

# التقنيات الأساسية لمعالجة البيانات في تعلم الآلة

رحلة شاملة لتحسين جودة البيانات وتهيئتها للنماذج الذكية

# معلومات التقرير

إعداد الطالب:

فادي كلش

المادة:

Machine Learning

التاريخ:

2026 / 2 / 16

# الملخص التنفيذي

تعتبر مرحلة معالجة البيانات الأولية (Data Preprocessing) من أهم المراحل في مشاريع تعلم الآلة، حيث تؤثر جودة البيانات بشكل مباشر على أداء النماذج. يتناول هذا التقرير شرحاً مفصلاً لأهم التقنيات المستخدمة في هذه المرحلة، بدءاً من التعامل مع القيم المتطرفة، مروراً بتحجيم الخصائص وتقليل الأبعاد، وصولاً إلى معالجة القيم المفقودة وترميز المتغيرات الفئوية.



تقليل الأبعاد  
ضغط المعلومات مع الحفاظ على  
أهم البيانات



تحجيم الخصائص  
توحيد مقاييس المتغيرات  
المختلفة لتحسين أداء النماذج



القيم المتطرفة  
اكتشافها واتخاذ القرارات حول  
الإبقاء عليها أو حذفها

يهدف التقرير إلى تقديم فهم شامل لهذه التقنيات مع توضيح الحالات التي تناسب كل منها، مما يمكن الباحثين والطلاب من اتخاذ قرارات مستنيرة في مشاريعهم.

# المقدمة

"القمامة تدخل، قمامنة تخرج" (Garbage In, Garbage Out)

في عالم تعلم الآلة، هناك مقوله شهيرة تلخص أهمية جودة البيانات: إذا كانت البيانات المدخلة للنموذج رديئة الجودة، فستكون المخرجات (النتائج) رديئة أيضاً، بغض النظر عن مدى تعقيد النموذج المستخدم.

تمثل مرحلة معالجة البيانات الأولية ما بين **60% إلى 80%** من الوقت الكلي لأي مشروع تعلم آلة. تتضمن هذه المرحلة تنظيف البيانات وتحويلها إلى شكل مناسب للنماذج الرياضية.

03

تقليل الأبعاد

ضغط المعلومات

02

تحجيم الخصائص

توحيد مقاييس المتغيرات

01

القيم المتطرفة

اكتشاف وإدارة البيانات الشاذة

05

المعالجة الإضافية

القيم المفقودة والترميز

التوزيع الطبيعي

تحسين توزيع البيانات

04

# أولاً: القيم المتطرفة (Outliers)

## تعريف القيم المتطرفة

القيم المتطرفة (Outliers) هي نقاط بيانات تختلف بشكل كبير عن بقية البيانات في المجموعة. يمكن تشبثها بشخص كبير في السن بين مجموعة من الشباب، أو منزل ضخم جداً بين منازل متوسطة الحجم في حي سكني.

مثال توضيحي: إذا كانت بيانات أعمار الموظفين في شركة ما كالتالي: [25, 26, 27, 28, 29, 30, 65] فإن القيمة 65 تعتبر قيمة متطرفة لأنها تختلف بشكل كبير عن بقية الأعمار.

### أشجار القرار

أقل تأثيراً، لكنها قد تسبب فروعًا عميقية غير مفيدة

### K-Means نموذج

تحرك مركز التجمعات بشكل غير صحيح

### الانحدار الخطى

تسحب خط الانحدار نحوها، مما يعطي تقديرات غير دقيقة

## طرق الكشف عن القيم المتطرفة

03

### مخطط الصندوق (Boxplot)

تمثيل بياني لطريقة IQR، حيث تظهر القيم المتطرفة كنقاط خارج شعيرات المخطط.

02

### المدى الربيعي (IQR)

يعتمد على تقسيم البيانات إلى أربع. تعتبر القيمة متطرفة إذا كانت أقل من  $Q1 - 1.5 \times IQR$  أو أكبر من  $Q3 + 1.5 \times IQR$ .

01

### Z-Score أسلوب

يحسب عدد الانحرافات المعيارية التي تبعدها القيمة عن المتوسط. تعتبر أي قيمة ذات Z-Score أكبر من 3 أو أقل من -3 قيمة متطرفة.

