مسابقه دومین کنگره بین المللی هوش مصنوعی در علوم پزشکی(AIMS2025)

عنوان: طراحی یک مدل غربالگری هوشمند برای تشخیص کبد چرب غیرالکلی

شرایط شرکت در مسابقه:

این چالش یک مسابقه آزاد است که امکان شرکت در آن تنها به صورت تیمی است. با توجه به ماهیت چالش، لزوم استفاده از تخصص های علوم پزشکی و مهندسی درکنار هم امری ضروری است لذا تیم های شرکت کننده حتما باید در کنار اعضای مهندسی، حداقل یک نفر از دانشجویان یا فارغ التحصیلان رشته های علوم پزشکی را به عنوان هم گروهی همراه داشته باشند.

جایزه ویژه:

به تیم اول مبلغ ۷۰۰ میلیون ریال به عنوان جایزه تعلق می گیرد که ۳۵۰ میلیون ریال مربوط به ارائه الگوریتم و ۳۵۰ میلیون ریال مابقی مربوط به چاپ مقاله Q1 در مجلاتscopus است.

پیش نیاز آماری مسئله:

الف- گمشدگی:

در داده‌های پزشکی، گمشدگی داده‌ها به سه دسته اصلی تقسیم می‌شود: **کاملاً تصادفی[1] نیمه تصادفی[2]**، و **غیر تصادفی[3]**.این دسته‌بندی‌ها نحوه و دلیل گمشدگی داده‌ها را توضیح می‌دهند و اهمیت ویژه‌ای در تحلیل داده‌های پزشکی دارند، زیرا نحوه برخورد با داده‌های گمشده می‌تواند روی نتایج تحلیل‌ها تأثیر بگذارد.

۱- گمشدگی کاملاً تصادفی MCAR

- در این حالت، داده‌ها بدون هیچ الگوی خاصی و به صورت کاملاً تصادفی گم شده‌اند. یعنی احتمال گم شدن یک داده هیچ ارتباطی با مقدار خود آن داده یا سایر متغیرهای موجود در مطالعه ندارد.
- به عبارت دیگر، اگر داده‌ای در یک متغیر خاص گم شده باشد، این گمشدگی هیچ وابستگی به مقادیر دیگر متغیرها یا حتی به مقدار خود آن متغیر ندارد.
- **مثال**: در یک مطالعه پزشکی، اگر پرسشنامه‌های برخی بیماران به دلیل خرابی تجهیزات به صورت تصادفی گم شوند و این گمشدگی هیچ ارتباطی با وضعیت سلامت بیماران نداشته باشد، این نوع گمشدگی کاملاً تصادفی است.
- **اهمیت**: این نوع گمشدگی کمترین مشکل را ایجاد می‌کند، زیرا می‌توان بدون تعصب از روش‌های جایگزینی (مانند میانگین‌گیری یا مدل‌های آماری) استفاده کرد.

---

[1] MCAR: Missing Completely At Random
[2] MAR: Missing At Random
[3] MNAR: Missing Not At Random

۲- گمشدگی نیمه تصادفی MAR

- در گمشدگی نیمه تصادفی، گم شدن داده‌ها به متغیرهایی غیر از خود متغیری که داده از آن گم شده است، وابسته است. یعنی اگر ما سایر متغیرها را در دست داشته باشیم، می‌توانیم دلیل گمشدگی را توضیح دهیم.
- به بیان ساده، احتمال گم شدن داده‌ها به سایر متغیرها بستگی دارد، ولی به مقدار خود متغیر گمشده ارتباط ندارد.
- **مثال**: فرض کنید داده‌های مربوط به سطح کلسترول برخی بیماران گم شده باشد، اما گمشدگی داده‌ها به دلیل سن بیمار باشد. اگر بیماران مسن‌تر داده‌های کلسترول خود را بیشتر از دست داده باشند، این گمشدگی نیمه تصادفی است.
- **اهمیت**: گمشدگی نیمه تصادفی می‌تواند با روش‌های پیشرفته‌تر آماری مدیریت شود، مثلاً از مدل‌های رگرسیون برای پیش‌بینی داده‌های گمشده استفاده شود.

