Aufgabe 10:

Die empirische Varianz ist ein Maß für die Streuung der Werte in einer Stichprobe. Sie gibt an, wie weit die einzelnen Werte vom Mittelwert entfernt sind.

Die Formel für die empirische Varianz lautet:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

- x_i: Einzelne Werte der Stichprobe
- $\bar{x}$: Mittelwert der Stichprobe
- n: Anzahl der Werte in der Stichprobe

Nach Carl Friedrich Gauß wird anstelle der Beträge der Abweichungen vom Mittelwert ($|x_i - \bar{x}|$) die Quadrate dieser Abweichungen ($(x_i - \bar{x})^2$) verwendet. Dies hat zwei Vorteile:

1. **Negative Abweichungen heben sich nicht auf**: Die Quadrate sorgen dafür, dass alle Abweichungen positiv sind.
2. **Einfachere mathematische Behandlung**: Mit Quadraten lassen sich mathematische Ableitungen und Optimierungen leichter durchführen.

Die **empirische Varianz** basiert auf dem Durchschnitt dieser quadrierten Abweichungen. Jedoch wird dabei nicht durch n (die Anzahl der Stichprobenwerte) geteilt, sondern durch n−1. Der Grund:

- Bei einer Stichprobe von n Werten gibt es nur n−1 "Zwischenräume", die die gegenseitige Lage der Punkte charakterisieren.
- Diese Korrektur, bekannt als **Bessel-Korrektur**, macht die Varianz zu einer unverzerrten Schätzung der Streuung der Grundgesamtheit.

Die Herleitung aus der Grundformel:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$(x_i - \bar{x})^2 = x_i^2 - 2x_i\bar{x} + \bar{x}^2$$

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum x_i^2 - 2\bar{x}\sum x_i + n\bar{x}^2$$

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

Aufgabe 11:

Datenvorverarbeitung ist ein essenzieller Schritt in der Datenanalyse und im maschinellen Lernen. Sie beinhaltet die Transformation und Bereinigung von Rohdaten, um sie für die Analyse oder Modellierung geeignet zu machen. Ziel ist es, die Qualität der Daten zu verbessern und sie in ein Format zu bringen, das von Algorithmen effizient genutzt werden kann.

Drei Methoden zum Umgang mit fehlenden Werten

Fehlende Werte entfernen:

Zeilen oder Spalten mit fehlenden Werten werden aus dem Datensatz gelöscht. Diese Methode eignet sich, wenn die Menge der fehlenden Daten gering ist, sodass der Informationsverlust minimal bleibt. Beispiel: Ein Datensatz mit 1000 Einträgen hat 5 Zeilen mit fehlenden Werten. Löschen dieser Zeilen beeinträchtigt die Analyse kaum.

Fehlende Werte auffüllen:

Fehlende Werte werden durch plausible Werte ersetzt, z. B. durch den Mittelwert, Median oder einen speziellen Wert wie 0. Diese Methode wird verwendet, wenn der Verlust von Daten durch Löschen vermieden werden soll. Beispiel: In einem Datensatz mit Einkommen wird ein fehlender Wert durch den Durchschnitt des Einkommens ersetzt.

Vorhersage der fehlenden Werte:

Fehlende Werte werden durch maschinelle Lernmodelle vorhergesagt, basierend auf den übrigen Daten. Diese Methode wird verwendet, wenn die Datenstruktur komplex ist und andere Merkmale (Features) Informationen über die fehlenden Werte liefern können. Beispiel: Fehlende Temperaturwerte in einer Wetterstation könnten durch ein Modell vorhergesagt werden, das andere Wetterparameter wie Luftdruck oder Luftfeuchtigkeit berücksichtigt.