

Data tiyding

Soheil Shirvani

810195416



**در این پروژه می خواهیم با استفاده از سایت ورزش 3 که یک سایت معتبر برای اخبار ورزشی در ایران است داده‌‌ها‌‌ای را به واسطه صفحه های وب آن استخراج کنیم. برای این کار 5 فایل اکسل حاوی داده های صفحه وب به ما داده شده است تا داده ها را استخراج کنیم. این داده‌های استخراج شده سپس به تیم یادگیری ماشین داده می شود تا با پیدا کردن مدل های مناسب خروجی های مورد نظر را پیدا کنند.**

* **جام حذفی:**

**داده های این جدول مانند زیر است:**



**این داده ها شامل:**

**Year: سال مسابقه، Stage: مرحله مسابقه در جام حذفی، team1: تیم میزبان،**

**team2: تیم مهمان، score1: امتیاز تیم میزبان، score2: امتیاز تیم مهمان، time: زمان شروع مسابقه، date: تاریخ شروع مسابقه، penalty1: امتیاز گرفته شده در پنالتی برای تیم میزبان، penalty2: امتیاز گرفته شده در پنالتی برای تیم مهمان در صورتی که بازی به پنالتی نرود عدد آن 0 است.**

**این داده ها شامل 401 ردیف که نشان دهنده تعدادی از مسابقات جام حذفی در سالهای گذشته است.**

* **تیم ها:**

**این داده ها شامل:**



**که در آن**

**Competition: نشان دهنده مسابقه انجام شده، تیم: نشان دهنده اسم تیمی که در مسابقه شرکت کرده، بازیها: نشاندهنده ی تعداد بازی های انجام شده توسط آن تیم در آن مسابقه، امتیاز: نشان دهنده امتیاز گرفته شده توسط آن تیم در آن مسابقه است.**

**این داده ها شامل 241 ردیف که نشان دهنده 240 تیم است که در مسابقات مختلف شرکت کرده‌اند و امتیاز و تعداد بازیهای آن ها در همان مسابقه را نشان می دهد. تیم های خارجی و داخلی ایران هر دو در این جدول حضور دارند.**

* **جدول ها:**

**این داده ها شامل:**



**که نشان دهنده بازی های فوتبال و فوتسال است. این جدول شامل:**

**Year: سال انجام بازی، Competition: نام مسابقه ی بازی، تیم: اسم تیمی که در ان مسابقه شرکت کرده، بازیها: تعداد بازی ای که آن تیم در آن مسابقه انجام داده است، برد: تعداد برد های آن تیم در آن مسابقه، مساوی: تعداد مساوی های آن تیم در آن مسابقه، باخت: تعداد باخت های آن تیم در آن مسابقه، گل زده و گل خورده: تعداد گل های زده و خورده ی آن تیم در آن لیگ، امتیاز: امتیاز گرفته شده توسط آن تیم در آن لیگ.**

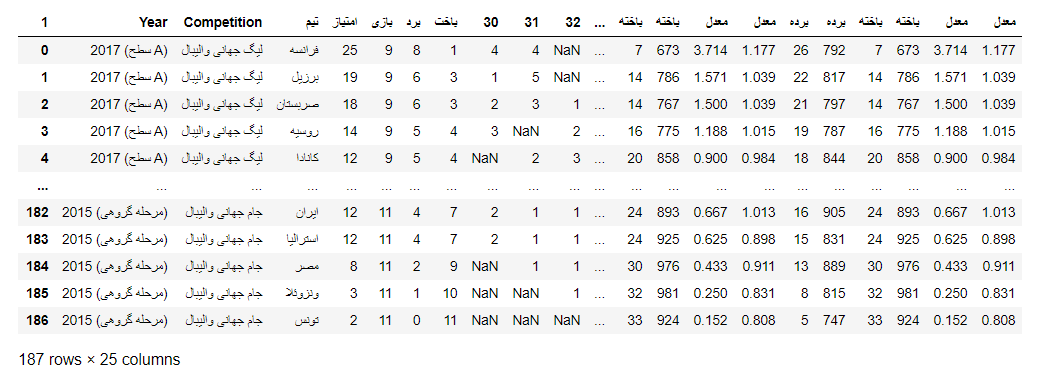
**این داده ها شامل 1960 تیم است که در لیگ های مختلف فوتبال و فوتسال شرکت کرده اند و امتیازات این تیم ها نوشته شده است.**