۳- گمشدگی غیرتصادفی MNAR

- در این نوع گمشدگی، احتمال گم شدن داده‌ها به خود داده‌های گمشده وابسته است. یعنی دلیل گمشدگی داده‌ها به مقدار آن متغیر خاص مرتبط است.
- این بدترین نوع گمشدگی از نظر تحلیل داده‌ها است، زیرا گمشدگی با همان داده‌ای که می‌خواهیم تحلیل کنیم ارتباط دارد و تعصب ایجاد می‌کند.
- **مثال** :فرض کنید بیمارانی که سطح قند خون بسیار بالایی دارند از ارائه این اطلاعات خودداری کنند، در این صورت داده‌های مربوط به قند خون این بیماران گم می‌شود و دلیل گمشدگی به خود مقدار قند خون مرتبط است.
- **اهمیت** :مدیریت این نوع گمشدگی بسیار دشوار است و نیاز به روش‌های پیچیده و گاهی فرضیات خاص دارد، زیرا نمی‌توان از روش‌های ساده برای جایگزینی داده‌ها استفاده کرد.

ب- داده‌های نویزی[۴] شامل سه نوع اصلی هستند که هر کدام ویژگی‌ها و اثرات خاص خود را دارند. این سه نوع عبارتند از **نویز**[۵]، **خطاها**[۶] و **داده‌های پرت**[۷]. در ادامه به توضیح هر یک از این موارد پرداخته می‌شود:

۱- نویز

- نویز به دو حالت اصلی تقسیم می‌شود: **نویز تصادفی** و **نویز غیرتصادفی**.

نویز تصادفی[۸]

- **تعریف**: نویز تصادفی به نوسانات و خطاهایی اطلاق می‌شود که به صورت تصادفی و بدون الگوی مشخص در داده‌ها ایجاد می‌شوند. این نوع نویز معمولاً ناشی از عوامل محیطی یا تصادفی است که نمی‌توان به راحتی آن‌ها را کنترل کرد.
- **ویژگی‌ها**
  - نویز تصادفی به صورت پراکنده و غیرقابل پیش‌بینی در داده‌ها ظاهر می‌شود.
  - این نویز می‌تواند ناشی از تغییرات طبیعی در فرآیند اندازه‌گیری یا شرایط محیطی باشد.

---

[4] Noisy
[5] Noise
[6] Errors
[7] Outliers
[8] Random Noise

- معمولاً با تکنیک‌های آماری، مانند هموارسازی یا فیلتر کردن، قابل کاهش است.

- **مثال**: در اندازه‌گیری فشار خون، تغییرات جزئی به دلیل حرکات ناخواسته بیمار می‌تواند به عنوان نویز تصادفی شناسایی شود.

نویز غیرتصادفی[9]

- **تعریف**: نویز غیرتصادفی به انحرافات و خطاهایی گفته می‌شود که به طور سیستماتیک و با الگوهای مشخص در داده‌ها ایجاد می‌شوند. این نوع نویز معمولاً به دلیل مشکلات در فرآیند جمع‌آوری داده‌ها یا ابزارهای اندازه‌گیری ایجاد می‌شود.

- **ویژگی‌ها**:
  - این نویز به طور مداوم و در جهتی خاص در داده‌ها وجود دارد و می‌تواند منجر به سوگیری در نتایج شود.
  - شناسایی و اصلاح این نوع نویز معمولاً نیاز به بررسی دقیق فرآیندها و ابزارهای جمع‌آوری داده‌ها دارد.
- **مثال**: اگر یک دستگاه اندازه‌گیری دما به طور مداوم دمایی بالاتر از مقدار واقعی ثبت کند، این نشان‌دهنده وجود نویز غیرتصادفی است.

۲-خطاها

- **تعریف**: خطاها به اشتباهات ناشی از ورود یا پردازش داده‌ها اطلاق می‌شود که ممکن است به دلایل انسانی، فنی یا محاسباتی ایجاد شوند.
- **ویژگی‌ها**: این خطاها معمولاً به صورت مقادیر نادرست یا غیرمنطقی در داده‌ها ظاهر می‌شوند و می‌توانند منجر به نتایج نادرست شوند.
- **مثال**: ورود مقدار ۱۰- به جای ۱۷۰ برای یک قد می‌تواند تأثیر جدی بر تحلیل‌ها داشته باشد.

۳- داده‌های پرت

- **تعریف**: داده‌های پرت به مقادیری گفته می‌شود که به طور قابل ملاحظه‌ای از سایر مقادیر موجود در مجموعه داده‌ها فاصله دارند و معمولاً نشانه‌ای از خطا یا پدیده‌های خاص هستند.
- **ویژگی‌ها**: این داده‌ها می‌توانند به دلایل مختلفی وجود داشته باشند و می‌توانند تأثیر زیادی بر روی نتایج تحلیل‌ها داشته باشند.
- **مثال**: اگر در یک مجموعه داده مربوط به قد افراد، مقدار ۲۰۰ سانتی‌متر ثبت شود، این مقدار به عنوان یک داده پرت شناسایی می‌شود.

