**الخطوة الأولى: فهم المشروع وتحديد الاهداف**

قبل أن نقوم بجمع البيانات، نحتاج إلى فهم أهداف المشروع. لتقليل وضع المتغيرات المناسبة، وتقليل وقت جمع البيانات وفهمها.

مجموعة بيانات الدخل هي الدخل السنوي للفرد الناتج عن عوامل مختلفة. قدمت مجموعة بيانات القطار ميزة تنبؤية مثل عدد التعليم والوظيفة وفئة العمل والحالة الاجتماعية وما إلى ذلك، للتنبؤ بما إذا كانت هذه العوامل يمكن أن تؤثر على دخل الموظفين أم لا.

استكشفنا إمكانية التنبؤ بمستوى الدخل (أكثر من 50 كلفن أو أقل) بناءً على المعلومات الشخصية للفرد من خلال ممارسة مشكلة التعلم الآلي مثل التصنيف والانحدار (توقع دخل العملاء)

هذه مشكلة تعلم آلي خاضعة للإشراف لأن لدينا هنا مجموعة بيانات مصنفة.

المتغير التابع (Y) هو الدخل.

- ينقسم الدخل إلى فئتين: <= 50 ألف و> 50 ألف

المتغيرات المستقلة (X) هي: العمر ونوع التعليم والمهنة وما إلى ذلك.

**الخطوة الثانية: التنقيب عن البيانات**

جمع البيانات التي احتاجها.. وفي ها المثال تم الحصول على قاعدة الدخل المطلوبة للتحليل من موقه Kaggle

قمنا بتحميل ملف البينات واستحضاره بالكود التالي

**الخطوة الثالثة: تنظيف وتحضير البيانات قبل البدء في تحليلها**

البيانات التي اخترتها هي مجموعة بيانات الدخل للتصنيف الثنائي (https://www.kaggle.com/mastmustu/income). تقدم مجموعة البيانات هذه مجموعة واسعة من الميزات. بعض الأعمدة لها قيمة مفقودة مثل؟ نحن بحاجة إلى تحويل لا شيء الفعلي.

كان لدينا 3 خصائص فيها نسب بسيطة من القيم المفقودة وهي: work class – occupation – native country وتم تعديلها. والحصول على null

وهذا يعني ان البيانات لدينا جاهزة

**الخطوة الرابعة: استكشاف البيانات**

الآن بعد أن حصلت على مجموعة بيانات رائعة ونظيفة، فأنت جاهز للبدء أخيرًا في تحليلك. تشبه مرحلة استكشاف البيانات عملية العصف الذهني لتحليل البيانات. باستخدام Pandas ، أو رسم مخطط بياني أو منحنى توزيع لمعرفة الاتجاه العام ، أو حتى إنشاء تصور تفاعلي يتيح لك الغوص في كل نقطة بيانات واستكشاف القصة وراء القيم المتطرفة .

باستخدام كل هذه المعلومات، تبدأ في تكوين فرضيات حول بياناتك والمشكلة التي تعالجها

في هذا المثال، قدمت مجموعة البيانات ميزة تنبؤية مثل العمر، وعدد التعليم، وحالة التوظيف، والحالة الاجتماعية للتنبؤ بما إذا كان الدخل أكبر من 50 ألف دولار. يمكن استخدامه لممارسة مشكلة التعلم الآلي مثل التصنيف الثنائي. لذلك، نحتاج إلى التنبؤ بدقة بما إذا كان شخص ما يربح أكثر أو أقل من 50000 دولار في السنة أم لا.

هنا مثال بين الذكور والاناث والدخل

نلاحظ من الرسمة أن الكور اعلى من الاناث في الدخل السنوي اكثر من 50 الف

**الخطوة الخامسة: هندسة الخصائص**

في التعلم الآلي، الميزة هي خاصية قابلة للقياس أو سمة لظاهرة يتم ملاحظتها. إذا كنا نتوقع دخل العملاء، فإن الميزة المحتملة هي مقدار الدخل الذي يحصلون عليه.

نقوم عادةً بنوعين من المهام في هندسة الميزات - اختيار الميزات والبناء. في هذا المثال، استخدمت اختيار الميزة.

واختا ر لي خمس خصائص للتوقع وهي: العمر، الوزن المالي fnlweight ,

كما ترونها في الرسمة

**الخطوة السادسة: النمذجة التنبؤية**

لقد استخدمت النمذجة التنبؤية للتأكد من أن نتائج النموذج منطقية وذات مغزى.

بعد ذلك، من الأهمية بمكان أن تقيم نجاحها. عملية تسمى التحقق من صحة k-fold. هذا يسمح للنموذج بالتدريب على جميع البيانات بدلاً من استخدام تقسيم نموذجي للاختبار.

في هذا المثال، استخدمت (logisticRegression (LR) ، KNeighboursClassifier (KNN) ، DecisionTreeClassifier (DTC) ، RandomForestClassifier (RFC)).

كما هو موضح أعلاه، فإن Random Forest لديها قيمة أعلى دقة ب RFC: 0.9243. ثم يأتي بعدها البقية

**الخطوة السابعة: عرض مرئي للمعلومات**

نحن هنا نمثل النموذج بطريقة يمكن أن يفهمها مختلف أصحاب المصلحة في المشروع.

ناحية ، العلاقة الإيجابية بالدخل هي: Age، Workclass Federal-، Workclass private، fnlwgt، .. جميع المتغيرات باللون الأزرق.

في حين أن العلاقة السلبية مع الدخل هي: Workclass لم يعمل أبدًا ، التعليم السابع - الثامن ، الحالة العسكرية لم تتزوج أبدًا ، المهنة إنهاء الأسرة…. جميع المتغيرات باللون الأخضر.

من ناحية أخرى ، هناك بعض المتغيرات مثل: التعليم قبل المدرسة ، والعلاقة غير المتزوجين ، والمهنة ، والنقل المتحرك ... جميع المتغيرات باللون الأسود. هذه لا تؤثر على الدخل على الإطلاق.

**خاتمة**

هناك عدد قليل من الموظفين يزيد دخلهم عن 50 ك. نوصي بزيادة الساعات الأسبوعية إلى 40 ك. ، كما يمكننا توظيف العامل الحاصل على درجة البكالوريوس والماجستير (بحاجة إلى الزيادة) بدخل أقل من 50 ك. فئة خاصة للدخل المرتفع. باختصار ، يجب أن يشجع الموظفون على زيادة دخلهم بالمتغيرات التي تتعلق بزيادة الدخل مثل العمر وتجنب جميع المتغيرات الأخرى التي لا تؤثر على الدخل أو لها تأثير سلبي عليه.