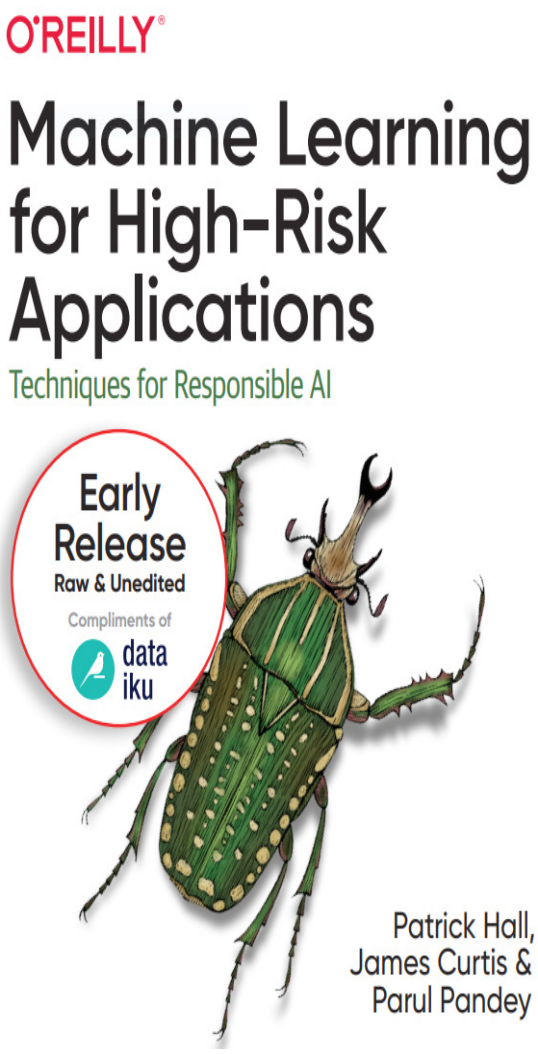
تعلم الآلة للتطبيقات عالية المخاطر

المؤلفون: بتريك هال، جيمـِس كورتيس ، و بارول باندي

الناشر: أورلي

تاريخ النشر: ٢٠٢٢

عدد الصفحات: ١٦٩ صفحة



يعتبر هذا الكتاب ممثلا للجزء الأول فقط ويحتوي علي خمسة فصول أساسية هي:

الفصل ١: حوكمة النموذج المعاصر، الذي يناقش الإلتزمات القانونية؛ حوادث الذكاء الاصطناعي؛ الكفاءات التنظيمية والثقافية للذكاء الاصطناعي المسئول؛ العمليات التنظيمية للذكاء الاصطناعي المسئول؛ ودراسة حالة صعود وهبوط Zillow›s iBuying

الفصل ٢: تعلم الآلة القابل للتفسير والشرح، يتعرض لكل من الأفكار المهمة لقابلية التفسير والشرح؛ النماذج القابلة للتفسير؛ الصعوبات العنيدة في الشرح اللاحق للسلامة والأداء؛ حالة متدرجة بواسطة الخوارزمية.

الفصل ٣: تصحيح تعلم الآلة للسلامة والأداء، يناقش كل من التدريب؛ تصحيح النموذج؛ النشر؛ ودراسة حالة: الموت بواسطة المركبات المستقلة ذاتية القيادة.

الفصل ٤: الأمن لتعلم الآلة، يتعرض لأساسيات

الأمن؛ هجمات تعلم الآلة؛ اهتمامات أمن الذكاء الاصطناعي العامة؛ التدابير المضادة؛ و دراسة حالة هجمات التهرب في العالم الحقيقي.

الفصل ٥: آلات التعزيز القابلة للتفسير والشرح XB Boost، يناقش تجديد المفهوم؛ مع إدارة الحساب العالمي GAM لعلاقات طويلة الأجل من أجل إنشاء فرص أعمال جديدة وعائد أعظم ، مع قيود ذكاء اصطناعي قابلة للشرح.

هذا الكتاب يناقش القضايا المختلفة من منظور تطبيق عملي مع توضيح ابعاد النظرية عندما يكون ذلك ضروريا. ويبدأ **الفصل الأول** من هذا الكتاب بالغوص العميق في اللوائح المعلقة؛ كما يناقش المسئولية عن المنتجات ؛ ومعالجة إدارة النموذج التقليدي. حيث أن كثيرا من الممارسات تفترض مدخلا مهنيا للنمذجة، فقد تم مناقشة كيفية تضمين الممارسات الأحسن لأمن الحاسب الآلي التي تفترض الفشل في حوكمة النموذج.

أما **الفصل الثاني** يعرض النظام البيئي المزدهر للنماذج القابلة للتفسير، كما يغطي فئة النموذج المضاف المعمم GAM بتعمق كبير، مع مناقشة أنواعا أخري كثيرة من مقدرات الشفافية أيضا، إلي جانب تحديد العديد من أساليب التفسير اللاحقة مع النظر نحو الصرامة والمشاكل المعروفة. وفي **الفصل الثالث** تعرض لمعالجة تصحيح النموذج لكن بطريقة تختبر افتراضات النموذج بالفعل مع أداء العالم الحقيقي. وعلي ذلك يتعرض لاختبار أساسيات البرمجيات بالإضافة لاستخلاص النقاط البارزة في مجال تصحيح أخطاء النموذج الجديد.

ويلقي **الفصل الرابع** نظرة عامة علي الجوانب الفنية للعدالة والتحيز في تعلم الآلة، بدءا بتصميم تجريبي ملائم، وجمع البيانات والوكلاء، ثم بعدئذ يعالج اختبار التحيز بالتفصيل متضمنا الاختبارات للتأثير المتباين، والصلاحية التفاضلية مع الكشف عن دوافع التحيز مع قيم شابلي Shapley Values إلي جانب أساليب شرح أخري. وبذلك يخاطب هذا الفصل كلا من الطرق المنشأة والمحافظة بجاناب أساليب المعالجة القبلية والبعدية، والمزيد من أحدث الأهداف المزدوجة والعدائية وأساليب المعالجة قبل وداخل وبعد المعالجة.

ويختتم الكتاب **بالفصل الخامس** الذي يضع كيف أن فريق نظم تعلم الآلة يبدأ بأساسيات أمن الحاسب الآلي ، ويناقش هجومات تعلم الألة المشترك، وتعلم الآلة العدائي والقوي أيضا.