## متى نحذف ومتى نبقي القيم المتطرفة؟

الحالات	القرار	السبب
خطأ في إدخال البيانات	حذف	البيانات لا تمثل الواقع
عطل في جهاز القياس	حذف	القيمة غير حقيقة
ظاهرة حقيقة نادرة	إبقاء	هذه الحالات هي هدف الدراسة
كمية قليلة جداً	حذف	تأثيرها ضئيل وقد تشوش النموذج

# ثانياً: تحجيم الخصائص (Feature Scaling)

## أهمية تحجيم الخصائص

تعتمد العديد من خوارزميات تعلم الآلة على حساب المسافات بين النقاط أو على النسب المئوية. عندما تكون الخصائص بمقاييس مختلفة، تهيمن الخصائص ذات القيم الكبيرة على العملية الحسابية.

مثال توضيحي: لدينا خاصيتين لوصف الموظفين: الراتب (3000-10000) والعمر (22-60). عند حساب المسافة بين موظفين، سيهيمن الراتب على المعادلة و يجعل تأثير العمر شبه معدوم، رغم أن العمر قد يكون مهمًا للتنبؤ.

### طرق التحجيم الرئيسية

نوع التحجيم	المعادلة	الصيغة	الناتج	الأمثلية
Min-Max Scaling (التطبيع)	$X_{scaled} = \frac{(X - X_{min})}{(X_{max} - X_{min})}$	الصيغة: $X_{scaled} = \frac{(X - X_{min})}{(X_{max} - X_{min})}$	الناتج: جميع القيم بين 0 و 1	الأمثلية: عندما يكون للبيانات حد أدنى وأعلى معروفيين
Standardization (التوحيد القياسي)	$X_{scaled} = \frac{(X - \text{Mean})}{\text{Standard Deviation}}$	الصيغة: $X_{scaled} = \frac{(X - \text{Mean})}{\text{Standard Deviation}}$	الناتج: متوسط = 0، انحراف معياري = 1	الأمثلية: الطريقة الأكثر شيوعاً، وتعمل جيداً مع وجود قيم متطرفة
Robust Scaling (التحجيم المتنين)	$X_{scaled} = \frac{(X - \text{Median})}{\text{IQR}}$	الصيغة: $X_{scaled} = \frac{(X - \text{Median})}{\text{IQR}}$	الناتج: يعتمد على الوسيط والمدى الرباعي	الأمثلية: عندما تكون القيم المتطرفة كثيرة

### جدول المقارنة

الطريقة	المعادلة	الناتج	متى نستخدمها؟
Min-Max	$(X-min)/(max-min)$	$[0, 1]$	بيانات ذات حدود معروفة
Standardization	$(X-mean)/std$	متوسط 0	الحالة العامة
Robust Scaling	$(X-median)/IQR$	غير محدد	بيانات بها قيم متطرفة كثيرة

# ثالثاً: تقليل الأبعاد (Reduction Dimensionality)

## تعريف وأهمية تقليل الأبعاد

تقليل الأبعاد هو عملية تقليل عدد المتغيرات (الأعمدة) في مجموعة البيانات مع الاحتفاظ بأكبر قدر ممكن من المعلومات الهامة. عند وجود بيانات عالية الأبعاد (مئات أوآلاف الأعمدة)، نواجه مشاكل عديدة: بطء النماذج، التجهيز المفرط (Overfitting)، وصعوبة التصور.

## تحليل المكونات الرئيسية (PCA)

هي أشهر تقنية لتقليل الأبعاد. تعمل على إيجاد اتجاهات التباين في البيانات وإنشاء مكونات رئيسية جديدة تمثل هذه الاتجاهات.



### إنشاء مكونات رئيسية

تصنع متغيرات جديدة تمثل هذه الاتجاهات



### إيجاد اتجاهات التباين

تبحث PCA عن الاتجاهات التي تنتشر فيها البيانات أكثر



### اختيار الأهم

نختار أول مكونين أو ثلاثة ونترك الباقي



### ترتيب المكونات

ترتب المكونات من الأهم للأقل أهمية حسب كمية المعلومات

مثال: إذا كان لديك 10 امتحانات لطالب (رياضيات، فيزياء، كيمياء، عربي، إنجليزي، تاريخ,...). يمكن لـ PCA أن تجمع المواد العلمية في مكون واحد، واللغات في مكون ثان، والمواد الاجتماعية في مكون ثالث.

## رابعاً: تقنيات معالجة إضافية

### معالجة القيم المفقودة (Missing Values)

تعتبر القيم المفقودة مشكلة شائعة في مجموعات البيانات الحقيقية. يجب معالجتها بطرق مناسبة لتجنب التأثير السلبي على أداء النماذج.

