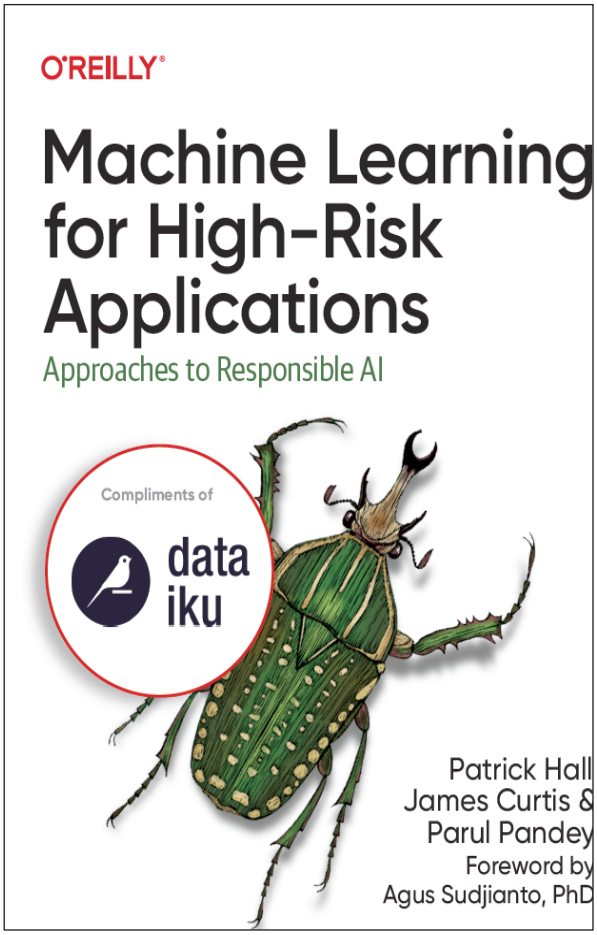
التعلم الآلي للتطبيقات عالية المخاطر: مداخل الذكاء الاصطناعي المسؤول



عنوان الكتاب: تعلم الآلة للتطبيقات عالية المخاطر: مداخل الذكاء الاصطناعي المسؤول.

المؤلفون: باتريك هول؛ جيمس كورتيز & بول باندي

الناشر: أورلي

تاريخ النشر: أبريل ٢٠٢٣

عدد الصفحات: ٤٦٩ صفحة.

يشتمل هذا الكتاب علي ثلاثة أجزاء رئيسية تتفرع جميعها الي ١٢ فصلا كما يلي

الجزء الأول: نظريات وتطبيقات عملية لإدارة مخاطر الذكاء الاصطناعي:

١. إدارة مخاطر التعلم الآلي المعاصر

٢. التعلم الآلي القابل للتفسير والشرح

٣. تصحيح أخطاء نظام التعلم الآلي للسلامة والأداء

٤. إدارة التحيز في التعلم الآلي

٥. الأمن للتعلم الآلي

الجزء الثاني: وضع مخاطر الذكاء الاصطناعي موضع التنفيذ:

٦. آلات التعزيز القابلة للتفسير وXGBoost القابلة

٧. شرح مصفي الصو PYTorch

٨. إختيار وتصحيح نماذج XGBoost

٩. تصحيح أخطاء مصنف الصور PyTorch

الجزء الثالث: الاستنتاج

١٢. كيفية النجاح في التعلم الآلي عالي المخاطر

ذات مرة قال الأحصائي الشهير جورج بوكس George Box أن «جميع النماذج خاطئة. لكن بعضها مفيد». ويشكل الاعتراف بهذه الحقيقة أساسي الإدارة الفعالة للمخاطر. وفي عالم حيث يعمل التعلم الآلي على نحو متزايد على آلية القرارات المهمة

المتعلقة بحياتنا، فإن العواقب المترتبة علي فشل النماذج يمكن أن تكون كارثية. ومن الأهمية بمكان اتخاذ خطوات مدروسة جيدا للتخفيف من المخاطر وتجنب الضرر غير المقصود.

