دانشگاه ملی مهارت

آموزشکده میناب

نام و نام خانوادگی : آرش زارعیان

واحد درسی : مباحث ویژه

رشته : مهندسی حرفه ای کامپیوتر

مدرس : محمد احمد زاده

بخش 4 Data Preprocessing :

A. چرا Data Cleaning در علم داده اهمیت دارد؟
B. Missing Valuesچگونه مدیریت می‌شوند؟
C. Outliersچیست و چگونه می‌توانید آن‌ها را تشخیص دهید؟
D. Data Transformationچرا کاربرد دارد؟
E. Encoding Techniques (One-Hot Encoding و Label Encoding) چه تفاوتی دارند؟
F. چرا Feature Selection در Model-building اهمیت دارد؟
G. Duplicate Dataچگونه در پایگاه داده‌ها حذف می‌شود؟
H. Irrelevant Dataچه مشکلاتی را در پیش‌بینی‌های Machine Learning ایجاد می‌کند؟
I. چرا Data Imputation برای پر کردن Missing Values کاربرد دارد؟
J. چگونه می‌توانید Normality را در داده‌های عددی بررسی کنید؟

مدرس: محمد احمدزاده ترم بهمن ۱۴۰۳ مباحث ویژه

Day.... Month.... Year..... Subject.....

بخش 4 Data Preprocessing

A: چرا Data cleaning در علم داده اهمیت دارد؟

Data cleaning یکی از مهم ترین مراحل در پردازش و تحلیل داده ها است. دلایل اهمیت آن عبارت اند از:

۱. افزایش دقت مدل ها ۲. کاهش خطا و تناقضات ۳. بهبود عملکرد الگوریتم ها

۴. تصمیم گیری بهتر ۵. صرفه جویی در زمان و هزینه.

به طور کلی، کیفیت داده ها تاثیر مستقیمی بر نتایج تحلیل و مدل های یادگیری ماشین دارد.

بدون پاکسازی مناسب، حتی قوی ترین الگوریتم ها نیز نمی توانند عملکرد مطلوبی داشته باشند.

B. Missing Values چگونه مدیریت می شود؟

مدیریت مقادیر گم شده در داده ها بخش مهمی از پیش پردازش داده ها است. روش های

مختلفی برای برخورد با این مقادیر وجود دارد که بسته به نوع داده و کاربرد،

انتخاب می شوند.

AVANGE

Day.... Month.... Year..... Subject.....

روش های مدیریت مقادیر گم شده

۱- حذف داده های دارای مقدار گم شده مزیت: ساده و سریع عیب: ممکن است

باعث از دست رفتن اطلاعات مهم شود.

۲- جایگزینی مزیت: داده ها حفظ می شوند عیب: ممکن است دقت مدل کاهش یابد

۳- استفاده از مدل های پیشبینی مزیت: جایگزینی دقیق تر عیب: زمان پردازش بالا

۴- استفاده از مقدار قبلی یا بعدی مزیت: مناسب برای داده های ترتیبی

عیب: در صورت وجود تغییرات ناگهانی، ممکن است داده ها نادرست شوند

۵- استفاده از الگوریتم های مقاوم به مقادیر گم شده مزیت: کاهش نیاز به پیش پردازش

عیب: همیشه بهینه نیست

۶. outliers چیست؟ چگونه می توان آنها را تشخیص داد؟

outliers به نقاط داده ای گفته می شود که به طور قابل توجهی با سایر داده های موجود در یک

مجموعه داده تفاوت دارند. این نقاط می توانند به دلایل مختلفی مانند خطاهای اندازه گیری،

Day.... Month.... Year..... Subject.....

ورود اشتباه داده‌ها یا پدیده‌های واقعی در داده‌ها به وجود بیایند. راه‌های مختلفی برای

شناسایی داده‌های پرت وجود دارد از جمله:

۱- روش آماری (محدوده بین چارکی - IQR) ۲- استفاده از انحراف معیار

۳- روش‌های مبتنی بر مدل‌های یادگیری ماشین ۴- استفاده از نمودارها

D- چرا Data transformation کاربرد دارد؟

Data Transformation فرآیند تبدیل، تغییر یا نگاشت داده‌ها به یک فرمت جدید

است که باعث بهبود تحلیل، مدل‌سازی و تفسیر داده‌ها می‌شود. این تبدیل می‌تواند شامل

استاندارد سازی، نرمال سازی، تبدیل لگاریتمی، دسته‌بندی، رمزگذاری و کاهش ابعاد باشد.

کاربردهای مهم: ۱- بهبود دقت مدل‌های یادگیری ماشین ۲- مدیریت داده‌های پرت

۳- افزایش کارایی الگوریتم‌های مبتنی بر فاصله ۴- بهبود توزیع داده‌ها ۵- کاهش ابعاد و

پیچیدگی داده‌ها ۶- سازگاری داده‌ها برای مدل‌های مختلف

E- one-Hot Encoding و Lable Encoding چه تفاوتی با هم دارند؟

AVANGE

Day.... Month.... Year..... Subject.....

هر دو روش برای تبدیل متغیرهای دسته‌بندی شده به داده‌های عددی استفاده می‌شوند اما

نحوه کاربرد و تبدیل آنها متفاوت است.

