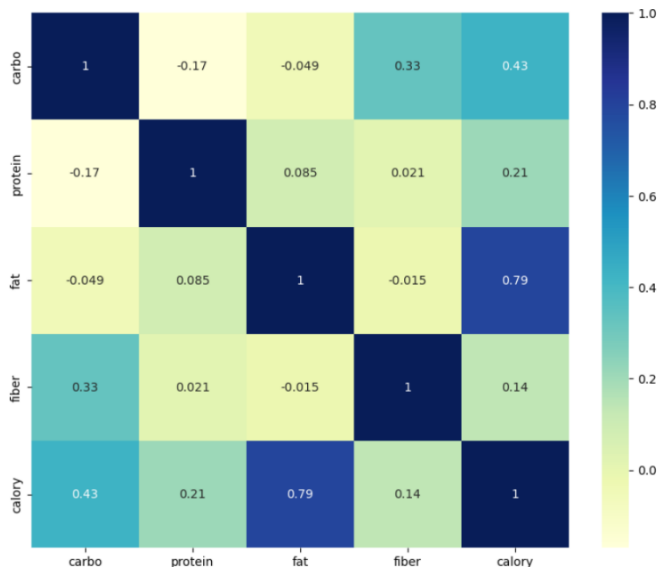
<class 'pandas.core.frame.DataFrame'> Int64Index: 1754 entries, 0 to 1753 Data columns (total 11 columns): # Column Non-Null Count Dtype --- 0 siteId 1754 non-null int64 1 name 1754 non-null object 2 calory 1754 non-null float64 3 carbo 1754 non-null float64 4 protein 1754 non-null float64 5 fat 1754 non-null float64 6 fiber 1754 non-null float64 7 activity1 1754 non-null float64 8 activity2 1754 non-null float64 9 activity3 1754 non-null float64 10 activity4 1754 non-null float64 dtypes: float64(9), int64(1), object(1) memory usage: 164.4+ KB					<class 'pandas.core.frame.DataFrame'> RangeIndex: 1821 entries, 0 to 1820 Data columns (total 11 columns): # Column Non-Null Count Dtype --- 0 siteId 1821 non-null int64 1 name 1754 non-null object 2 calory 1754 non-null float64 3 carbo 1754 non-null float64 4 protein 1754 non-null float64 5 fat 1754 non-null float64 6 fiber 1754 non-null float64 7 activity1 1754 non-null float64 8 activity2 1754 non-null float64 9 activity3 1754 non-null float64 10 activity4 1754 non-null float64 dtypes: float64(9), int64(1), object(1) memory usage: 156.6+ KB				
--	--	--	--	--	--	--	--	--	--

اطلاعات استخراج شده شامل 10 ستون و 1821 رکورد می باشد. ستون ها حاوی اطلاعات آیدی آیتم، نام آیتم، مقدار کالری، کربوهیدرات، پروتئین، چربی و فیبر آیتم در واحد 100 گرم و همچنین مدت زمان فعالیت ها (به ترتیب دوچرخه سواری، دویدن، پیاده روی، نظافت خانه) می باشد.

داده ها در فایل Mankan_dataset.csv ذخیره شده اند.

شکل 3 : اطاعات تعداد رکورد موجود در مجموعه داده و اطلاعات ستون ها قبل و بعد از پاکسازی مجموعه داده.

بعد از ذخیره سازی داده های استخراج شده، داده ها را برای پردازش آماده میکنیم که، بعد از پاکسازی داده ها ، تعداد رکورد ها به 1754 رکورد کاهش پیدا میکند (رکوردهای حاوی مقادیر خالی و داده های پرت).



چالش های این مرحله :

- رکورد های حاوی مقادیر خالی
- رکورد های حاوی داده های پرت (outlier data)
- رکورد های کم برای آموزش و تست

2. همبستگی ستون ها و ارتباط آن ها با یکدیگر

سعی میکنیم تا رابطه میان ستون ها و تاثیر آن ها با یکدیگر را بررسی بکنیم. شکل 4 نشان دهنده ی میزان وابستگی ستون ها به یکدیگر را نشان می دهد. مشهود می باشد که میزان چربی موجود در مواد تاثیر بسیار زیادی در میزان کالری آن دارد. البته لازم به ذکر است که بقیه ستون ها نیز تاثیر دارند.

