# دسته بندی

**تمرین درس داده کاوی بخش سوم - استاد مهردوست**

**کیان رضایی**

# دسته بندی – طبقه بندی (Classification)

- طبقه بندی (Classification) یکی از روش های یادگیری ماشین است که برای یادگیری چگونگی تخصیص برچسب کلاس به یک نمونه ورودی، استفاده میشود. در این بخش به الگوریتم های مختلف دسته بندی نظیر KNN، درخت تصمیم، رگرسیون خطی و لجستیک در دسته بندی و 'svm' پرداخته میشود. در انتهای این فایل پیاده سازی هر کدام از این موارد در پایتون انجام خواهد شد و خواهیم دید که هر کدام از این الگوریتم ها چه نقاط قوت و ضعفی دارند و موقعیت استفاده هر کدام چیست.

# مقدمه

- درک درست مفهوم طبقه بندی
- فهمیدن روش های مختلف مانند KNN ، Decision Trees، Logistic Regression، SVM
- پیاده داده سازی این الگوریتم ها در دیتاست های مختلف ( در این بخش و تمامی بخش ها از دیتاست مربوط به شرکت IBM استفاده میکنیم.)
- ارزیابی الگوریتم ها

# مقدمه

دسته بندی

- یک روش یادگیری نظارت شده است.

یادآوری میکنیم که منظور از یادگیری با نظارت (Supervised learning) یا یادگیری نظارت شده این است که به الگوریتم داده‌هایی را می‌دهیم که بعنوان پاسخ‌های صحیح در نظر گرفته می‌شوند.

- اشیا میباست حتما در یکی از کلاس ها قرار داده شود.
- دو مقداره یا چند کلاسه (Multi Class)

مثال: اینکه آیا کیان دانشجوی خوبی است یا خیر یک دسته بندی دو مقداره (Binary) است. اما آیا عکس داده شده شبیه به کدام یک از حروف الفبا میباشد مربوط به دسته بندی چند کلاسه میباشد.

# دسته بندی -- مثال

- فرض کنید که شما رئیس یک بانک هستید و میخواهید مشتریان خودتان را به سه دسته :

۱. مشتریان خوشحساب
۲. مشتریان معمولی
۳. مشتریان بدحساب

تقسیم کنید. وقتی مشتری وارد بانک شما شد احتمالا اگر در دسته اول قرار داشت خیلی او را تحویل بگیرید و نوبت سفارشی برای او قرار دهید و یا اگر با دسته سوم از مشتریان مواجه شدید از دادن وام مجدد به این دسته از مشتریان صرف نظر کنید.

# دسته بندی -- مثال

- یا در نظر بگیرید شما به عنوان مسئول حفظ مشتریان یک شرکت هستید و میخواهید ببینید که کدام مشتریان از خدمات گرفته شده راضی نیستند و احتمالا سال آینده دیگر به مشتری شرکت نخواهند بود. در صورت شناسایی (مشتری راضی / مشتری ناراضی) احتمالا شما میتوانید با دادن تخفیف مناسب به مشتریان ناراضی از دست دادن این مشتریان جلـوگیری کنید.

# دسته بندی -- مثال

- دادن وام ( سن ، درآمد، رکورد های قبلی مشتری و .... )
- اسپم یا مهم بودن یک ایمیل
- شناسایی دست خط
- شناسایی گفتار
- تشخیص اثر انگشت ( آیا این اثر انگشت کیان رضایی است یا نه!)