Label Encoding (رمزگذاری برچسبی)

در این روش هر دسته (کلاس) به یک عدد صحیح اختصاص داده می‌شود.

مزایا: سادگی و کاهش ابعاد داده‌ها - مناسب برای مدل‌هایی که روابط ترتیبی را در نظر می‌گیرند

معایب: ایجاد ارتباط‌های عددی نادرست بین دسته‌ها

One-Hot Encoding (رمزگذاری تک حالتی)

در این روش برای هر مقدار از متغیر دسته‌بندی شده، یک ویژگی جدید ایجاد شده و مقدار ۱ یا ۰

به آن اختصاص می‌یابد

مزایا: از بین بردن ارتباط عددی نادرست بین دسته‌ها - مناسب برای مدل‌هایی که فرض می‌کنند

داده‌های ورودی مستقل از یکدیگرند

معایب: افزایش ابعاد داده، به خصوص در متغیرهای دسته‌بندی شده با مقادیر زیاد

Day.... Month.... Year..... Subject.....

F- چرا Feature selection در Model-building اهمیت دارد؟

Feature selection فرآیند انتخاب مهم‌ترین ویژگی‌ها از میان تمام ویژگی‌های موجود در مجموعه داده است. این کار باعث بهبود کارایی مدل، کاهش پیچیدگی و جلوگیری از overfitting می‌شود. مزایای Feature selection در مدل‌سازی عبارت است از:

۱- کاهش ~~overfitting~~ overfitting ۲- بهبود دقت و عملکرد مدل

۳- کاهش زمان پردازش و هزینه محاسباتی ۴- بهبود تفسیرپذیری مدل

G- Duplicate Data چگونه در پایگاه داده‌ها حذف می‌شود؟

داده‌های تکراری زمانی رخ می‌دهند که ردیف‌های مشابه یا یکسان در پایگاه داده ذخیره شوند. حذف این داده‌ها برای بهبود کارایی، کاهش حجم داده‌ها و جلوگیری از تحلیل‌های نادرست ضروری است.

روش‌ها: ۱- استفاده از SQL برای حذف داده‌های تکراری

۲- حذف داده‌های تکراری در Pandas

۳- حذف داده‌های تکراری در Excel

AVANGE

Day.... Month.... Year..... Subject.....

H. Irrelevant Data چه مشکلاتی را در پیش‌بینی‌های Machine Learning ایجاد می‌کند؟

Irrelevant Data (داده‌های نامربوط) به ویژگی‌هایی گفته می‌شود که اطلاعات مفیدی برای مدل یادگیری ماشین ارائه نمی‌دهند و حتی ممکن است باعث کاهش دقت مدل شوند.

این نوع داده‌ها معمولاً نویز ایجاد کرده و باعث کاهش کارایی، افزایش پیچیدگی و ایجاد overfitting می‌شوند. موارد زیر مشکلات به وجود آمده می‌باشند:

۱- کاهش دقت مدل، ۲- افزایش overfitting ۳- افزایش پیچیدگی محاسباتی

۴- سخت‌تر شدن تفسیر مدل ۵- کاهش عملکرد مدل‌های مبتنی بر فاصله

I. چرا Data Imputation برای پر کردن Missing Values کاربرد دارد؟

Data Imputation یک تکنیک مهم در پیش‌پردازش داده‌ها است که به جای حذف داده‌های ناقص مقدارهای مناسب را جایگزین می‌کند. این کار باعث می‌شود داده‌های از دست رفته تأثیر منفی روی دقت مدل نداشته باشند و به حفظ کیفیت و کامل بودن مجموعه داده کمک کند.

دلایل اهمیت و کاربرد: ۱- جلوگیری از کاهش حجم داده‌ها ۲- بهبود دقت مدل یادگیری ماشین

Day.... Month.... Year..... Subject.....

۳- جلوگیری از Bias در داده‌ها ۴- افزایش پایداری و قابلیت تعمیم مدل

۱- چگونه می‌توانید Normality را در داده‌های عددی بررسی کنید؟

Normality به این معنی است که داده‌ها از یک توزیع نرمال پیروی می‌کنند. بسیاری از الگوریتم‌های یادگیری ماشین (مانند رگرسیون خطی، ANOVA و آزمون‌های آماری پارامتریک) فرض می‌کنند که داده‌ها دارای توزیع نرمال هستند. برای بررسی نرمال بودن داده‌ها می‌توان از روش‌های آماری و گرافیکی استفاده کرد.

۱- روش‌های گرافیکی

۱- رسم Histogram ۲- رسم Q-Q Plot (Quantile-Quantile Plot)

۲- روش‌های آماری

۱- آزمون شاپیرو-ویلک ۲- آزمون کولموگروف-اسمیرنوف

۳- آزمون جارک-برا

۳- بررسی Skewness و Kurtosis

AVANGE

Day.... Month.... Year..... Subject.....

Skewness (چولگی): اگر مقدار چولگی نزدیک به صفر باشد، داده ها متقارن و نرمال هستند.

Kurtosis (کشیدگی): اگر مقدار کشیدگی نزدیک به ۳ باشد، داده ها دارای توزیع نرمال هستند.