شکل 4 : همبستگی میان چندین ستون (کالری، پروتئین، چربی، فیبر و کربوهیدرات).

3. پیش بینی مقدار کالری خوراکی یا نوشیدنی بر حسب میزان مواد موجود در آن

برای پیش بینی مقدار کالری بر حسب میزان مواد موجود در آن، از دو مدل رگرسیون خطی (LinearRegression) و جنگل تصادفی (RandomForestRegressor) استفاده شده است. با توجه به دقت های به دست آمده (درج شده در جدول زیر) ما جنگل تصادفی را برای پیش بینی استفاده میکنیم.

مدل	RandomForestRegressor	LinearRegression
دقت	93.81 %	91.22 %

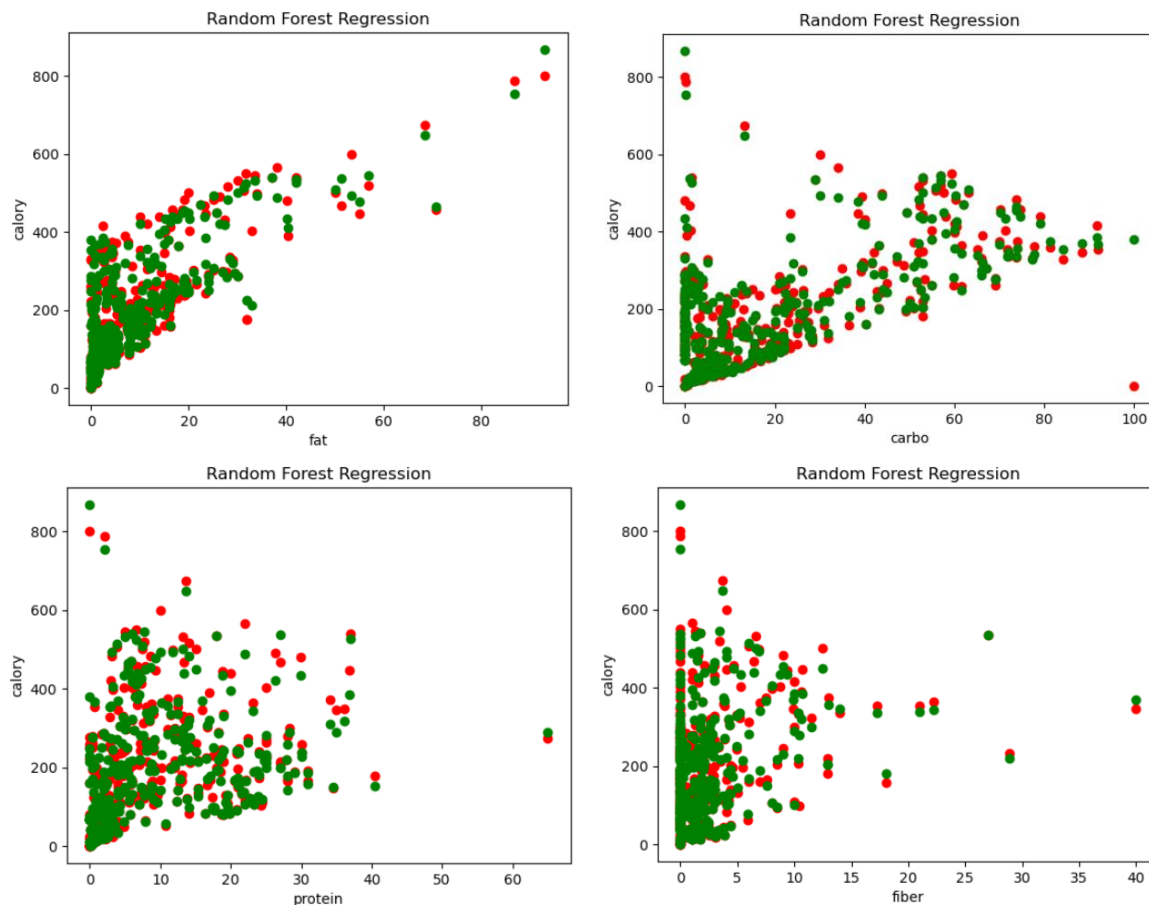
ارزیابی مدل با استفاده داده های تست :

شکل های این صفحه نشان دهنده ی خطا و مقدار اختلاف پیش بینی با مقدار اصلی داده ها می باشند.

