

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

4 جی: Data Preprocessing

- A. ? درجے = داداں کے طبق Data Cleaning
- B. ? جو میرے میڈیا میں نہیں ہے Missing values
- C. ? وجد نہیں کرنے والے داداں کا اسٹنڈرڈ کرنا Outliers
- D. ? جو کاربود داری Data Transformation
- E. ? پر مقابلہ کرنے والے Encoding Techniques (One-Hot Encoding, Label Encoding)
- F. ? داداں کے طبق Model building, Feature Selection
- G. ? جو میرے میڈیا میں تکرار ہے Duplicate Data
- H. ? ایجاد کرنا Machine Learning program کے لئے Irrelevant Data
- I. ? بار بار کرنا Missing Values Data Imputation
- J. ? جو میرے میڈیا میں نہیں ہے Normality EOT

4: جنگی Data Preprocessing

جناح A: Data Cleaning (علم داده های نجات)

Data Cleaning یا کاری است که در علم داده های اولیه انجام می شود تا داده های خوب را از داده های ناقص و غیر مفید بفرمایی برخیارود.

۱. لغتی - داده های خام عوامل خطاها و مادردستی های فناوری های داده های خارجی را برداشته باشند. یا کاری داده های خارجی را برداشته باشند تا این داده های خارجی خود را بتوان لغتی برداشتنی کرد. این عمل هایی دقیق و معاصر ضروری اند.

۲. دقت عمل های سلسی: عمل هایی یادگیری و اسکن به داده های دستور و بدلون نفعی نیز طرز داده های آنلاین کرد و بکمود دقت عمل های اکثر فیلتری و بایانی می شوند تا بعدها بکمتری شامل شود.

۳. کاهش زمان هزینه: آنر داده های بزرگی یا کارهای ناشوند، عمل اندک در عمل هایی کلیل و عمل کاری مطالعه هایی سلسی برای آن بسیار زمان و هزینه بسیاری میگیرند و باید کاری اولیه فیلتر از این بروز این مطالعه هایی را بکاهش کری اند.

۴. تکمیل در تکمیل داده های آغاز و ختم: فرآیند تکمیل و تعمیی کاری داده های کاهشی از این تکمیل نهان و داشتن بی اهمیت داده های اولیه و زوردها، این را سازی کنند.

۵. **تصیم تیری**: متری: تصیمی کالا است. متری برداره، زمانی عویض تر هستند که دارند هارتفق و غیره باشند، پس از آن داده های برابر با عالم ها هستند. متری آنها را می بگیرند.

خطوکاری: مرحله ضروری ای - که تأثیر نداشته باشد بر لغت نتایج کمال های دارند ای در این مرحله ای داده های خود را می بخواهند.

چگونه دریافت فیلتردید missing values .B

مشیر داده: عکس از مقدار اعیانه (Missing Value) ای: عکس داده هایی که جایش های خود را که ای داشتند مختلفی برای دارند. این عکس ای داده وجود دارد که بعنوان داده و نزدیکی دارد.

۱. **زنگ داروها**: آن داده های مقادیر اعیانه در عکس داده هایی که باید داشتند ای دارند. داده های این داده هایی که ای داشتند ای داده هایی که ای داشتند ای دارند.

۲. **جاگایزی**: با مقادیر قابلیت، ویانه یافته: فیکر کن مقادیر اعیانه ای با فیکر این رسانیده باشند. سایر مقادیر جاگایزی کرد که این داده های دارند. این داده های دارند. این داده های دارند.

۳. **جاگایزی**: باسیں بینی: از داده هایی که این داده هایی دارند. فیکر کن برای سیسی مقادیر اعیانه ای استفاده کرد، این داده هایی که این داده هایی دارند.

۴. اسکالر دسته ها: برای مقادیر افسوسه فیتوال کیل درست جیسا اینجا نظر (فناهه "عمر افسوسه") که به عنوان اجازه های دادن اطمینان دارد، تغذیر نمود.