<b>التعويض بالوسيط</b>	<b>التعويض بالمتوسط</b>	<b>الحذف (Deletion)</b>
تعبئة القيمة المفقودة بالوسيط	تعبئة القيمة المفقودة بمتوسط العمود	حذف الصفوف التي تحتوي على قيم مفقودة
بيانات رقمية مع قيم متطرفة	بيانات رقمية، توزيع متماثل	عندما تكون النسبة قليلة (< 5%)
<b>التعويض المتقدم (KNN)</b>		<b>التعويض بالمنوال</b>
استخدام أقرب الجيران للتبؤ عندما يكون التبؤ الدقيق مهمًا		تعبئة بأكثر قيمة تكراراً (بيانات فئوية (نصوص))

### ترميز المتغيرات الفئوية (Encoding Categorical Variables)

لا تفهم خوارزميات تعلم الآلة النصوص، لذا يجب تحويلها إلى أرقام.

#### Label Encoding (الترميز الرقمي البسيط)

تحويل كل فئة إلى رقم فريد

مثال: أحمر=1، أخضر=2، أزرق=3

**متى نستخدم؟** عندما يكون هناك ترتيب طبيعي (صغر، وسط، كبير)

#### One-Hot Encoding (الترميز الواحد-)

إنشاء عمود مستقل لكل فئة

مثال: عمود أحمر (1/0)، عمود أخضر (0/1)، عمود أزرق (0/1)

**متى نستخدم؟** عندما لا يوجد ترتيب بين الفئات

### طرق أخذ العينات (Sampling Methods)

عندما تكون البيانات غير متوازنة (فئة واحدة أقل بكثير من الأخرى)، يميل النموذج لتجاهل الفئة الصغيرة.

الطريقة	الوصف	الميزة	العيوب
Oversampling	زيادة عينات الفئة القليلة	نستفيد من كل البيانات	قد يسبب تعمية
Under-sampling	تقليل عينات الفئة الكبيرة	تدريب أسرع	خسر بيانات مهمة
SMOTE	صنع عينات جديدة ذكية	بيانات متنوعة	معقد حسابياً

# خامساً: أهمية تحويل البيانات للتوزيع الطبيعي

## ما هو التوزيع الطبيعي؟

التوزيع الطبيعي (Normal Distribution) هو توزيع إحصائي على شكل جرس، حيث تتركز معظم القيم حول المتوسط وتقل تدريجياً باتجاه الأطراف. هذا التوزيع مهم جداً في تعلم الآلة والإحصاء.

### لماذا نفضل التوزيع الطبيعي؟

استقرار التباين  
 يجعل التباين ثابتاً عبر القيم



الانحدار الخطي واللوجيستي  
يفترض أن الأخطاء موزعة طبيعياً



تحسين المسافات  
 يجعل المسافات أكثر استقراراً



تقليل القيم المتطرفة  
يقلل من تأثير القيم الكبيرة



### طرق تحويل البيانات للتوزيع الطبيعي

تحويل بوكس-كوكس  
 $(x^{\lambda} - 1)/\lambda$

الحالة العامة، يجد أفضل تحويل  
تلقاءياً

الجذر التربيعي  
 $\sqrt{x}$

بيانات عدّة (Count Data)

التحويل اللوغاريتمي  
 $\log(x)$

بيانات منحرفة بشدة (قيم كبيرة جداً)

**ملاحظة:** تحويل البيانات يساعد في تحسين أداء النماذج الإحصائية والمبنية على المسافات، ويجعل التباين أكثر استقراراً عبر قيم الخاصية.

# الخاتمة

## النتيجة الرئيسية

الآن، في الوقت الحاضر، أصبحت تقنيات معالجة البيانات الأولية ضرورية للنجاح في تعلم الآلة. البيانات النظيفة والمنظمة هي الطريق الأكيد لنماذج دقيقة.

تعتبر مرحلة معالجة البيانات الأولية (Data Preprocessing) حجر الأساس لأي مشروع تعلم آلة ناجح. من خلال هذا التقرير، استعرضنا أهم التقنيات المستخدمة في هذه المرحلة:

### القيم المتطرفة

اكتشاف باستخدام Z-Score و IQR



### تحجيم الخصائص

Min-Max و Standardization و Robust Scaling



### تقليل الأبعاد

دور PCA في ضغط المعلومات



### تقنيات إضافية

القيم المفقودة، الترميز، أخذ العينات



### التوزيع الطبيعي

تحسين أداء النماذج



"في النهاية، لا يوجد حل واحد يناسب جميع الحالات."

يجب على محلل البيانات فهم طبيعة بياناته ومتطلبات المشروع لاختيار التقنيات المناسبة. البيانات النظيفة والمنظمة هي الطريق الأكيد لنماذج تعلم آلة دقيقة وموثوقة.