1. **Những vấn đề thách thức của machine learning là gì? Nêu ví dụ cho mỗi loại.**

**+ Bad data**

- Non prepregetative data:

Noisy data: dử liệu thu được có một số phần tử bất thường có thể ảnh hưởng đến kết quả thuật toán.

Sampling bias: dử liệu chỉ được thu thập ở một vùng nào đó và không trải đều ra các trường hợp của bài toán

- Irrelevant feature: dử liệu có các feature không liên quan tới bài toán.

- Missing data:

Không thu được một số các feature củ dử liệu. Có thể loại bỏ các feature có quá nhiều ô trống hoặc bỏ các sample có quá nhiều feature trống. Ngoài ra, có thể thêm dử liệu vào feature trống nếu thấy hợp lý.

**+ Bad algorithsm:**

- Overfitting: Sử dụng model quá phức tạp với bài toán.

- Underfitting: Sử dụng model quá đơn giản với bài toán.

1. **Nếu model của bạn hoạt động tốt trên dữ liệu huấn luyện nhưng kém hiệu quả trên dữ liệu mới thì hiện tượng gì có thể đã xảy ra?**

Có thể hiện tượng overfitting đã xảy ra do chọn model có bật quá cao. Ngoài ra, vì đã sử dụng dử liệu tranning để test nên kết quả tốt nhưng ra thực tế thì kết quả kém.

1. **[Bài này không bắt buộc làm, nhưng làm có điểm thưởng] Tìm hiểu những cách xử lý overfitting.**
2. **Test set là gì? Tại sao cần dùng test set để đánh giá model mà không dùng training set.**

Test set là một phần của dử liệu được cung cấp, nó được dùng để kiểm tra kết quả training.

Nếu ta sử dụng tranning set để đánh giá thì ta sẻ không phản ánh được tính hiệu quả của model vì đây không phải là dử liệu mới.

1. **Hyperparameter tuning là gì? Validation set là gì? Tại sao cần dùng nó.**

Hyperprameters là các tham số mà thuật toán cần để chạy (trừ dử liệu). Hyperprameters có thể ảnh hưởng trực tiếp tới độ hiệu quả của thuật toán.

Hyperparameter tunning là điều chỉnh hyperparameter để thuật toán chạy hiệu quả. Để là được việc đó ta cần tạo một tập dử liệu gọi là Validation Set để kiểm tra độ hiệu quả của Hyperparameter.

Việc sử dụng Hyperparameter tunning sẻ giúp ta tránh được overfitting và underfitting.