**که شامل بازی های تیم ها در لیگ های بسکتبال است. این جدول شامل:**

**Year: سالی که مسابقه برگزار شده است، Competition: نام لیگ ای که مسابقه برگزار شده است، تیم: نام تیمی که در آن مسابقه شرکت کرده است، بازی، برد و باخت: شامل تعداد کل بازی های آن تیم، تعداد برد ها و تعداد باخت های آن تیم که در آن مسابقه بوده است، گل خورده و گل زده: نشان دهنده تعداد گل زده و خورده ی آن تیم در آن لیگ، تفاضل گل: نشان دهنده تفاضل گل خورده و زده ی آن تیم در آن لیگ، امتیاز: نشان دهنده ی امتیاز گرفته شده توسط آن تیم در آن مسابقه.**

**این جدول شامل 94 تیم است که در لیگ های بسکتبال شرکت کرده اند و امتیاز های این تیم ها در این جدول جمع آوری شده است.**



* **نقل و انتقالات**

**جدول شامل زیر است:**



**که در آن:**

**نقل و انتقال: نام لیگ ای که نقل و انتقال انجام شده است، سال: سالی که این نقل و انتقال در آن لیگ انجام شده است، تیم: نام تیمی که این نقل و انتقال را انجام داده است، ورودی و قطعی: بازیکنان ورودی به تیم که قطعی شده اند، ورودی قرضی بازگشت: بازیکنان ای که به تیم های دیگر قرض داده شده بودند و بازگشتند، ورودی قرضی: بازیکنان ای که از تیم های دیگر قرض گرفته شده اند و به تیم ورود کردند، ورودی احتمالی: بازیکنان ای که احتمالا قرار است در این فصل به این تیم ورود پیدا کنند، خروجی ها: همان حالت های ورودی را دارد با این تفاوت که این بازیکنان از تیک خارج شده اند.**

**این جدول شامل 320 ردیف شامل نقل و انتقالات تیم های مختلف در سال ها و لیگ های مختلف را نشان می دهد.**

* **جام جهانی:**

**جدول آن به شکل زیر است:**



**این جدول شامل ستون هایی با همان تعریف های ستون های جام حذفی است با این تفاوت که جام آن یک جام جهانی در سال های 2014 و 2018 است. بقیه ستون ها همانند جام حذفی است.**

1. **برای پیشبینی قهرمان یک لیگ، از داده هایی که تیم قبلی به ما داده است چه اطلاعاتی بدست آوردیم؟ چه ویژگی هایی را استخراج کردیم و چه داده هایی را کنار گذاشتیم؟**

داده های استخراج شده از صفحات مختلف در بالا توضیح داده شد. لازم به ذکر است که داده هایی مانند فیلم و عکس های بازیکنان، بازی های گذشته برخی تیم ها که در جدول جام حذفی گزارش شده بود، خبر های ورزشی که در رابطه با تیم و لیگ ها در صفحات مختلف گذاشته شده بود استخراج نشده اند و در صورت نیاز می توان آنها را نیز استخراج کرد ولی ممون است تاثیر چندانی در پیشبینی فینال ندارند.

1. **هزینه استخراج ویژگی ها چقدر بوده است؟**

در قسمت قبل در رابطه با داده های استخراج شده و داده هایی که تصمیم گرفته شد استخراج نشوند صحبت شد. هزینه این استخراج برای داده های بالا هزینه زمانی بوده است و زمان خوبی برای استخراج گرفته اند. برای استخراج داده های بیشتر از جمله داده های خبر ها و یا فیلم ها نیاز به حافظه بیشتر و در نتیجه دیتابیس بزگر تر نیز داریم که مثلا برای داده های فیلم و دانلود کردن آنها ممکن است هزینه بیشتری نیاز داشته باشیم. برای استخراج جزیات بیشتر هزینه زمانی زیادی نیاز داشتیم چرا که برای اینکار باید به صفحان مربوط به هر کدام از صفحات داده شده نیز مراجعه می کردیم و جداول مرتبط را استخراج می کردیم. در صورت بزرگ تر شدن جداول نیاز داریم تا آنها را وارد دیتابیس کنیم که کوری نوشتن آن نیز یک هزینه زمانی دیگر تحمیل می کند.

