احتمال گُل

فرض کنید که در راه رسیدن به مسابقات جهانی فوتبال، کاپیتان سوباسا تصمیم به استفاده از تحلیل داده برای ارتقای سطح فنی تیم خود گرفته است و به همین منظور **شما** به عنوان دانشمند داده تیم جذب شده‌اید. در هفته اول کاری خود، شما از تعدادی *کارآموز* درخواست می‌کنید که فیلم تمامی بازی‌های فصل قبل لیگ را مشاهده کنند و اطلاعات مرتبط با شوت‌ها در هر بازی را به صورت دستی ثبت کنند. خروجی کار در قالب csv ، تحویل شما شده است. شما می‌خواهید اقدام به ساختن مدل احتمال گُل (ارزیابی موقعیت شوت) با استفاده از دادگان آموزش (train.csv) بکنید. این مدل بایستی با دریافت مشخصات مربوط به شوت به عنوان ورودی، احتمال گُل شدن (عددی بین صفر و یک) آن را به عنوان خروجی برگرداند.

به عنوان مثال، شما فکر می‌کنید مُدل شما احتمال گُل شدن صحنه زیر را چند درصد اعلام می کند؟

با توجه به این که شما، به دنبال ساخت یک *مُدل مستقل از بازیکن و بازی* هستید، در دادگان آزمایش (test.csv) به ستون‌های playerId و matchId دسترسی ندارید. همچنین در نظر داشته باشید که در این مرحله، *گُل به خودی* به عنوان خروجی گُل برای شوت در نظر گرفته **می‌شود**.

دادگان

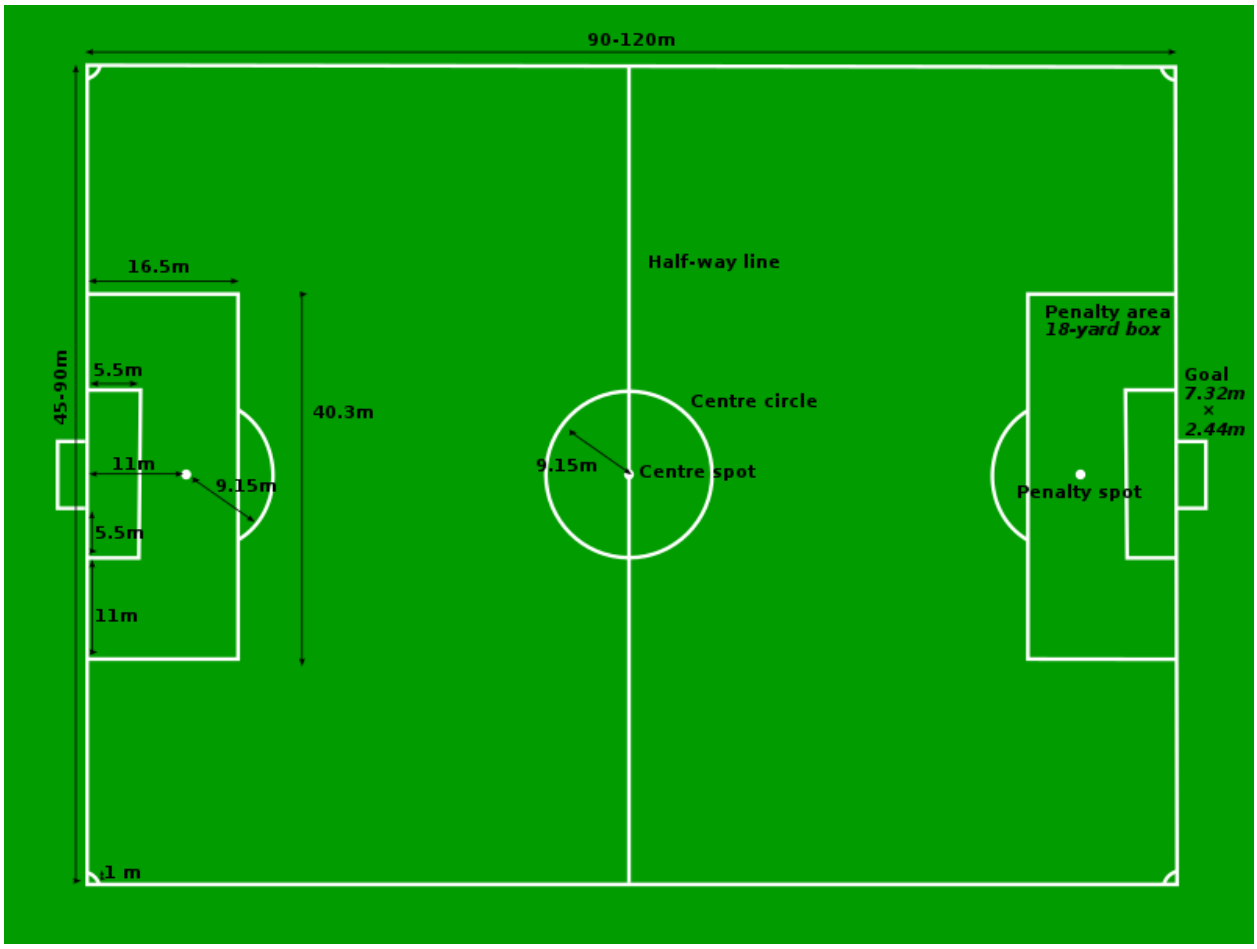
داده های مربوط به سوال را می توانید از این لینک دریافت کنید.