[9] Systematic Noise

ج- الگوریتم جعبه سفید و جعبه سیاه: الگوریتم‌ها در علم داده و یادگیری ماشین به دو دسته اصلی **جعبه سفید**[10] و **جعبه سیاه**[11] تقسیم می‌شوند. هر کدام از این دسته‌ها ویژگی‌ها، مزایا و معایب خاص خود را دارند. در ادامه به توضیح این دو نوع الگوریتم پرداخته می‌شود.

۱- الگوریتم‌های جعبه سفید

- **تعریف**: الگوریتم‌های جعبه سفید به الگوریتم‌هایی اطلاق می‌شود که شفافیت و قابلیت تفسیر بالایی دارند. در این نوع الگوریتم‌ها، فرآیند تصمیم‌گیری به راحتی قابل مشاهده و درک است، و می‌توان به راحتی نحوه رسیدن به نتایج را توضیح داد.
- **ویژگی‌ها**:
  - **قابلیت تفسیر**: این الگوریتم‌ها معمولاً به کاربران اجازه می‌دهند تا بفهمند که چگونه و چرا یک تصمیم خاص اتخاذ شده است.
  - **شفافیت**: مراحل و روند تصمیم‌گیری در این الگوریتم‌ها کاملاً مشخص و قابل پیگیری است.
  - **کنترل**: کاربران می‌توانند تغییرات را در ورودی‌ها مشاهده کرده و تأثیر آن‌ها را بر خروجی‌ها تحلیل کنند.
- **مثال‌ها**:
  - **درخت تصمیم**[12]: این الگوریتم با استفاده از ساختار درختی، فرآیند تصمیم‌گیری را به صورت سلسله‌مراتبی نمایش می‌دهد و به راحتی قابل تفسیر است.
  - **رگرسیون خطی**[13]: در این روش، رابطه بین متغیرها به صورت یک معادله ریاضی قابل مشاهده است.

۲- الگوریتم‌های جعبه سیاه

- **تعریف**: الگوریتم‌های جعبه سیاه به الگوریتم‌هایی گفته می‌شود که در آن‌ها فرآیند تصمیم‌گیری و نحوه رسیدن به نتایج به وضوح قابل مشاهده نیست. در این نوع الگوریتم‌ها، کاربران نمی‌توانند به راحتی بفهمند که چرا یک تصمیم خاص اتخاذ شده است.
- **ویژگی‌ها:**
  - **پیچیدگی**: این الگوریتم‌ها معمولاً پیچیده‌تر از الگوریتم‌های جعبه سفید هستند و فرآیند تصمیم‌گیری به صورت غیرخطی و غیرقابل پیش‌بینی است.
  - **عدم شفافیت:** کاربران نمی‌توانند به راحتی از مراحل داخلی و تصمیم‌گیری‌های الگوریتم مطلع شوند.
  - **دقت بالا:** در برخی موارد، این الگوریتم‌ها می‌توانند دقت بالاتری نسبت به الگوریتم‌های جعبه سفید داشته باشند، به ویژه در مسائل پیچیده.
- **مثال‌ها:**
  - **شبکه‌های عصبی**[14]: این الگوریتم‌ها با ساختارهای پیچیده‌ای از نورون‌ها کار می‌کنند و نتایج آن‌ها به راحتی قابل تفسیر نیستند.

---

[10] White Box
[11] Black Box
[12] Decision Tree
[13] Linear Regression
[14] Neural Networks

- **ماشین‌های بردار پشتیبان**[15]: این الگوریتم‌ها برای مسائل پیچیده طبقه‌بندی استفاده می‌شوند و معمولاً به عنوان یک جعبه سیاه در نظر گرفته می‌شوند.