1. **چه توجیهی برای این هزینه دارید؟ در آینده چه کمکی می تواند به ما بکند؟**

این داده ها علارقم داشتن هزینه زمانی می توانند اطلاعات زیادی برای ما داشته باشند. مثلا در صورت داشتن اخبار مربوط به یک تیم می توانیم نقل و انتقالات تیم، معروف بودن بازیکن های مختلف، میزان رضایت مردم از تیم و ... را تشخیص داد. این اطلاعات می تواند هزینه ناشی از استخراج اطلاعات را جبران کند. هزینه داشتن دیتابیس های خوب نیز با داده API های مناسب به شرکت ها و موسسه ها می تواند هزینه را جبران کند. این داده‌‌ها چون هزینه بیشتری می طلب اند نیاز داریم تا با تیم یادگیری مدل نیز ارتباط داشته باشیم تا در صورت نیاز آنها را استخراج کنیم.

1. **چگونه می سنجید که داده های استخراج شده قابل اعتماد است؟ فرایند کیفیت سنجی را توضیح دهید؟**

می توانیم داده ها را با سایت های معتبر دیگر چک کنیم که آیا نتایج بدست آماده با واقعیت بقیه سایت ها هم خوانی دارد یا خیر. می توانیم از یک متخصص در حوزه ی داده ها بخواهیم تا داده ها را چک کند و بگوید آیا این داده ها می توانند درست باشند یا خیر. می توانیم داده های پرت مربوط به هر جدول را استخراج نمودار های پراکندگی و هیستوگرام مربوط به آن ها را کشیده و از روی این داده های آماری ببینیم که آیا داده های بدست آماده قابلیت استفاده در واقعیت و صنعت را دارند یا خیر. نیاز داریم تا این داده ها را تمیز کنیم و با واقعیت چک کنیم سپس بعد از دادن داده ها به تیم یادگیری مدل مطمئن شویم به دقت بالایی رسیده اند در غیر این صورت داده های بیشتر و تمیز تری را استخراج کنیم.

1. **چگونه می توان این پروژه را در ابعاد بزرگ تر و در یک تیم انجام داد؟ چه وظایفی و چه فرایند کاری لازم است؟**

برای این کار نیاز داریم تا مثلا به هر کس تعدادی از قسمت های داده ها را بدهیم تا با دقت و جزییات بیشتری این کار انجام دهد. باید یک رابط یا مدیر بین دو بخش استخراج داده و یادگیری مدل باشد تا به گروه بگوید که چه داده های برای این گروه واجب است را استخراج کنند. نیاز به یک مدیر برای تامین هزینه ها و افراد داریم تا بتوانیم افراد مناسب را با کمترین هزینه پیدا کنیم. تسک ها باید طوری بین افراد تقسیم شود که همپوشانی نداشته باشند و بتوانند با توجه به توانایی های خود داده های خوب را استخراج کنند. تیم استخراج داده باید با تیم یادگیری مدل تعامل داشته باشد تا بتواند داده های مناسب را استخراج کند. در این راه می توانیم یک گروه نیز برای تایید داده ها داشته باشیم تا بتوانیم با صراحت بگوییم داده قابل اعتماد هستند. برای این کار می توانیم متخصص مربوط به کار را در گروه داشته باشیم و از سایت ها و صفحه های دیگر این داده ها را چک کنیم تا مغایرت با واقعیت نداشته باشند.

1. **نتایج حاصل از فعالیت تیم را چگونه به تیم بعد ارائه می دهید؟ توضیح دهید؟**

ابتدا جداول و دیتابیس های استخراج شده را با یک گزارش از هر کدام با توضیح ویژگی های استخراج شده توضیح می دهیم. این توضیحات به زبان ساده و عام هست و نیاز به تخصص ندارند و قابل فهم هستند. سپس با توضیح این که داده ها قابل اعتماد هستند و چه ویژگی های دیگری نیز قابل استخراج هستند که به دلیل هزینه آن فقط در صورت نیاز استخراج می شوند ادامه می دهیم. سپس در رابطه با تمیز کردن داده ها و این که چه پروسه ای برای این کار انجام دادیم صحبت می کنیم تا در صورت نیاز بتوانند با روش های دیگر این کار را انجام دهند. و در آخر نمودار هایی مربوط به ویژگی های آماری داده نشان می دهیم و داده ها و پراکندگی آن ها که مثلا چه تعداد داده ی پرت موجود است گزارش می کنیم.

جدول های گفته شده به پیوست آمده اند و در صورت قبول هزینه بیشتر می توان داده های بیشتری مخصوصا داده های متنی بهتری نیز از این صفحات بدست اورد.