لازم به ذکر است که با اضافه کردن یک مرحله ی دیگر به پیش پردازش داده ها، دقت پیش بینی افزایش داشته و میزان اختلاف پیش بینی کاهش داشته است (شکل ها نشان دهنده ی مرحله ی اول پیش پردازش داده می باشند تا داده های پرت در شکل گزارش نشان داده شوند).

Mean Absolute Error : 17.718561475785048
Mean Squared Error : 1366.9100181436536
R-squared : 0.9381796282650858

شکل 5 : میزان خطای پیش بینی مدل در داده های تست بر اساس سه معیار ذکر شده.



شکل 6 : نمایش مقدار پیش بینی کالری بر حسب چربی، کربوهیدرات، پروتئین و فیبر و مقدار اصلی کالری موجود بر حسب مواد ذکر شده که نشان دهنده ی میزان نزدیکی یا دور بودن مقدار پیش بینی با مقدار واقعی می باشد. مشاهده می شود که مقادیر پیش بینی شده بسیار نزدیک به مقادیر واقعی هستند .

برای نمونه :

مقدار کالری را برای مقادیر 8 گرم کربوهیدرات، 20 گرم پروتئین، 16 گرم چربی و 1.4 گرم فیبر پیش بینی میکنیم.

مقدار کالری پیش بینی شده برابر : 275.9 کیلوکالری می باشد، که میتوان نمونه های نزدیک به آن و مقادیر اصلی آن را در مجموعه داده ی اصلی مشاهده کرد (شکل 7).

calory for 8.0g carbo 20.0 protein 16.0g fat and 1.4g fiber :: 275.9										
[81]:										
	name	calory	carbo	protein	fat	fiber	activity1	activity2	activity3	activity4
1100	خوراک قارچ و گوشت	285.0	8.0	16.0	16.0	1.4	40.0	30.0	81.0	69.0
1121	ران مرغ سوخاری	277.0	9.0	22.0	17.0	0.3	39.0	29.0	78.0	67.0
1198	پنیر سویا (توفو) سرخ شده	271.0	10.0	17.0	20.0	3.9	38.0	28.0	77.0	66.0
1298	خوراک گوشت	285.0	8.0	16.0	16.0	1.4	40.0	30.0	81.0	69.0

شکل 7 : میزان کالری پیش بینی شده و نمایش خوراکی ها و نوشیدنی هایی که مقادیر کربوهیدرات، پروتئین، چربی و فیبر آن ها با مقادیر ورودی برای پیش بینی نزدیک هستند جهت مقایسه مقادیر کالری آن ها با کالری پیش بینی شده.

4. پیش بینی مقدار کالری سوزانده شده در بدن بر حسب میزان فعالیت های متداول

برای پیش بینی مقدار کالری سوزانده شده بر حسب میزان فعالیت های متداول (دوچرخه سواری، دویدن، پیاده روی و نظافت خانه) از مدل جنگل تصادفی (RandomForestRegressor) استفاده شده است. دقت به دست آمده برای این مدل برابر 99.9 % بوده است.

ازیابی مدل با استفاده از داده های تست :

شکل های زیر نشان دهنده ی خطا و مقدار اختلاف پیش بینی با مقدار اصلی داده ها می باشند.

Mean Absolute Error: 0.6358511369801789
Mean Squared Error: 2.7473165255252514
R-squared: 0.9998757488593784

شکل 8 : میزان خطای پیش بینی مدل در داده های تست بر اساس سه معیار ذکر شده.

