A group of colored dots

Description automatically generated with medium confidence

Các giá trị có sự phân tán tự nhiên và nguyên vẹn theo từng đặc trưng. Tuy nhiên, có sự khác biệt về phạm vi giữa các đặc trưng, điều này có thể gây ảnh hưởng đến độ chính xác khi huấn luyện mô hình.

**Giải thích**:

* **Lượng mưa** (precipitation) có một số giá trị ngoại lệ rất lớn, khiến dữ liệu trở nên phân tán không đồng đều, với nhiều điểm tập trung gần giá trị 0.
* Các biến nhiệt độ và gió có sự phân bố đều hơn, với các mối quan hệ khá rõ ràng, như sự tương quan giữa temp\_max và temp\_min (mối quan hệ gần như tuyến tính).

Điều này cho thấy rằng trước khi chuẩn hóa, dữ liệu có sự phân tán lớn, đặc biệt là ở biến lượng mưa, có thể ảnh hưởng đến hiệu suất của mô hình mạng nơ-ron nếu không được xử lý đúng cách.

A diagram of different colored dots

Description automatically generated with medium confidence

Chuẩn hóa giúp các đặc trưng được đưa về cùng một phạm vi giá trị, dễ dàng để mô hình học và tăng cường khả năng hội tụ nhanh hơn. Việc chuẩn hóa không thay đổi cấu trúc mối quan hệ giữa các đặc trưng, chỉ thay đổi phạm vi giá trị, giúp dữ liệu dễ dàng được sử dụng trong các mô hình máy học mà không ảnh hưởng đến đặc điểm ban đầu của dữ liệu.

**Giải thích**:

* Sau khi chuẩn hóa, các điểm dữ liệu được thu gọn lại trong phạm vi hẹp hơn, giúp dữ liệu trở nên đồng đều hơn và dễ dàng cho mô hình mạng nơ-ron trong việc học các mẫu.
* Lượng mưa (precipitation), mặc dù có một số giá trị ngoại lệ lớn trước khi chuẩn hóa, sau chuẩn hóa đã được điều chỉnh để nằm trong một khoảng hợp lý hơn.
* Các mối quan hệ giữa các biến, như giữa temp\_max và temp\_min, vẫn được bảo toàn sau khi chuẩn hóa.

Việc chuẩn hóa giúp mô hình mạng nơ-ron huấn luyện tốt hơn, bởi vì nó loại bỏ các sự khác biệt lớn giữa các thang đo của các biến, giúp tăng tốc quá trình học và cải thiện hiệu suất mô hình.

**Tại sao biểu đồ không thay đổi nhiều về cấu trúc sau khi chuẩn hóa?**

* **Giữ nguyên mối tương quan giữa các đặc trưng**: Khi chuẩn hóa, chỉ có **phạm vi** giá trị của các đặc trưng thay đổi, còn **cấu trúc** của dữ liệu và mối tương quan giữa các đặc trưng vẫn được giữ nguyên. Đó là lý do tại sao các điểm trên sơ đồ vẫn giữ nguyên hình dạng phân tán sau khi chuẩn hóa.
  + Ví dụ: Mối tương quan dương giữa nhiệt độ tối đa và tối thiểu được giữ nguyên vì chuẩn hóa không thay đổi tương quan giữa các đặc trưng, chỉ thay đổi kích thước giá trị.
* **Chuẩn hóa không làm thay đổi phân bố**: Chuẩn hóa chỉ đơn giản là điều chỉnh các giá trị về cùng một thang đo, không ảnh hưởng đến phân bố của các đặc trưng. Ví dụ, trong biểu đồ "Nhiệt độ Tối Đa vs Nhiệt độ Tối Thiểu," các điểm phân tán vẫn theo một đường chéo sau khi chuẩn hóa, điều này chứng tỏ rằng mối tương quan dương giữa hai biến này không bị thay đổi bởi quá trình chuẩn hóa.

A black screen with white text

Description automatically generated

Trong bước thử nghiệm nhiều cấu hình khác nhau của lớp ẩn, bạn đã huấn luyện mạng nơ-ron với các cấu hình lớp ẩn sau:

* (10,)
* (20,)
* (30, 15)
* (40, 20, 10)

Điểm đánh giá cho mỗi cấu hình được tính bằng phương pháp **Cross-validation** (xác thực chéo), tức là dữ liệu được chia thành nhiều tập nhỏ, và mỗi tập sẽ lần lượt được sử dụng để kiểm tra, trong khi các tập còn lại sẽ được dùng để huấn luyện.

Kết quả:

* Cấu hình (10,) có điểm Cross-validation là **0.7701**
* Cấu hình (20,) có điểm Cross-validation là **0.7681**
* Cấu hình (30, 15) đạt **0.8043**, cao nhất trong tất cả các cấu hình.
* Cấu hình (40, 20, 10) đạt **0.8004**, gần bằng với (30, 15), nhưng vẫn thấp hơn.

**Giải thích**:

* Khi cấu hình lớp ẩn có quá ít nơ-ron, mô hình có thể không đủ khả năng học được các mẫu phức tạp, dẫn đến hiệu suất thấp (như trường hợp (10,) và (20,)).
* Khi cấu hình lớp ẩn có nhiều nơ-ron hơn, mô hình có khả năng học các mẫu tốt hơn, nhưng có thể dẫn đến tình trạng **overfitting** (quá khớp) nếu có quá nhiều nơ-ron, làm cho mô hình học quá kỹ các mẫu huấn luyện mà không tổng quát hóa tốt cho dữ liệu mới. Điều này có thể lý giải tại sao (30, 15) có hiệu suất tốt hơn so với (40, 20, 10).

Như vậy, cấu hình **(30, 15)** được chọn là cấu hình tốt nhất do nó có khả năng cân bằng giữa việc học đủ các mẫu phức tạp và không bị quá khớp.