۵. استارتاپ روش های آنالیز (PCA): روشن هایی عاست کیل عوامل های اصلی (PCA) با این پیشنهاد طبقه بندی میگردند. این پیشنهاد MULTIPLE IMPUTATION میگردد که برای عیوب مقادیر افسوسه ۲۰ کاربرد دارد.

۶. کم و بیش به مقادیر افسوسه: در بین از کیل علاوه علوف این دست نیازی به دریافت مقادیر افسوسه نباشد و کیل براساس داده های موجود انجام شود.

اپنال روشن بسته به نوع داده ها صرف کیل و میزان داده های افسوسه درد.

C. Outliers

Outliers به نقاط داده ای اشاره می شود که به طور غایب تلقی نمی شوند ای دلایل نقاط داده های غایب دارند، این نقاط در تابعیت داده های غایب (متغیر) عاست خواهد داشت زیرا داده های غایب معمولاً داده های معمولی ندارند این داده های غایب اینها را از داده های غایب تبعیض می کنند. برای این تفاوت تائید کارهای outliers را بر کیل داده ها و نتایج نهایی داشته باشند.

۱. نمودار جعبه ای (Box Plot): این نمودار، می توان نقاط داده ای را بوضوح نشان دهد که فاصله این داده های بین عاده (IQR) قرار نداشند. بین این نقاط داده های غایب می شوند.

۲. نمودار پرالنگی (scatter plot): با استفاده از این نمودار، فیتوان مقاطعه ب طور غیر معمولی از الگوی مخصوص داروهای شناسایی کرد.

روش Z score: این روش برای محاسبه نکل داده Z score را می‌سازد اگر score را ب سمت از -3 تا کمتر از 3 باشد آن نکل به عنوان Outlier و غایر رفتاری می‌شود.

روش R: با محاسبه $R = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$ فیتوان حدود مابغایی داروهای شناسایی را خارج از این محدود قرار دارند به عنوان Outlier و غایر رفتاری می‌شوند.

۵. عملهای آماری و یادگیری ماژن: برای از الگوی سه ماتریسی پایه کردن ماژن، از این دسته تحلیلی تجزیه ماتریسی یا الگوی سه ماتریسی فوکوس بندی فیتواند به عنوان ایکسی Outlier را انتخاب کند.

استفاده از این روش با خود به شماره شناسایی و همایش Outliers کرد.

۶. Data Transformation: مرا کجا ببرد؟

Data Transformation با فرآیند تغییر فرم یا راستار، داروهایی به صفاتی بینوکته و گاییده استفاده آنها را در داده می‌کند.

۷. نسبت داده: با این دادهها فیتوان خواصی اقتصادی مادوناهمانی عاری شناسایی و اصلاح کرد.

2. پیامی سازی دادهها: در این قاعدهای کل دادهها از فرایند فتح مختلف جمع آوری شوند، تفسیر فعل دادهها میتواند بسیاری از پیام و ترکیب اصلی کارکرد را نشان دهد.

3. تخلیق و بحثینه لازمی: دادههای سیل نهاد فحوضه برای پژوهشگران
بحثینه لازمی فراخواسته خواهد بود.

4. دستیاری ارزیابی: با تخلیق دادهها به تخلیق قابل فهم تر و سادهتر و دیرباره
و تخلیق لغزان حیثیاتی تعمیمهای تکمیلی بخوبی بخوبی بخوبی.

5. تطبیق بایانی سیم های مختلف: دادهها علاوه بر است نیاز به تخلیق داشته
باشند تا با سیم های افزایشی افزایشی مختلف سازگار شوند.

در نتیجه Data Transformation،
تخلیق هایی که برای دادهها و تغییرات دادهها
متغیرهایی متألفه هیئت شوند.

(Label Encoding , One-Hot Encoding) Encoding Techniques.

نحوی داری؟

One-Hot Encoding و Label Encoding
دو تابعی که مسائل برای تبدیل طبقه های
گذشتگی به فرم عددی میکنند، اما هر دو از آنها به شیوه های مختلف عمل میکنند.

