۳۰ Day ۱ Month ۱۴۰۲ Year

Subject: «بنام خدا»

نام درس:

سری تمرین: پنجم ۲۴

Data Preprocessing

نام دانشجو: معصومه منزع

واژه ترم: ورودی بهمن ۱۴۰۳

نام مدرس: مهندس استاد احمدزاده

Subject: تمرینات مباحث ویژه Year ۱۴۰۴ Month ۱ Day ۳۰

A. چرا Data cleaning در علم داده اهمیت دارد؟

پاک سازی داده ها مرحله ای اساسی در علم داده است، زیرا داده های Data cleaning. نادرست یا ناقص می‌توانند منجر به نتایج نادرست و تصمیم‌گیری های غلط شوند. پاک سازی داده ها شامل شناسایی و اصلاح یا حذف داده های نادرست، ناقص، تکراری یا غیر مرتبط است.

با وجود داده های تمیز و بهینه، مدل های یادگیری ماشین می‌توانند عملکرد بهتری داشته باشند و پیش بینی های دقیق تری ارائه دهند.

B. missing values چگونه مدیریت می‌شود؟

Missing values یا مقادیر گمشده به داده هایی اشاره دارد که در یک مجموعه داده وجود ندارد. مدیریت این مقادیر می‌تواند شامل چندین روش باشد:

حذف رکوردها: رکورد هایی که شامل مقادیر گمشده هستند، می‌توانند حذف شوند.

جایگزینی: مقادیر گمشده می‌توانند با میانگین، میانه یا مد داده های موجود جایگزین شوند.

مدل سازی: استفاده از الگوریتم های خاص برای پیش بینی مقادیر گمشده بر اساس سایر ویژگی ها.

C. outliers چیست و چگونه می‌توانید آن ها را تشخیص دهید؟

outliers یا نقاط دور افتاده، داده هایی هستند که به طور قابل توجهی از سایر نقاط داده متفاوت هستند. این نقاط می‌توانند ناشی از خطا در جمع آوری داده ها یا موارد واقعی باشند. تشخیص outliers می‌تواند با استفاده از روش های زیر انجام شود

نمودار جعبه‌ای (Box Plot): برای شناسایی نقاط دورافتاده. روش Z-Score: محاسبه Z-Score برای شناسایی داده‌هایی که بیشتر از ۳ انحراف استاندارد از میانگین فاصله دارند.

روش IQR: محاسبه فاصله بین کوانتایل‌های اول و سوم و شناسایی نقاطی که خارج از محدوده مشخص شده قرار دارند.

D. DataTransformation چه کاربردی دارد؟

یا تبدیل داده‌ها به فرآیند تغییر شکل داده‌ها به منظور بهبود کیفیت داده‌ها و قابلیت استفاده آن‌ها اشاره دارد. این فرآیند شامل استانداردسازی، نرمال‌سازی و دسته‌بندی به مدل‌ها کمک می‌کند تا بهتر یاد بگیرند و عملکرد DataTransformation داده‌ها می‌شود. بهتر دانشته باشند، همچنین می‌تواند سرعت پردازش را نیز افزایش دهد.

E. تفاوت Label Encoding, Encoding Techniques, One-Hot Encoding چیست؟

Label Encoding: در این روش هر دسته به یک عدد صحیح اختصاص داده می‌شود. این روش برای داده‌های ترتیبی مناسب است.

One-Hot Encoding: در این روش هر دسته به یک بردار باینری تبدیل می‌شود که تنها یک عنصر آن ۱ و بقیه ۰ هستند. این روش برای داده‌های غیر ترتیبی مناسب است و مانع از ایجاد ترتیب نادرست بین دسته‌ها می‌شود.

F. چرا Feature Selection در Model-building اهمیت دارد؟

Feature Selection : یا انتخاب ویژگی ها به فرآیند انتخاب زیرمجموعه‌ای از ویژگی ها مرتبط با هدف برای استفاده در مدل‌سازی اشاره دارد.

این عمل به کاهش ابعاد داده‌ها افزایش دقت مدل و کاهش زمان آموزش کمک می‌کند.

همچنین با کاهش نویز و جلوگیری از Overfitting به بهبود کارایی مدل کمک می‌کند.

6. چگونه Duplicate Data در پایگاه داده‌ها حذف می‌شود؟

Duplicate Data یا داده‌های تکراری می‌توانند با استفاده از روش‌های زیر در پایگاه داده‌ها حذف شوند:

استفاده از توابع گروهی: مانند GROUP BY در SQL جهت شناسایی و حذف رکوردهای تکراری.

استفاده از روش‌های برنامه‌نویسی: مانند استفاده از پانداس در پایتون جهت شناسایی و حذف داده‌های تکراری DataFrame.

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

داده‌های غیرمرتبط می‌توانند باعث کاهش دقت مدل‌های یادگیری ماشین Irrelevant Data شوند و زمان آموزش را افزایش دهند Overfitting شوند. این داده‌ها می‌توانند موجب همچنین، مدل را دچار سردرگمی می‌کنند و ممکن است به نتایج نادرست منجر شوند.

8. چرا Data Imputation جهت برخورد با Missing Values کاربرد دارد؟

جایگزین کردن مقادیر گمشده تکنیکی است که جهت بهبود کیفیت داده‌ها و Data Imputation جلوگیری از مشکلات ناشی از مقادیر گمشده استفاده می‌شود. این کار باعث

Subject: Year 1401 Month 1 Day ۳۰

حفظ اطلاعات و افزایش دقت مدل های یادگیری ماشین می شود.

ل. چگونه می توانید Normalize را در داده های عددی بررسی کردی

برای بررسی Normalize یا نرمال بودن داده های عددی می توان از روش های زیر استفاده کرد:

نمودار هیستوگرام: برای مشاهده توزیع داده ها.

نمودار Q-Q: برای مقایسه توزیع داده ها با توزیع نرمال.

آزمون های آماری: مانند آزمون shapiro_wilk یا Kolmogorov-Smirnov برای بررسی نرمال بودن داده ها.

در نهایت Duplicate و مدیریت داده ها به عنوان مراحل حیاتی در علم داده یاد گیری ماشین به شمار می آیند و تاثیر قابل توجهی بر عملکرد مدل ها دارند.