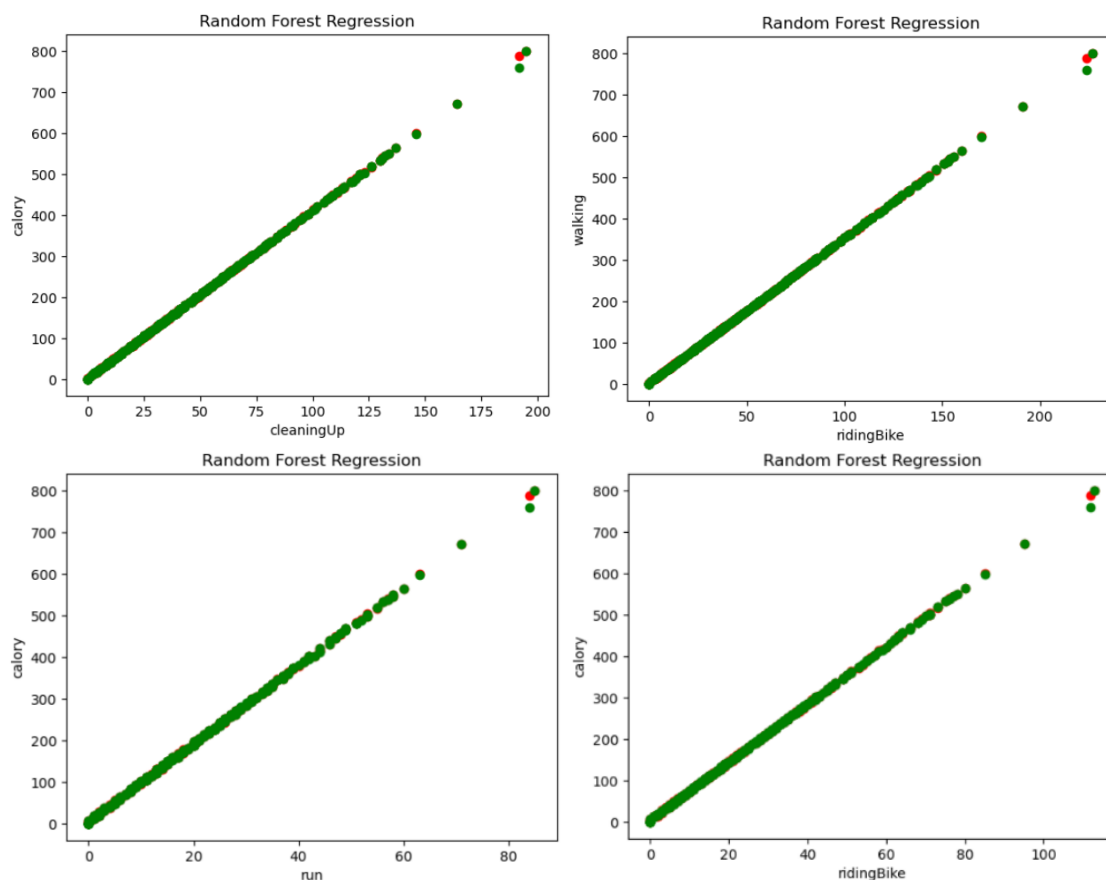
برای نمونه :

مقدار کالری سوزانده شده برای 10 دقیقه پیاده روی، 30 دقیقه نظافت منزل پیش بینی میکنیم.

مقدار کالری پیش بینی شده برابر : 52.23 کیلوکالری می باشد، که میتوان نمونه های نزدیک به آن و مقادیر اصلی آن را در مجموعه داده ی اصلی مشاهده کرد (شکل 9).

calory for 0m ridingBike 0m run 10m walking and 30m cleaningUp :: 52.22910108604845										
[100]:										
	name	calory	carbo	protein	fat	fiber	ridingBike	run	walking	cleaningUp
4	تخم مرغ سفیده خام	52.0	0.7	11.0	0.2	0.0	7.0	5.0	14.0	12.0
15	سوپ خامه با مرغ	64.0	10.4	8.9	6.5	1.0	9.0	6.0	18.0	15.0
82	ماهی چهارگوش کباب شده با استخوان	59.0	0.0	14.0	0.3	0.0	8.0	6.0	16.0	14.0
---	---	---	---	---	---	---	---	---	---	---
1711	کلم بروکلی سرخ شده/ سوخاری	58.0	6.5	2.1	7.4	3.0	8.0	6.0	16.0	14.0
1726	کلم بروکسل سرخ شده	59.0	6.4	2.3	3.5	2.3	8.0	6.0	16.0	14.0
1737	لبو پخته	44.0	10.0	1.7	0.2	2.0	6.0	4.0	12.0	10.0
223 rows x 10 columns										

شکل 9 : میزان کالری پیش بینی شده و نمایش خوراکی ها و نوشیدنی هایی که میزان فعالیت آن ها با مقادیر ورودی برای پیش بینی نزدیک هستند جهت مقایسه مقادیر کالری آن ها با کالری پیش بینی شده.