Label Encoding

مَعْرِفَةٌ : Label Encoding هر دسته (table) بِيُوْنِيلِ عِنْدَهُ فِي قَوْدِيْسِيلِ مَعْرِفَةٍ شُونَسِ بِهِ عِنْدَهُ مُتَّالِ آثَرَفِ وَيَزْلِي شَاطِلِ سَمَوَاتِ "قَرْفَزٌ" ، "بَزْ" وَ "آكْبَى" بَاشِ عَطَانِ اَسَتِ بِهِ عِنْدَهُ 1.0 وَ 0 لَدَلَانَارِي شَوَّدِ.

عِزَابَا : سَادِه وَ سَرِيعِ اَسَتِ وَ نَيَازِي بِهِ فَنَاقِي اَنَافِي نَفَارِدِ.

عِوَادِي : اَنَّ رَوْنِي عَطَانِ اَسَتِ بَاعِرَتِ اِيجَادِ رَابِطَهَايِ نَادِرِسِ بِهِ عِنْدَهُ دَسَهَا شَوَّدِ، بِهِ عِنْدَهُ مُتَّالِ "عَتَالِ" وَ "عَتَالِ بَالِ" عِدَدِ 1 عَطَانِ اَسَتِ بِهِ فَحَجَيِ بِهِ تَرَى لَانِدَتِ بِهِ مَا وَنَتَلِ دَهَرَكِ وَ وَاقِعِ جِنْلِ نَيَدِ.

One-Hot Encoding

مَعْرِفَةٌ : One-Hot Encoding هُوَ مَرْدَسَه بِيُوْنِيلِ فَلَوْرِ بَانِزِي كَبِيلِ فِي شُونَدِهِ دَهَنِ تَهْمَايِلِ عَنْصِرِ بِرَبِّرِ بَا 1 اَسَتِ وَ بَعْدِهِ مَخَاسِرِ بِرَبِّرِ بَا 0 هَسَرِ، بِهِ عِنْدَهُ مُتَّالِ "عَتَالِ" وَ "عَتَالِ" وَ "يَزْلِي" "قَرْفَزٌ" ، "بَزْ" وَ "آكْبَى" عَطَانِ اَسَتِ بِهِ مُورَدِتِ زَيْرِ لَدَلَانَارِي شَوَّدِ.

"قَرْفَزٌ : [0, 0, 1]

"بَزْ : [0, 1, 0]

"آكْبَى : [1, 0, 0]

عِزَابَا : اَنَّ رَوْنِي اِيجَادِ رَابِطَهَايِ نَادِرِسِ بِهِ عِنْدَهُ دَسَهَا جَلُوكِيِ حِيَانِ وَ بَهِ عَلِلِ لَكِ كَنِهِ تَابِعَهُ يَادِ بَلِيرِدِ.

مدل ایجاد کردن : معلمات ایست فضایی خصیرو سازی بیسٹری نیار داشتہ باشندہ دیگرہ اگر تعداد
دستور علاوه بر ایاد پائیں.

نتیجہ کری: استعمال بسی این روشنی بنتی دارہ دارہ مکالمہ عمل بنتی دارہ
برائی مکالمہ صافی کے بر روابط بسی دارہ دارہ مکالمہ معمولی ترجیح دارہ دی خود در
حالی کے عمل ایست بدل صافی سادھریا در شرایط خاص مناسب باشندہ

Model building ایسیت دارہ F جو Feature Selection

1. کاہنن بیسٹری عمل : با انتدا ب ویزکی های معمولی توان بیسٹری عمل
کاہنن ایست، این ب فضای سر عمل بیسٹری اکثریت و میں بینی ایست.

2. بہبود وقت عمل : حذف ویزکی های غیر ضروری و زائدی توان بہبود
وقت عمل لکھ کن، ویزکی های بی ربط عقلان ایست نویز ب عمل افراحت
کردو، عالمد آن کاہنن دھنے.