# دسته بندی – انواع روش ها

- K همسایه نزدیک
- درخت تصمیم
- بیز ساده (Naïve bayes)
- آنالیز افتراقی خطی (Linear discriminate analysis)
- رگرسیون لجستیک
- شبکه های عصبی
- ماشین بردار پشتیبان

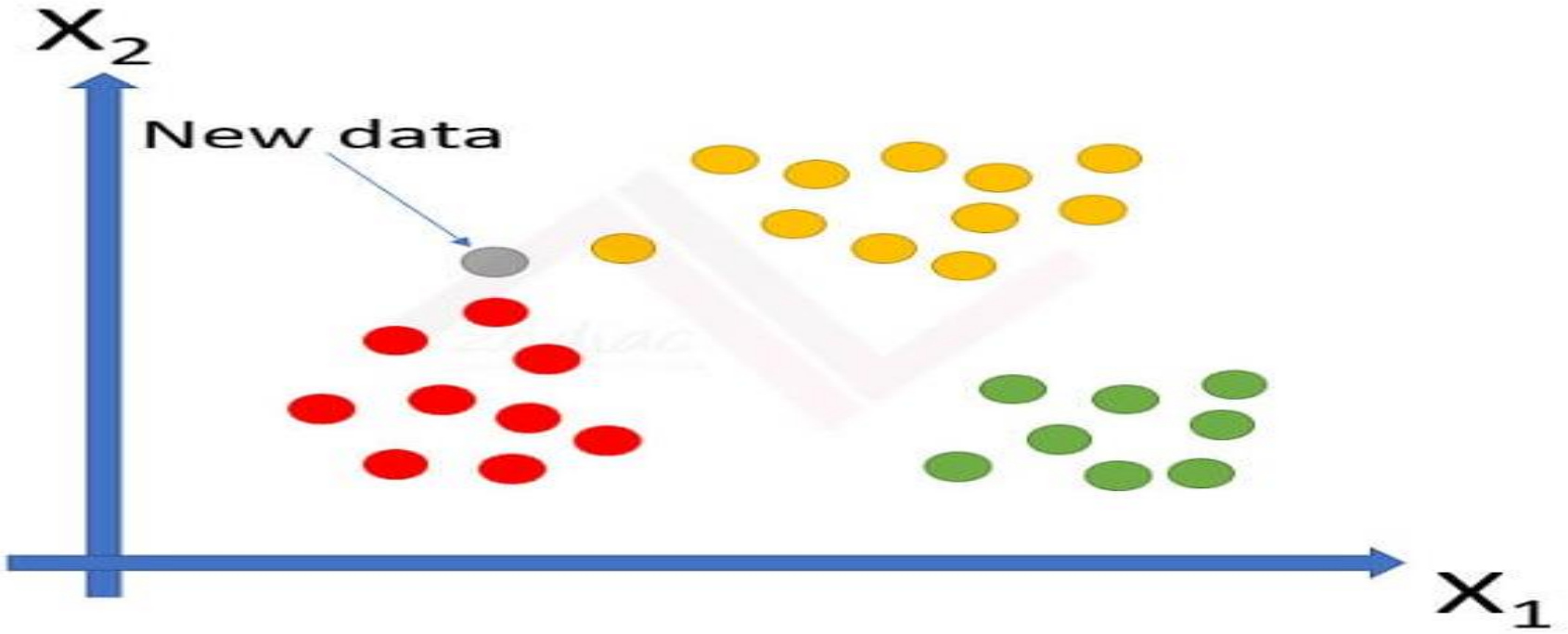
...

# KNN چیست؟

- یک متد آمار ناپارامتری است که برای طبقه‌بندی آماری و رگرسیون استفاده می شود. در هر دو حالت کی شامل نزدیک ترین مثال آموزشی در فضای داده ای می باشد و خروجی آن بسته به نوع مورد استفاده در طبقه بندی و رگرسیون متغیر است. در حالت طبقه بندی با توجه به مقدار مشخص شده برای کی، به محاسبه فاصله نقطه ای که میخواهیم برچسب آن را مشخص کنیم با نزدیک ترین نقاط میپردازد و با توجه به تعداد رای حداکثری این نقاط همسایه، در رابطه با برچسب نقطه مورد نظر تصمیم گیری می‌کنیم. برای محاسبه این فاصله میتوان از روش های مختلفی استفاده کرد که یکی از مطرح ترین این روش ها، فاصله اقلیدسی است. در حالت رگرسیون نیز میانگین مقادیر بدست آمده از k خروجی آن می باشد. از آنجا که محاسبات این الگوریتم بر اساس فاصله است نرمال‌سازی داده‌ها می‌تواند به بهبود عملکرد آن کمک کند.