نام ستون	توضیح ستون

نام ستون	توضیح ستون
matchId	شناسه بازی
playerId	شناسه بازیکن شوت‌زننده
playType	موقعیت بازی که در آن ضربه زده شد (جریان بازی، پنالتی، ضربه آزاد مستقیم، مستقیم از کرنر)
bodyPart	بخشی از بدن که با آن شوت زده شده‌است (پای چپ، پای راست، سَر، سایر)
x	موقعیت شوت در زمین به متر (مختصات x)
y	موقعیت شوت در زمین به متر (مختصات y)
interveningOpponents	تعداد بازیکنان حریف که در لحظه شوت‌زدن مانع دید شوت‌زننده به دروازه شده بودند
interveningTeammates	تعداد هم‌تیمی‌هایی که در لحظه شوت زدن مانع دید شوت‌زننده به دروازه شده بودند
interferenceOnShooter	میزان دخالت مستقیم تیم مدافع بر روی شوت‌زننده (کم - هیچ بازیکن تیم حریف در یک متری شوت‌زننده نیست، متوسط - یک بازیکن حریف در امتری شوت‌زننده قرار دارد، زیاد - بیشتر از یک بازیکن حریف در امتری شوت‌زننده قرار دارند)
minute	دقیقه زدن شوت
second	ثانیه زدن شوت
outcome	نتیجه شوت (برخورد به دفاع، موقعیت از دست رفته، برخورد به تیردروازه، مهار توسط دروازه‌بان، گُل، گُل به‌خودی)

مبدا مختصات (۰,۰) مرکز دروازه‌ی تحت شوت می‌باشد و مختصات (x,y)، فاصله طولی (x) و عرضی (y) محل زدن شوت تا مبدا مختصات را به متر تعیین می‌کند، به عنوان مثال، موقعیت پنالتی در مختصات (۱۱,۰) می‌باشد.

برای آشنایی بیشتر با زمین فوتبال و ابعاد قسمت‌های مختلف آن، شما به عکس زیر از ویکی‌پدیای فارسی دسترسی دارید.



ارزیابی

برای ارزیابی مُدل شما از سطح زیر ناحیه نمودار ROC استفاده می‌شود. برای مطالعه بیشتر در مورد این نمودار می‌توانید ویکی‌پدیا یا راهنمای کوتاه نکات و ترفندهای یادگیری ماشین را مطالعه کنید.

نتیجه AUC ROC مُدل شما بر روی دادگان آزمایش در عدد ۱۰۰۰ ضرب شده و به عنوان امتیاز این مرحله در نظر گرفته می‌شود (بالاترین امتیاز ممکن از این مرحله ۱۰۰۰ می‌باشد).

داوری این سوال قبل از پایان مسابقه، تنها بر اساس ۳۰ درصد از دادگان آزمایش (test) خواهد بود. پس از اتمام مسابقه، برای به‌روزرسانی نهایی جدول امتیازات از ۱۰۰ درصد دادگان آزمایش استفاده خواهد شد؛

این کار برای جلوگیری از بیش‌برازش (overfit) روی دادگان آزمایش انجام می‌شود.

خروجی

پیش‌بینی‌های مدل خود بر روی دادگان آزمایش (test.csv) را در فایلی با نام output.csv قرار دهید. این فایل باید دارای یک ستون با نام prediction باشد که ردیف i ام آن پیش‌بینی شما (احتمال گُل‌شدن - عددی بین صفر و یک) برای شوت ردیف i ام از دادگان آزمایش باشد (دقت کنید که ستون باید حتما دارای header باشد). بعد از آماده‌سازی فایل output.csv ، آن را برای ما بارگذاری کنید.

▼ توجه

انتظار می‌رود افرادی که دارای توانایی آشنایی با حوزه جدید و مسلط به پیش‌پردازش، feature engineering و اصول اولیه یادگیری‌ماشین هستند، بتوانند این سوال را حل کنند.

▼ هشدار ارسال کد

فراموش نکنید کد این سوال را در تمرین آخر (بارگذاری کد)، بارگذاری کنید. در صورتی که پس از پایان زمان مسابقه، فایل کدها توسط شما بارگذاری نشده باشد، از جدول مسابقات حذف خواهید شد.