3. کاہنن خطہ Overfitting : با استفادہ از تعداد لغتی از ویزکی ها، احتمال
اینکہ عمل بے دارہ های اکثریتی میں از جد تطبیق یا (Fit) کاہنن ہی یا یا۔

4. افرائیں قابلیت تفسیر : عمل مداری کے ویزکی های لغتی داری معمولہ
قابل تفسیر ہنسن این افراد ریاضی فانس پڑائی یا عالی کے نیاز بے تحفہ
دقیق دارہ بسیار عجم ایست.

5. کامن حی ساختی: کار با ویژگی هایی است که بازی طبقه بندی می شود، این امر به ویژه در داده های بزرگ و پیوسته اهمیت دارد.

به طور کلی، فرآیند Feature Selection بین چنین ترتیب و اینها محال و قابل اعمال نمی باشد.

5. حذف داده های تکراری (Duplicate Data):

حذف داده های تکراری در پایگاه های داده عموماً با استفاده از (زون) هایی زیر آید.

می شود:

1. استفاده از دستور DELETE: میتوان از دستور DELETE به عنوان یک زیر پرسشی (Subquery) برای انتخاب و حذف کردن داده های تکراری استفاده کرد، به عنوان مثال:

`DELETE FROM table_name`

`WHERE id NOT IN (`

`SELECT MIN(id)`

`FROM table_name`

`GROUP BY column1, column2`

`);`

2. استفاده از تابع MAX: برای از پایگاه های داده ای استفاده کننده و آن تابع تحریک دارند، میتواند برای انتخاب و حذف داده های تکراری مورد استفاده قرار گیرد. به عنوان مثال

۱۶- ساختار از نابع (ROM NUMBER) هست که در توان دارد همچنانکه کد و نسخه
کوادھای تکراری را حذف کرد.

۱۷- ساختار از اینکه قدرتی بایگانی دارد: بایگانی عدیدت پایگاه

داده (DBMS) مانند SQL Server management studio MySQL workbench
حذف این از این نزینه سایر ترافلی برای عنایتی و حذف داده های تکراری از آن هی دهن.

۱۸- ایجاد قیدیتی یکتا (unique constraints): برای جلوگیری از ورود داده های

تکراری دوچاره اول رفعی توان قیدیتی یکتا (ایروی) ساختن های خاص از جدول اعمال
کرد. این کار بخط و خود کار از ورود داده های تکراری جلوگیری فرمی کرد.

۱۹- برسی (سی): در بینی صادرات معلم این است لازم باشد که داده ها به صورت دستی

بررسی و تکمیلی هایی کنند که اساسی و حذف شوند، به قدری که داده های سیمده با باساختار
های خاص.

۲۰- ساختار این روش ها، فهرزان داده های تکراری را در بایگانی داده هایی و
حذف کرد.

۲۱- مفهومی که معملاً ترتیب ادریسی سی های Machine Learning

ایجاد فیلتر

۱. کاهش دقت فعل : داده های نا مربوط فی تواند به فعل اصلی نادرست ارزی دهن، که عینکه کاهش دقت سشن بسیارها فی شود.

۲. افزایش پرسیکی فعل : افزون داده های نا مربوط فی تواند بجزئی فعل افزایش دهد و باعث شدنکه فعل به قویی generalize شاند.

۳. نمان و منابع پشتیر: پژوازش داده های نا مربوط را عال و منابع پشتیری را می طلب و فعل ایس کاری بسیم را باعث تأثیر آفرادند.

۴. اختلال fitting : فعل معلم ایس به داده های نا مربوط سشن ایحد حساس شود و بجهای یادگیری الگوهای واقعی، به یادگیری نوین پژوازد.

۵. تفسیر نادرست نتایج : وجود داده های نا مربوط فی تواند عینکه به تفسیر نادرست نتایج و آنکه تفسیر نادرست شود.

۶. تداخل با دیگری هایی صفهم : داده های نا مربوط فعل ایس- یا ویژگی هایی صفهم ترافل لتو و باعث شوندنکه فعل نتواند بررسی آن هارا نمایی کند.