# مثال

- به شکل زیر نگاه کنید.
- در شکل زیر ما ۳ دسته قرمز ، زرد و سبز را در اختیار داریم.
- این ۳ دسته ، دسته بندی شده اند کاملا از یکدیگر مجزا هستند.
- حال نقطه خاکستری که یک داده جدید است وارد می شود.

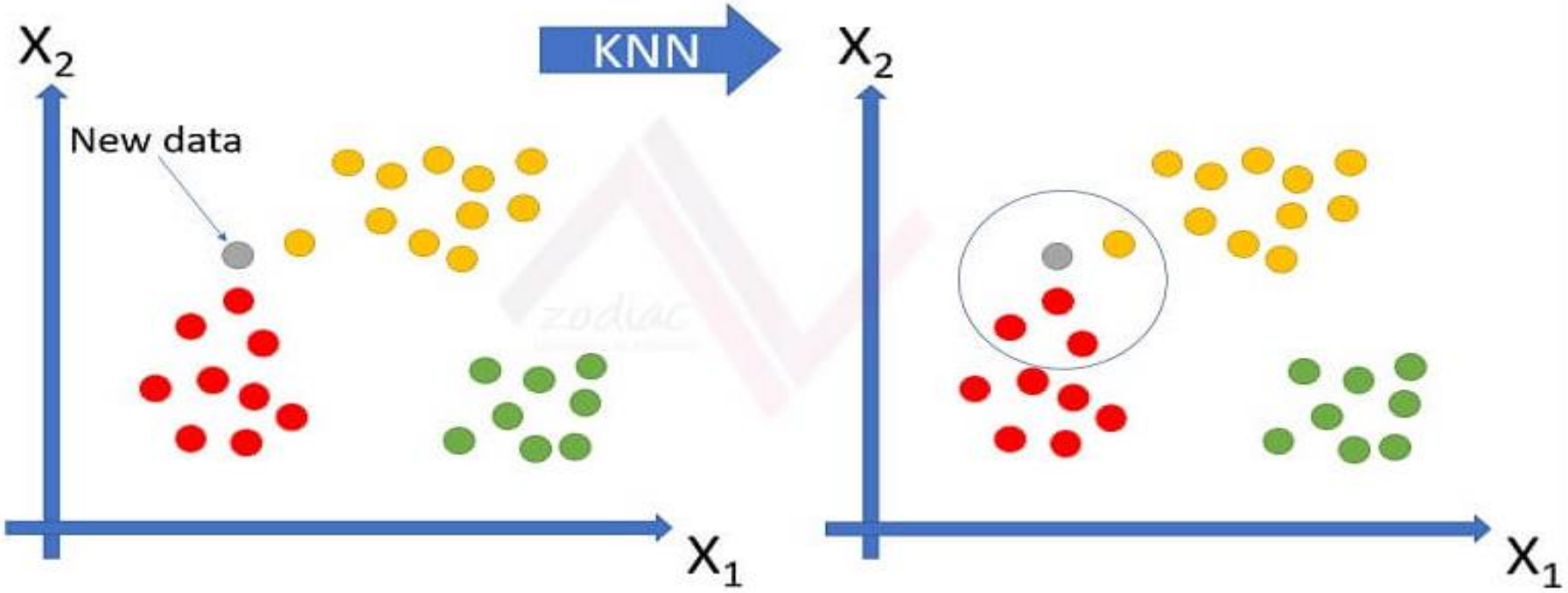


# ادامه مثال...

- سوال این است که این نقطه خاکستری مربوط به کدام دسته است ؟
- اولین چیزی که به ذهنمان خطور می کند این است که این نقطه نزدیک به کدام دسته است ؟
- یا بهتر است مانند مهندسین علوم داده بگوییم ویژگی های این نقطه به کدام دسته بندی نزدیک تر است.
- دقیقا الگوریتم KNN هم همینکار را می کند. الگوریتم KNN هم به همسایگان نقاط جدید نگاه می کند. مثلا در مثال بالا ما بر اساس ۴ همسایه قصد داریم تشخیش دهیم که نقطه جدید مربوط به کدام دسته است.

# ادامه مثال...

- از میان ۴ همسایه ، یک همسایه زرد و ۳ همسایه قرمز است.




پس میتوان پیشبینی کرد که خصوصیات نقطه جدید بسیار شبیه دسته قرمز است.

# تفاوت الگوریتم KNN طبقه بندی و رگرسیون

- همانگونه که پیش تر عنوان شد این الگوریتم برای طبقه بندی و رگرسیون استفاده می شود.
- در حالت طبقه بندی ما به سراغ فاصله نقاط می رویم.
- اما در حالت رگرسیون به سراغ میانگین مقادیر می رویم.
- در اصل در حالت طبقه بندی فاصله نقطه جدید با همسایگان انداز گیری می شود و سپس رای گیری می شود که همسایگان نزدیک در کدام دسته اند اما در حالت رگرسیون به سراغ میانگین نزدیکترین همسایه ها می رویم.
- چون الگوریتم KNN بر اساس فاصله است ، نرمال سازی داده ها عملکرد الگوریتم را بهبود می بخشد.

# معیار انتخاب همسایه

همانگونه که پیش تر عنوان شد در حالت طبقه بندی از معیار فاصله نقطه با همسایه و در حالت رگرسیون میانگین نزدیکترین همسایه ها مورد بررسی قرار می گیرد. برای فاصله نقطه و همسایه (طبقه بندی ) روش هاس متعددی مانند روش فاصله اقلیدسی ، فاصله منهتن ، معیار مینکوفسکی ، فاصله Mahala و فاصله Nobis استفاده میشود و اگر نوع داده گسسته باشد یعنی حالت رگرسیون و میانگین نزدیکترین همسایه ، معمولا از فاصله همینگ استفاده می شود.

مثال :

فاصله اقلیدسی :

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

فاصله منهتن :

$$d_1(\mathbf{p}, \mathbf{q}) = \|\mathbf{p} - \mathbf{q}\|_1 = \sum_{i=1}^{n} |p_i - q_i|,$$

# ارزیابی KNN

- ارزیابی عملکرد مدل ما را توصیف میکند.
- برای ارزیابی الگوریتم KNN مدل های ارزیابی متریک متفاوتی وجود دارند :