د-متغیر وابسته و متغیر پیامد:

در علوم پزشکی، مفاهیم **متغیر پیش‌بینی**[16]، **متغیر وابسته**[17] **و متغیر پیامد**[18] از اهمیت بالایی برخوردارند. در ادامه به توضیح هر یک از این متغیرها پرداخته می‌شود

۱. متغیر پیش‌بینی

- **تعریف**: متغیر پیش‌بینی به ویژگی‌ها، عوامل یا متغیرهایی اطلاق می‌شود که برای پیش‌بینی یا توضیح یک نتیجه خاص مورد استفاده قرار می‌گیرند. این متغیرها معمولاً به عنوان ورودی در مدل‌های آماری یا یادگیری ماشین استفاده می‌شوند.
- **ویژگی‌ها**:
  - می‌توانند شامل عوامل جمعیتی، بیوشیمیایی، یا بالینی باشند.
  - معمولاً به بررسی رابطه بین خود و متغیر پیامد کمک می‌کنند.
- **مثال**: در یک مطالعه برای پیش‌بینی بروز دیابت، متغیرهای پیش‌بینی می‌توانند شامل سن، وزن، سابقه خانوادگی، و سطح فعالیت بدنی باشند.

۲- متغیر وابسته

- **تعریف**: متغیر وابسته به متغیری اشاره دارد که تحت تأثیر متغیرهای دیگر قرار می‌گیرد و نتیجه‌ای از تغییرات در این متغیرها است. در واقع، این متغیر معیاری برای سنجش اثرات متغیرهای پیش‌بینی است.
- **ویژگی‌ها:**
  - معمولاً با متغیر پیامد همپوشانی دارد، زیرا هر دو به تغییرات در متغیرهای پیش‌بینی وابسته‌اند.
  - در تجزیه و تحلیل داده‌ها، متغیر وابسته به عنوان نتیجه نهایی که مورد بررسی قرار می‌گیرد، مورد توجه است.

مثال: در مطالعه‌ای که به بررسی اثرات داروی خاصی بر فشار خون می‌پردازد، فشار خون می‌تواند به عنوان متغیر وابسته در نظر گرفته شود.

۳- متغیر پیامد

- **تعریف** :متغیر پیامد به نتیجه‌ای اشاره دارد که در مطالعه یا تحقیق بررسی می‌شود. این متغیر به طور مستقیم تحت تأثیر متغیرهای پیش‌بینی قرار می‌گیرد و نشان‌دهنده اثرات یا تغییرات ناشی از آن‌هاست.
- **ویژگی‌ها:**
  - معمولاً هدف اصلی مطالعه یا تحقیق است و برای ارزیابی اثربخشی درمان‌ها یا پیش‌بینی وضعیت سلامتی مورد بررسی قرار می‌گیرد.
- **مثال:** در همان مطالعه دیابت، متغیر پیامد ممکن است بروز یا عدم بروز دیابت در افراد باشد.

---

[15] Support Vector Machines
[16] Predictor Variable
[17] Dependent Variable
[18] Outcome Variable

پیش نیاز بالینی

روش تشخیص کبد چرب:

کبد یکی از حیاتی‌ترین اندام‌های بدن است که وظایفی همچون متابولیسم مواد مغذی، سم‌زدایی، تولید صفرا، و ذخیره مواد ضروری را بر عهده دارد. بیماری کبد چرب غیرالکلی (MASLD[19]) به تجمع چربی در کبد بدون مصرف زیاد الکل اشاره دارد و عمدتاً در افرادی با چاقی، دیابت نوع ۲، و سندرم متابولیک مشاهده می‌شود. شیوع MASLD در کشور بین ۳۰ تا ۴۰ درصد است (حدود ۲۵ تا ۳۵ میلیون نفر)، که این امر آن را به یکی از شایع‌ترین اختلالات کبدی تبدیل کرده است.

تشخیص بیماری‌های کبدی معمولاً به دو روش غیرتهاجمی و تهاجمی صورت می‌گیرد:

۱- روش‌های غیرتهاجمی:

- آزمایش آنزیم‌های کبدی: این آزمایش‌ها شامل اندازه‌گیری سطح آنزیم‌هایی مانند ALT ، AST و ALP در خون است. افزایش این آنزیم‌ها می‌تواند نشانه‌ای از آسیب کبدی باشد، اما به‌تنهایی برای تشخیص قطعی کافی نیست، زیرا افزایش این آنزیم‌ها ممکن است در سایر بیماری‌های کبدی نیز رخ دهد و محدود به کبد چرب نیست.
- سونوگرافی: سونوگرافی یک روش تصویربرداری غیرتهاجمی است که می‌تواند تجمع چربی در کبد را شناسایی کند و به ارزیابی شدت کبد چرب کمک کند. این روش رایج‌ترین ابزار تشخیصی برای کبد چرب است. این آزمایش حساسیت [60~90] و ویژگی [65~90] را دارا است.