وفي هذا الكتاب عن التعلم الآلي للتطبيقات عالية المخاطر، عرض مؤلفوه إطار عمل لتطبيق التعلم الآلي في اتخا القرار عالي المخاطر. فهم يقدمون أدلة دامغة من خلال حالات موثقة لفشل النماذج واللوائح التنظيمية الناشئة التي تسلط الضوء علي أهمية الحوكمة والثقافة القوية. إلا أنه لسوء الحظ، لا تزال هذه المبادئ نادرة وما يتم تنفيذها خارج الصناعات الخاضعة للتنظيم، مثل البنوك. ويغطي هذ الكتاب موضوعات مهمة تتراوح بين شفافية النموذج والحوكمة والأمن وإدارة التحيز والمزيد.

كما أن اختبار الأداء وحده لا يكفي في التعلم الآلي، حيث يمكن أن يكون للنماذج المختلفة نفس الأداء بسبب تعدد النماذج. كما يجب أيضا أن تكون النماذج قابلة للتفسير وآمنة وعادلة. وبذلك، يعتبر هذا الكتاب الأول من نوعه الذي يؤكد علي النماذج القابلة للتغيير بطبيعتها وتطورها وتطبيقها الحديث، خاصة في الحالات التي تؤثر فيها النماذج علي الأفراد، مثل التمويل الاستهلاكي. وفي هذه السيناريوهات، حيث تكون معايير ولوئح قابلية التفسير صارمة من الناحية العملية. وغالبا ما يواجه مدخل قابلية التفسير للذكاء الاصطناعي القابل للتفسير (XAI) تحديات كبيرة.

ويتطلب تطوير نظم التعلم الآلي إلي موثوقية وأمن أيضا تقييما دقيقا لنقاط ضعف النموذج. ويقدم هذا الكتاب عينيتين من خلال منهجية تصحيح أخطأ أو التقطيع المتبقي، وتقييم متانة النموذج في فساد المدخلات، وتقييممموثوقية مخرجات النموذج، واختبار مرونة النموذج في ظل تحول التوزيع من خلال اختبار الإجهاد. وتعتبر هذه الموضوعات حاسمة لتطوير ونشر التعلمالآلي في البيئات عالية المخاطر.

كما أن نماذج التعلم الآلي لديها القدرة علي الإضرار بشكل غير متناسب بالمجموعات المهمشة تاريخيا، وتقديم هذا القدر بسرعة وعلي نطاق واسع من خلال الآلية، والقرارات النموذجية المتحيزة لها أثارا ضارة علي المجموعات المحمية، مما يؤدي إلي إدامة الفوارق الاجتماعية والاقتصادية. وفي هذا الكتاب سيتعلم القارئ كيفية التعامل مع مسألة عدالة النموذج من خلال عدسة اجتماعية فنية. وقد قام المؤلفون أيضا بالتفصيل من خلال دراسة تأثيرات أساليب إزالة النماذج، وتقديم لمشورة العملية حول تطبيق هذه الأساليب في قطاعات مختلفة كثيرة ومنظمة.

وبذلك، يمكن القول أن التعلم الآلي للتطبيقات عالية المخاطر صار منتشرا لحد كبير. وبذلك، فإن هذا هو كتاب عملي إلي حد كبير، وفي الوقت نفسه سيجد القراء من جميع المواقع أبعادا مهمة حول هذا الموضوع، سواء كانوا يمثلون عالم يانات مهتما بفهم النماذج بشكل أفضل، أو مديرا مسؤولاعن ضمان الامتثال للمعايير الحالية، أو مسؤول تنفيذي يحاول تخسين ضوابط المخاطر في منظمته.

وقد أعد هذا الكتاب لمساعدة القراء علي إنجاز أعمالهم. وبشكل عام قد تم تقديم نموذج تعليمات البرمجية مع هذا العمل. وبذلك يمكن استخدامها في البرامج والتوثيق.