۱. اندیس جاکارد ( ساده ترین )
۲. محاسبه نمره F۱ (F۱-score)
۳. Log Loss

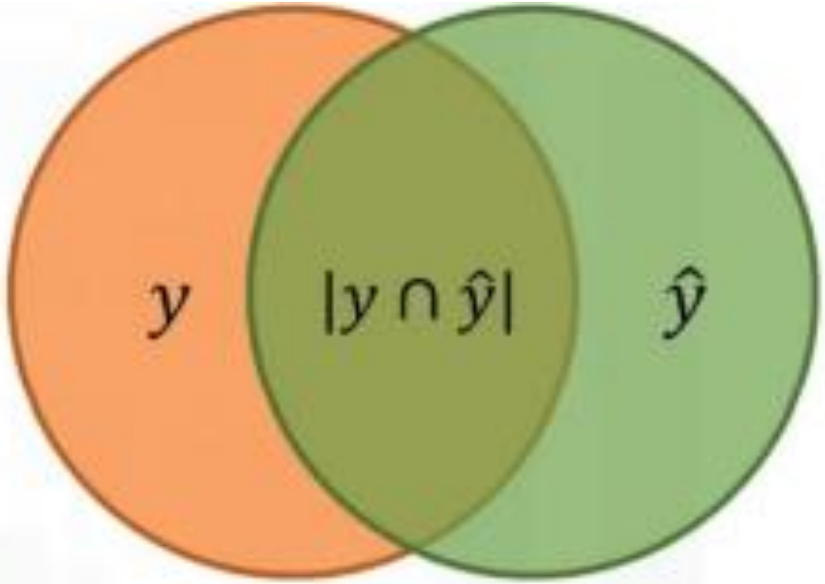
# ارزیابی به روش اندیس جاکارد

- این ارزیابی مبتنی بر نظریه مجموعه ها و مفاهیم مربوط به آن است.

$y$: Actual labels

$\hat{y}$: Predicted labels

$$J(y,\hat{y}) = \frac{|y \cap \hat{y}|}{|y \cup \hat{y}|} = \frac{|y \cap \hat{y}|}{|y| + |\hat{y}| - |y \cap \hat{y}|}$$




حدس نادرست

حدس درست

y: [0, 0, 0, 0, 0, 1, 1, 1, ,1 ,1 ]

ŷ: [1, 1, 0, 0, 0, 1, 1, 1, ,1 ,1 ]