شکل 10 : نمایش مقدار پیش بینی کالری بر حسب مقدار فعالیت های دوچرخه سواری، دویدن، پیاده روی و نظافت خانه و مقدار اصلی کالری موجود بر حسب فعالیت های ذکر شده که نشان دهنده ی میزان نزدیکی یا دور بودن مقدار پیش بینی با مقدار واقعی می باشد. مشاهده می شود که مقادیر پیش بینی شده بسیار نزدیک به مقادیر واقعی هستند.

کد های مربوط به این نتایج در فایل [MankanResults.ipynb](#) قرار دارد.

5. چالش :

چالش هایی که در طول پیاده سازی با آن ها روبرو شدیم :

- مشکل کپچا در هنگام ارسال مکرر درخواست زیاد به سایت
 - حل : استفاده از تاخیر برای ارسال درخواست ها به سرور و همچنین در صورت بن شدن استفاده از پروکسی های مختلف برای ارسال درخواست ها.
- رکورد های حاوی مقادیر خالی
 - حل : بررسی سطرها و پاک کردن (یا مقدار دهی میانگین) رکورد مربوطه.
- رکورد های حاوی داده های پرت (outlier data)
 - حل : ویرایش کردن مقدار و بروزرسانی با مقدار میانگین آن ستون.
- رکورد هایی که کد گذاری آن ها UTF-8 نبود
 - حل : مقدار دهی دستی اسم رکورد.

- وجود رکورد های کم برای آموزش
 - حل : میتوان از k-fold استفاده کرد.
- ایجاد درخت های بسیار بزرگ
 - حل : انتخاب بچینه ترین و ساده ترین درخت ممکن.
- وجود ستون های اضافی و بدون استفاده در مجموعه
 - حل : حذف ستون های بدون استفاده در پردازش های مربوط.
- وجود چندین معیار مقداری از خوراکی ها و نوشیدنی ها
 - حل : استفاده از یک واحد مشترک برای تمامی خوراکی ها و نوشیدنی ها.

6. توجهات :

این مجموعه داده ای از سایت مانکن استخراج شده است که :

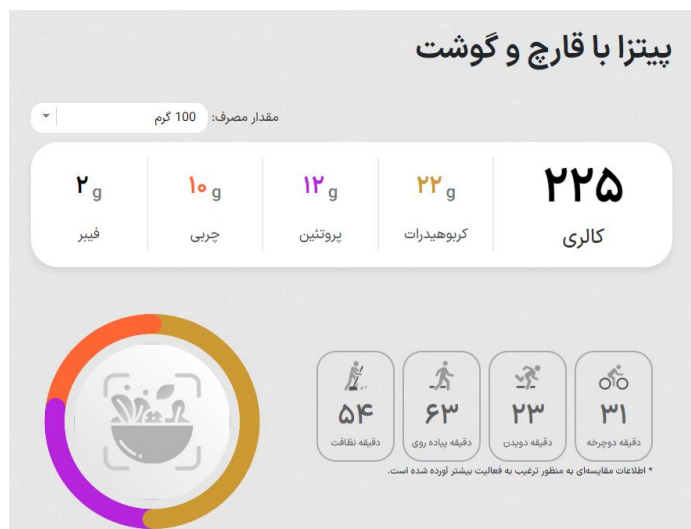
- برای پروتئین، فیبر، چربی، کربوهیدرات از واحد گرم
- برای کالری از واحد کیلو کالری
- برای فعالیت ها از واحد دقیقه

استفاده شده است.

قابل ذکر است که تمامی مقادیر به ازای 100 گرم از خوراکی یا نوشیدنی می باشند (این واحد به دلیل وجود واحد های مختلف مانند، عدد، پیمانه، فنجان برش خورده، قاشق و انتخاب شده است).

خروجی scrapy به صورت پیش فرض json می باشد.

به دلیل مشخص بودن صفحات مربوط به خوراکی ها و نوشیدنی ها، در این پروژه صرفاً scrapping استفاده شده است. برای نمایش رکورد های مشابه از هاپیرامتری تحت عنوان alpha و beta استفاده شده است که به وسیله ی آن محدوده ی خطای نمایش تنظیم می شود تا برای بیننده موارد بسیار نزدیک و مشابه برای ارزیابی دقت مدل نشان داده شود.