۲- روش تهاجمی:

- نمونه‌برداری کبد (بیوپسی): در این روش، یک نمونه کوچک از بافت کبد برداشته شده و تحت میکروسکوپ بررسی می‌شود. بیوپسی دقیق‌ترین روش برای تشخیص و ارزیابی شدت بیماری‌های کبدی، از جمله کبد چرب و فیبروز، است. با این حال، به دلیل تهاجمی بودن، معمولاً تنها در موارد خاص و زمانی که تشخیص قطعی نیاز است، انجام می‌شود. این روش دارای حساسیت و ویژگی بالای ۹۰ درصد است.

درمان کبد چرب غیرالکلی (MASLD) برای تمامی بیماران ضروری نیست و بسیاری از افراد مبتلا در مراحل اولیه فقط به تغییرات سبک زندگی نیاز دارند. بنابراین، تشخیص بیمارانی که نیازمند درمان هستند در این مراحل اهمیت ویژه‌ای دارد. اما به دلیل تعداد زیاد بیماران، انجام آزمایش‌های آنزیم کبدی و سونوگرافی به‌طور گسترده برای میلیون‌ها نفر امکان‌پذیر نیست.

[19] Metabolic dysfunction-associated steatotic liver disease

موضوع چالش:

داده‌های گذشته نگر مربوط به بیماران مبتلا به کبد چرب جمع‌آوری شده است. این داده‌ها شامل اندازه‌گیری آنزیم‌های کبدی است. علاوه بر این، داده‌ها شامل چندین متغیر بالینی نیز هستند.

متغیر وابسته به دو کلاس تقسیم می‌شود:

کلاس No شامل افرادی است که به کبد چرب گرید ۱ و ۲ مبتلا هستند.

کلاس Yes شامل افرادی است که به کبد چرب گرید ۳ و ۴ مبتلا شده‌اند.

هدف این است که یک مدل غربالگری برای تشخیص احتمال ابتلای افراد به کبد چرب با گرید بالا است.

الگوریتم پیشنهادی باید از نوع جعبه سفید (قابل توضیح) باشد (انتخاب الگوریتم به دلخواه فرد است) و حداقل چهار قانون پیش‌بینی ارائه دهد:

دو قانون برای تشخیص کلاسNo (برای هر قانون میزان NPV و Suppoert گزارش شود)

دو قانون برای تشخیص کلاسYes (برای هر قانون میزان PPV و Suppoert گزارش شود)

همچنین، هر قانون باید حداکثر از چهار متغیر به‌صورت هم‌زمان استفاده کند.

متغیرهای مستقل کبد چرب:

- سن
- جنسیت
- خستگی مفرد
- درجه مختلفی زرد
- عدم هضم غذا
- تغییر رنگ شدید پوست (پوست کدر شده است)
- AST(SGOT)[20]
- ALT(SGPT)[21]
- ALP[22]
- Bilirubin
- کبد چرب

برنده مسابقه کسی است که بتواند قوانین بالینی صحیحی برای غربالگری کبد چرب غیرالکلی ارائه دهد و همچنین قوانینی را پیشنهاد کند که از نظر پوشش[23] و ارزش اخباری مثبت یا منفی[24] عملکرد بهتری داشته باشند.

[20] Aspartate Aminotransferase Serum Glutamic-Oxaloacetic Transaminase
[21] Alanine Aminotransferase Serum Glutamic-Pyruvic Transaminase
[22] Alkaline Phosphatase
[23] Support
[24] PPV or NPV

مجموعه داده های سوال :

https://B2n.ir/t77346

موارد ارسالی:

- انتخاب زبان برنامه نویسی آزاد می باشد.
- قوانین به همراه پارامترهای ارزیابی مشخص شده در یک فایل ورد یا اکسل گزارش گردد.
- می بایست کد پیشنهادی ارسال گردد.
- کلیه موارد ذکر شده را در فایل zip با نام match_team name_AIMS2025 قرار دهید و به ایمیل competition.aims@smums.ac.ir ارسال کنید. (team name نام تیم انتخابی شماست)

زمان بندی های چالش:

مهلت ارسال نتایج اولیه الگوریتم پیشنهادی: ۱۴۰۴/۰۲/۱۰

اعلام نتایج مرحله مقدماتی:۱۴۰۴/۰۲/۱۵

زمان برگزاری مرحله نهایی: ۱۴۰۴/۰۲/۲۴