$$J(y,\hat{y}) = \frac{8}{10+10-8} = 0.66$$

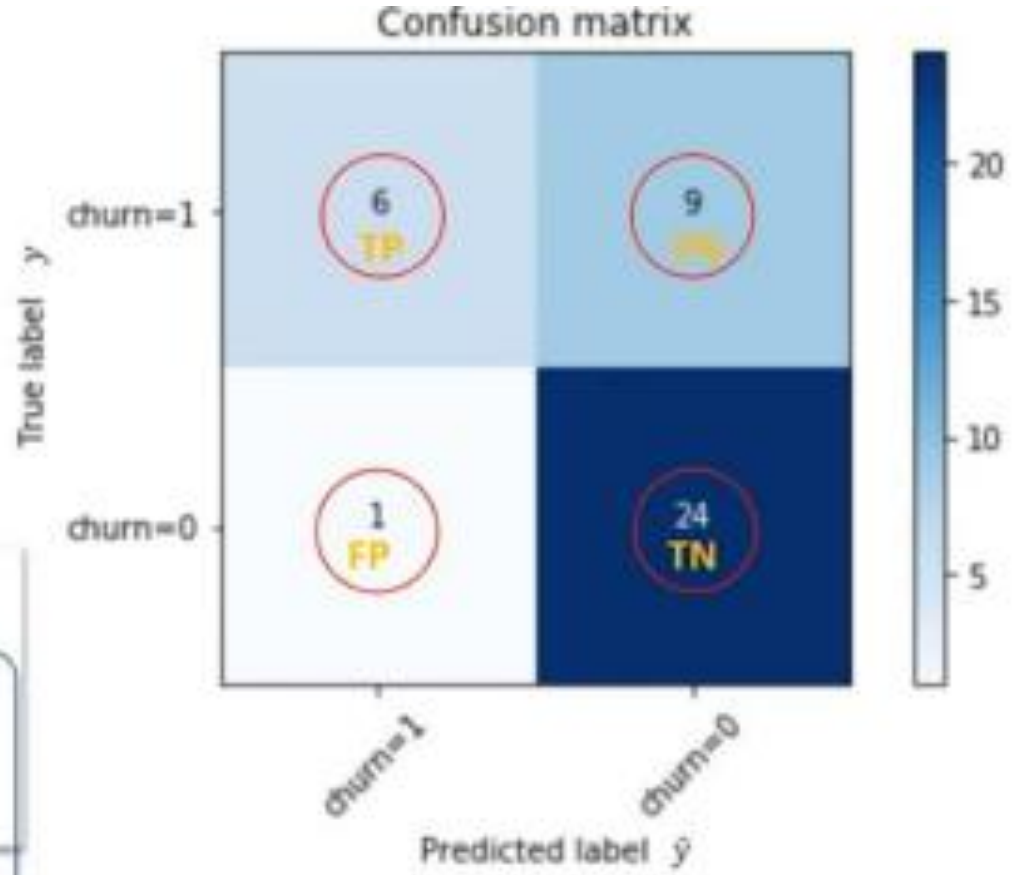
# ارزیابی به روش نمره F1

- Precision = TP / (TP + FP)
- Recall = TP / (TP + FN)
- F1-score = 2x (prc x rec) / (prc+rec)

**F1-score**: 0.00 ... 0.20 .... 0.55 .... 0.83 ... 1.00

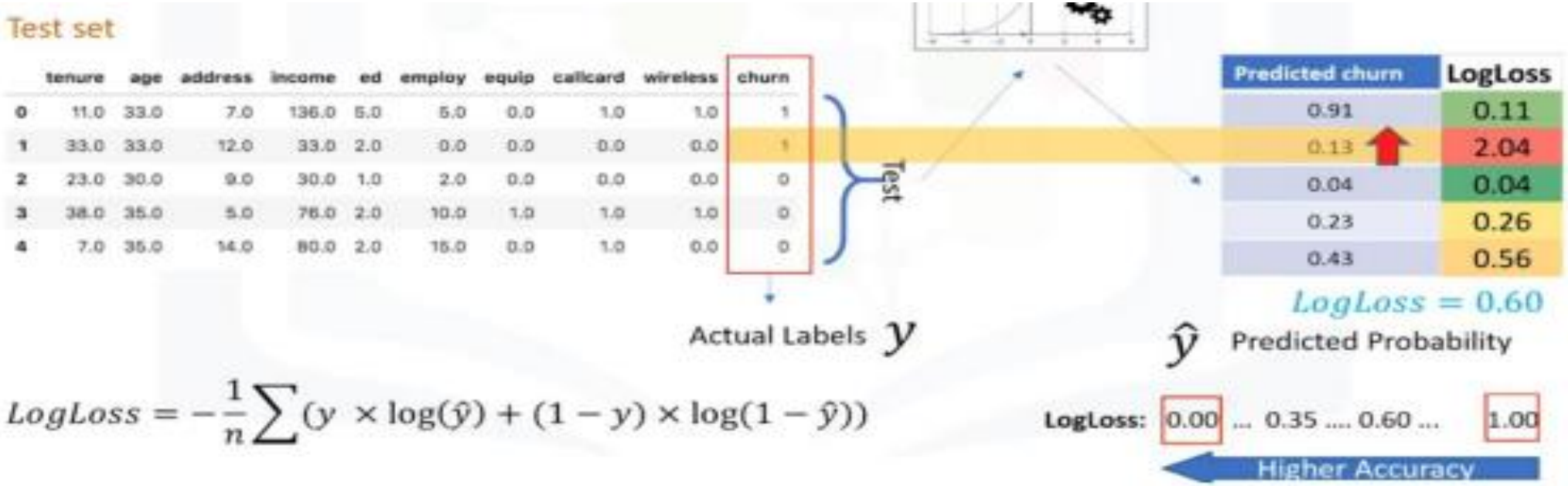
Higher Accuracy

| | precision | recall | f1-score |
|---|---|---|---|
| Churn = 0 | 0.73 | 0.96 | 0.83 |
| Churn = 1 | 0.86 | 0.40 | 0.55 |
| Avg Accuracy = | | | 0.72 |



# ارزیابی به روش Log loss

- عملکرد دسته ها بر اساس پیشبینی های خروجی یک مقدار احتمالاتی از صفر تا یک خواهد بود.