در تسمیه، تغییر کردن داده ها و اطلاعات از عربی به دون آنها یکی از عوامل کلیدی در فرآیند یادگیری محسنه ایس است.

Missing Values براي پر كردن Data Imputation ۱۲.I

ایجاد کسری و از پنهان آمدن داده ها در مجموعه داده برای آنها در این قسمت داده ها را با عنوان Data Imputation

۱. بیکار دلخیز - داده ها: مقادیر اعینی هی تواند باشد که هنوز درست عمل نمایند سپس بینی شوند، با این کار ان عقاید لطفی - داده ها بگردند هی یابد و عمل ها بگیر عمل می شوند.

۶. اجتناب از حذف داده ها: دلایل وجود مقادیر اعشاری، کلی از ترین عوامل حذف کل داده های ویژگی هاست. این کار می تواند عصب بخوبی را از داده های اصلی برداشته باشد. از روش IMPUTATION بحالی امکان رفع دهنده کردن داده های صورت دارد. بحث در نحو استفاده از آن.

۳- حقیقی توزیع داده‌ها: با استفاده از اوشی های فناوری Impputation توزیع داده‌ها، حقیقی شم و از بروز سوگیری های ناشی از حذف داده‌ها جلوگیری کنیم.

۱۰. افزایش قابلیت استفاده از اکو-تئم های پارهیزی ممکن است در برخی از این اکو-تئم ها این انتقاد را کنم و کمیل های بصری آنچه در همین

۵. کاهش عدم قطعیت : فرآیندی اینده می‌تواند به عنوان قطعیت در یک داده های ساخته شده باشد. غیر شوند IMPUTATION می‌تواند به کاهش این نوع عدم قطعیت امکان دهد.

و، کل، این ابزاری حیاتی برای کلیه و پردازش داده های را در عین حال، که غایب می‌گردند را در داده های ناقص بخوبی استفاده را بسیار خوبی دارد.

۶. جلوگیری توانی Normality را در داده های کوچک بررسی کنید

برای بررسی نرمالیتی در داده های کوچک، می‌توان روش های آنچه می‌گفتیم استفاده کنید.

۷. نمایش تصویری (VISUAL INSPECTION) :

هستوگرام : با رسم هستوگرام داده های می توانید توزیع آنها را مشاهده کنید. اگر داده های نرمال باشد باید بغل زنگولایی داشته باشند.

نحوه Q-Q : با رسم نمودار Quantile-Quantile plot Q-Q می توانید مقادیر

اعیانی داده های خود را با مقادیر اعیانی توزیع نرمال مقایسه کنید. آن نقاط بخوبی خط 45 درجه قدرتگیرند. داده های نرمال هستند.

۲. آنچه عنوان می‌شود اکواری:

آنچه عنوان شایسته و معتبر است shapiro-wilk Test (آنچه آنچه عنوان برای بررسی نرمالیتی داده ها از توزیع نرمال پیروی فیلتر نمایند) خوب است.

آنچه عنوان کوتوگوف- اسمیرنوف (kolmogorov-smirnov Test): آنچه آنچه عنوان نیز صیغه ای برای بررسی نرمالیتی استفاده شود، هرچند که حسابات کمتری نیست - بعنوان های کوچک دردست است.

۳. معیارهای توزیع:

معیارهای آمارهای عالیه که در آنکه شکستگی (skewness) و کورتیسیس (kurtosis) را می‌نمایند، برای توزیع نرمال تعیین می‌نمایند. معیارهای kurtosis تردیدی به ۳ باشد.

۴. تحلیل های دیگر:

تحلیل های دیگر: معیارهای روش های سبیل داده واند سبیل لگاریتمی یا سبیل جذر برای نرمال کردن داده ها استفاده کنند.

ترکیب:

آنچه روش های تواند به شاگرد آنستایاب درک بخوبی از نرمالیتی داده های خود بررسی کند.