Subject : _____ Date : _____

بخش چهارم :

A : چون داده‌های کثیف (ناقص، نگارش، اشتباه) دقت مدل را کم می‌کند و Data cleaning کثیف : داده را بالاتر می‌برد، خطاها را کم می‌کند و نتایج قابل اعتماد می‌دهد.

B :

حذف => ردیف یا ستون با داده گمشده حذف می‌شود (درصد گم‌شدن ؟)

پرکردن => با میانگین، میانه، یا مقدار ثابت (مثل 0) پر می‌شود.

پیش‌بینی => از مدل‌های ساده برای حدس زدن استفاده می‌شود.

C : داده‌هایی که از بقیه خیلی متفاوت‌اند (مثلاً خیلی بزرگ یا خیلی کوچک)

راه تشخیص : 1- IQR => داده خارج از 1.5 برابر IQR 2- Z-score => داده‌هایی که بیشتر از 3 انحراف معیار از میانگین دور هستند. 3- رسم Box Plot یا scatter هم کمک می‌کند

D : 1- داده‌ها را به فرمت مناسب برای مدل تبدیل می‌کند 2- مقیاس‌ها را یکسان می‌کند

3- روابط غیرخطی را ساده‌تر می‌کند تا مدل بهتر یاد بگیرد.

PAN

Date : _____ Subject : _____

E:

Label Encoding ← دسته‌ها را به اعداد تبدیل می‌کند (مثلاً: قرمز = ۰، آبی = ۱)

برای داده‌های ترتیبی خوب است.

One-hot Encoding ← برای هر دسته یک ستون باینری می‌سازد به طور مثال

(قرمز = [۱، ۰، ۰]، آبی = [۰، ۱، ۰]) برای داده‌های غیر ترتیبی مناسب است

به طور خلاصه: Label ترتیب فرض می‌کند ولی one-hot نه.

F: ۱- ویژگی‌های غیرضروری را حذف و پیچیدگی مدل را کم می‌کند. ۲- سرعت یادگیری را

بالا می‌برد و از overfiting جلوگیری می‌کند ۳- دقت پیش‌بینی را با تمرکز روی داده‌های مهم بیشتر می‌کند

G: ۱- با دستورات SQL مثل Group By یا Distinct.

۲- در پایتون با () drop_duplicates.

۳- ابتدا داده‌های تکراری شناسایی می‌شوند و براساس ستون و یا کلید خاص حذف می‌شوند.

H: ۱- مدل را کیج می‌کند و الگوهای اشتباه یاد می‌گیرد. ۲- زمان پردازش و مصرف منابع

را بی‌دلیل زیاد می‌کند. ۳- دقت پیش‌بینی را کم می‌کند چون داده بی‌ربط نویز ایجاد می‌کند.

PAN

Subject : _____ Date : _____

آ: ۱- با برگردن منطق (مثل میانگین یا مدل پیش‌بینی) داده را کامل می‌کند ۲- مدل‌های ML به داده کامل نیاز دارند، پس Imputation ضروری می‌شود ۳- حذف داده گم‌شده همیشه

عکس مثبت

ک:

۱- رسم نمودار هیستوگرام یا Plot Q-Q (اگر نقاط روی یک خط صاف باشند، نرمال است)

۲- آزمون آماری، مثل Shapiro-Wilk یا Kolmogorov-Smirnov بالا یعنی نرمال است

۳- بررسی Skewness و Kurtosis: اگر نزدیک صفر باشد، داده نرمال.