# مزایا الگوریتم KNN

- سادگی الگوریتم ( در جزوه درس این الگوریتم آورده شده برای همین از آوردن مجدد آن صرف نظر میکنیم.)
- تفسیر بسیار ساده
- دقت بالا
- چند منظوره بودن
- قابلیت استفاده در طیف وسیعی از مسائل

# معایب الگوریتم KNN

- زمان متوسط
- محاسبه کمی گران است.
- نیازمند حافظه زیاد چون باید تمامی داده های قبلی را ذخیره کند.
- حساس به مقیاس داده
- اگر K عدد بزرگی شود پیش بینی کند و زمان افزایش پیدا می کند.

- توجه : در کنار این فایل، یک فایل پایتون (KNN.py) قرار دارد که پیاده سازی الگوریتم KNN با دیتاست های جمع آوری شده از شرکت IBM است. باقی الگوریتم ها نیز در فایل های جداگانه پیاده سازی شده اند.

# درخت تصمیم

**مقدمه**

- تعریف درخت تصمیم
- کاربرد های درخت تصمیم
- انواع درخت تصمیم

.

.

# مقدمه

- در نظر داشته باشید چون در جزوه درس کاملا مفاهیم و الگـوریتم بـه دقـت توضـیح داده شده، از تکرار این مطالب صرف نظر میکنیم. در این بخش به توضیحات کلی و یک مثال از الگوریتم درخت تصمیم بسنده میکنیم.

# درخت تصمیم چیست؟

- درخت تصمیم (decision tree) یکی از پرکاربردترین الگوریتم‌ها در بین الگوریتم‌های داده‌کاوی است. درخت تصمیم دقیقا مانند یک درخت است با این تفاوت که از ریشه به سمت پایین (برگ) رشد کرده است. در الگوریتم درخت تصمیم نمونه‌ها را دسته‌بندی می‌کنیم که در واقع دسته‌ها در انتهای گره‌های برگ قرار دارد. درخت تصمیم در مسائلی کاربرد دارد که بتوان آنها را به صورتی مطرح نمود که پاسخ واحدی به صورت نام یک دسته یا کلاس ارائه دهند.

# درخت تصمیم چیست؟

- به طور خلاصه، درخت تصمیم نقشه‌ای از نتایج احتمالی یکسری از انتخاب‌ها یا گزینه‌های مرتبط بهم است به طوری که به یک فرد یا سازمان اجازه می‌دهد تا اقدامات محتمل را از لحاظ هزینه‌ها، احتمالات و مزایا بسنجد. از درخت تصمیم می‌توان یا برای پیشبرد اهداف و برنامه‌های شخصی و غیررسمی یا ترسیم الگوریتمی که بر اساس ریاضیات بهترین گزینه را پیش‌بینی می‌کند، استفاده کرد.
- یک درخت تصمیم‌گیری به طور معمول با یک نُود اولیه شروع می‌شود که پس از آن پیامدهای احتمالی به صورت شاخه‌هایی از آن منشعب شده و هر کدام از آن پیامدها به نودهای دیگری منجر شده که آن‌ها هم به نوبه خود شاخه‌هایی از احتمالات دیگر را ایجاد می‌کنند که این ساختار شاخه‌شاخه سرانجام به نموداری شبیه به یک درخت مبدل می‌شود.

# درخت تصمیم چیست؟

- درخت تصمیم یک ابزار پشتیبانی از تصمیم گیری است که همانگونه که از اسم آن مشخص است از نموداری درخت مانند استفاده می‌کند. این نمودار به صورت بیان تصمیم و ارزیابی پیامدهای احتمالی هر تصمیم ، از جمله نتایج رویدادهای احتمالی، هزینه‌های منابع و ابزارها ایجاد می‌شود. درخت تصمیم یک راه برای نمایش یک الگوریتم است که شامل عبارات کنترلی و شرطی است.

