**2-Data Cleaning**

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

* **باختصار العمليه دى مهمه ف البدايه انك ت check على ان ال dataset الى انتا هتعمل train بيها مفهاش مشاكل ؟**
* **مشاكل من انهى نوع يعنى ؟**
* **1- samples with None value in features**
* **يعنى ال data فاضيه بعض ال libs لما تشوف ان ال data فاضيه بيعتبرها ب 0 واحيانا التعامل ال 0 ده بيئذينى لانوا بيعمل مشكله معايا ف انا محتاج املى الفراغات دى بحاجه معينه**
* **فالمرحله دى اسمها data cleaning**
* **صحيح ان مصطلح ال data cleaning يشمل مشاكل وطرق كتير على حسب نوع ال data عندى ممكن تكون صور فمهوم ال data cleaning هيختلف شويه من عمليات image processing ك preprocessing phase ولكن هنشوف الكلام ده لما نيجى نتعامل مع image data زى مهنشوف**
* **ولكن من احد مشاكل الى بنواجهها ف ال phase دى هيا اال missing values فتعالى نشوف هنتعامل ازاى مع المشكله دى**
* **ومش شرط انها تكون صفر المقصود يعنى قيم لل features غير مقبوله بمعنى لو انا عندى salary feature هل مقبول ان يكون عندى واحد بياخد none value او 0 كمرتب لا طبعا فالقيم دى هتاثر بالسلب على عمليه ال train لو دخلت لل model بالشكل ده**
* **طيب تعالى نشوف بقى ال sk بتتعامل مع ال missing values ازاى ؟**
* **عندى method اسمها imputer دى بتاخد ايه ؟**
* **1-القيمه الى هيا بنسبالك تعتبر missing value**
* **2-هتتبع انهى strategy للتعامل مع الموضوع والحقيقه عندى كذا طريقه ؟**
* **Mean, median, most\_frequent, constant**
* **طيب هل هوا لما يجى يجيب المتوسط مثلا للقيم هل على مستوى ايه طبعا على مستوى العمود يعنى على مستوى ال feature نفسه مش features تانيه**
* **طيب هيقسم على العدد هيعد القيم ال missing ولا لا لا مش هيعد ال missing**