شکل 11 : نشان دهنده ی تفاوت مقادیر کالری، کربوهیدرات، پروتئین، چربی، فیبر و میزان فعالیت های متداول بر حسب تغییر معیار مقدار مصرفی در سایت (در اینجا وای پیتزا از واحد 100 گرم و همچنین از یک تکه مثلثی استفاده شده است).

7. آینده ی پیش رو :

- تلاش است که بتوانیم مقدار داده های زیادی از خوراکی ها و نوشیدنی های جهان به دست آوریم تا نتایج نزدیک به واقعیت تری داشته باشیم.
- اطلاعاتی درمورد مقدار تاثیر پختن و گرمادیدگی خوراکی ها در مقدار کالری ها به دست آوریم تا با استفاده از مواد اولیه و مواد اصلی تشکیل دهنده ی غذاها و با اضافه کردن مواد دیگر ، مقدار کالری موجود را پیش بینی بکنیم.
- برچسب زدن داده های به دست آمده برای دسته بندی آن ها به دسته های مفید، مضر.
- استخراج دستور پخت غذاها و یادگیری آن با جزئیات (مانند میزان گرما، دمای مورد نظر، میزان مواد استفاده شده) و ترکیب با مواد اولیه و ایجاد غذاها و دستور غذاهای جدید (هزنوع غذایی باهزنوع طعمی که شامل خوش مزه یا بد مزه یا ... خواهد بود)

8. مخزن کد ها و اجرا و شبیه سازی :

کد ها در مخزن گیت هاب پروژه² موجود می باشد.

	saeidEmadi Merge branch 'main' of https://github.com/saeidEmadi/scrapeForMankan	c445f5a · 12 hours ago	🕒 16 Commits
📄 LICENSE	Create LICENSE	last week	
📄 MankanResults.ipynb	remove Kmeans	13 hours ago	
📄 Mankan_dataset.csv	clean data deleted duplicated rows	last week	
📄 README.md	Update README.md	last week	
📄 requirement.txt	requirements for simulate	12 hours ago	
📄 spider-man.py	he is spider-man LOL scrapy runspider codes	last week	

برای شبیه سازی لازم است که ابتدا **requirements.txt** را نصب کنید.

اجرای فایل **spider-man.py** با استفاده از **scrapy** امکان پذیر می باشد.

تمامی قطعه کد ها و نتایج به دست آمده در **MankanResults.ipynb** ذخیره شده است.

مجموعه داده های به دست آمده قبل پاکسازی برای پردازش، در فایل **Mankan_dataset.csv** ذخیره شده است.

siteid		name	calory	carbo	protein	fat	fiber	activity1	activity2	activity3	activity4
0	0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	8	جگر سیاه گوساله، خام	104.0	0.0	18.3	3.4	0.0	14.0	11.0	29.0	25.0
2	9	جگر سیاه گوسفند، خام	137.0	0.0	20.3	6.2	0.0	19.0	14.0	38.0	33.0
3	3	تخم مرغ آب پز	155.0	1.1	12.6	10.6	0.0	22.0	16.0	44.0	37.0
4	2	تخم مرغ سفیده خشک	295.0	0.0	73.7	0.0	0.0	41.0	31.0	83.0	71.0
5	1	تخم مرغ سفیده خام	52.0	0.7	11.0	0.2	0.0	7.0	5.0	14.0	12.0
6	10	جگر سیاه سرخ شده گوساله	176.0	0.0	22.3	9.6	0.0	25.0	18.0	50.0	42.0
7	4	تخم مرغ نیمرو با روغن	191.0	0.7	11.7	15.4	0.0	27.0	20.0	54.0	46.0
8	12	جگر سیاه (به آرامی پخته) گاو تر	198.0	3.0	24.8	9.5	0.0	28.0	21.0	56.0	48.0
9	13	مرغ کامل، خام	201.0	0.0	19.1	13.8	0.0	28.0	21.0	57.0	49.0

نمونه مجموعه داده ی

پیش پردازش نشده.

² <https://github.com/saeidEmadi/scrapeForMankan>