# کاربرد های درخت تصمیم

- **درخت تصمیم در مدیریت ریسک**

اصطلاح تجزیه و تحلیل تصمیم، در سال ۱۹۶۴ توسط رونالد ای هاوارد مطرح شد. یکی از کاربردهای درخت تصمیم گیری در مدیریت است. در بحث مدَیریت ریسک پروژه ، درخت تصمیم به تحلیل گر کمک می‌کند تا با در نظر گرفتن وقایع آینده، بهترین تصمیم گیری را در زمان حال داشته باشد. بنابراین، خروجی اصلی مورد انتظار ما در بحث مدیریت ریسک ، تحلیل و ارزیابی منطقی ریسک است یعنی تمامی گزینه‌هایی را که می‌توان انتخاب کرد و همچنین ارزش انتظاری تمامی خروجی‌های ممکن (مانند میزان سود یا برگشت سرمایه) را نمایش دهد.

# کاربرد های درخت تصمیم - ادامه

- **تعریف استاندارد PMBOK از درخت تصمیم**

PMBOK درخت تصمیم را به عنوان نموداری تعریف می‌کند که تعاملات کلیدی بین تصمیم ها و حوادث احتمالی مربوط به آن را به شکلی که تصمیم گیرندگان درک کنند به تصویر می‌کشد. درخت تصمیم گیری یک دنباله‌ای از تصمیمات و نتایج مورد انتظار از اتخاذ این تصمیمات را نشان می‌دهد. در جایی که احتمال رویدادها و مقادیر خروجی مشخص است، به‌عنوان مدل های کمی تصمیم گیری در فرآیند تصمیم گیری مورد استفاده قرار می گیرند.

# کاربرد های درخت تصمیم - ادامه

- **ارزیابی فرصت های رشد آینده نگر**
- **استفاده از داده های جمعیتی برای یافتن مشتریان احتمالی**
- **به عنوان یک ابزار پشتیبانی در زمینه های مختلف کاری**

**...**

# مزایا و معایب درخت تصمیم

- مزایا :

۱. برای فهمیدن بسیار آسان است به دلیل آنکه از یک فرایند مشابه که انسان برای تصمیم گیری در زندگی واقعی خود استفاده میکند، پیروی میکند.
۲. میتواند برای حل مسائل تصمیم محور بسیار کارا باشد.
۳. برای در نظر گرفتن به تمامی خروجی های احتمالی یک مسئله به ما کمک میکند.
۴. نیاز کمتری به پاکسازی داده در مقایسه با سایر الگوریتم ها دارد.

# مزایا و معایب درخت تصمیم

- معایب :

۱. درخت تصمیم شامل لایه های زیاد است، که این لایه های زیاد باعث پیچیدگی زیاد میشود.
۲. ممکن است ما را دچار معضل overfitting کند. البته برای حل این مشکل میتوانیم از الگوریتم جنگل تصادفی (Random Forest algorithm) استفاده کرد.
۳. برای کلاس های با برچسب زیاد، پیچیدگی زمانی الگوریتم به شدت بالا است.

# انواع درخت تصمیم

- ID3 ( Iterative Dichotomiser 3 )
- ID 4.5 ( Successor of ID3 )
- CART ( Classification and Regression Tree )
- CHAID ( Chi-squared Automatic Interaction Detector )
- …

# منابع


- Chapman, Hall . (2010).”Data Mining with R Learning with Case Studies” , Luis Torgo. LIACC-FEP , University of Porto. R. Campo Alegre, 823 – 4150.
- Han, Jiawei ., Kamber, Micheline. (2006). “Data Mining: Concepts and Techniques”, Morgan Kaufmann Publishers,
- https://corporatefinanceinstitute.com/resources/knowledge/other/decision-tree/
- https://www.datascienceprophet.com/understanding-the-mathematics-behind-the-decision-tree-algorithm-part-i/