

Wissenschaftliche Kriterien von Explainability – Was macht ein Modell transparent?		
Simulierbarkeit	Zerlegbarkeit	Algorithmische Transparenz
Das Bayes'sche Modell ist für die Menschen leicht zu verstehen und zu kopieren. <i>Bei kleineren Daten können Menschen bedingte Wahrscheinlichkeiten von Klassen und dessen Ausprägungen errechnen und Vorhersagen treffen.</i>	Die Aufschlüsselung des Modells in verständliche Teile, wie Eingaben und Berechnungen, hilft bei der Erklärung seiner Funktionsweise. <i>Der Datensatz kann in seine einzelnen Merkmale zerlegt werden. Wahrscheinlichkeiten können für einzelne und mehrere Merkmale berechnet werden.</i>	Das Modell zeigt uns genau, wie es Entscheidungen trifft, so dass der Prozess nachvollziehbar ist. <i>Die fit()-Funktion von Scitkitlearn verhindert Einblicke in den Trainingsprozess, allerdings können Wahrscheinlichkeiten abgefragt werden.</i>

Vorgehen & Accuracy	
<i>Es wurden Methoden angewendet, um die Accuracy zu verbessern.</i>	
Nur Gaussian Naive	79,84%
Reduziert auf die wichtigsten Merkmale	79,97%
Mit Hyperparametertuning	82,72%
Mit Principal Component Analysis	80,69%
Der Datensatz wurde auf die wichtigsten Merkmale reduziert.	

Bayes Theorem & Bayes Classifier

Das Bayes Theorem basiert auf dem Konzept der bedingten Wahrscheinlichkeiten.

Posterior

Likelihood

Prior Class

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Prior Feature

Der Naive Bayes Classifier verwendet dieses Theorem, um im Trainingsprozess die Wahrscheinlichkeit für die Zugehörigkeit zu einer Klasse A für eines bestimmtes Merkmales B zu bestimmen. Eine Prämisse hierfür ist die Unabhängigkeit der Merkmale.

Das Theorem im Trainingsprozess:

- Wahrscheinlichkeit der Klasse P(A)* – **Prior Class**
- Inverse bedingte Wahrscheinlichkeit P(B|A)* – **Likelihood**
- Wahrscheinlichkeit des Merkmales P(B)* – **Prior Feature**
- Berechnung der gesuchte Wahrscheinlichkeit P(A|B) – **Posterior**

*anfängliche Wahrscheinlichkeit (Historische Daten, Annahmen)

Explainable Boosting Machine (EBM)
Explainable Boosting Machine (EBM) ist ein baumbasiertes, zyklisches Gradient Boosting Generalized Additive Model. EBM ist ein Glassbox-Modell mit hoher Genauigkeit, das mit den modernsten Methoden des maschinellen Lernens wie Random Forest vergleichbar ist, und gleichzeitig hochgradig intelligent und erklärbar ist.
<i>Als erklärbares Modell wird dieses zum Vergleich herangezogen.</i>

Bayes vs. EBM	
$P(A B) = \frac{P(B A) \cdot P(A)}{P(B)}$	$g(E[y] A B) = \frac{P(B A) \cdot P(A)}{P(B)}$
Accuracy bei Dataset: 80%	Accuracy bei Dataset: 87%
Verteilung der Daten kann mit sns.countplot analysiert werden, für jede einzelne Variable	Funktion Class Histogram: Analyse der Verteilung der Daten
Marginal Plot mit seaborn Funktion und Berechnung des Pearson-Korrelationskoeffizienten, für jede einzelne Variable	Funktion Marginal: Verteilung der Datenmerkmale, Berechnung des Pearson-Korrelationskoeffizienten

Bayes vs. EBM	
Feature Importance kann mit plt gezeigt werden, Beitrag, den jedes Merkmal zu den Vorhersagen leistet, nicht	Explain.global(): zeigt Feature Importance + Beitrag, den jedes Merkmal zu den Vorhersagen leistet
Keine ähnliche Funktion	Explain.local(): zeigt zeigt die Aufschlüsselung, wie viel jeder Begriff zur Vorhersage für eine einzelne Stichprobe beiträgt
Da EBM speziell auf die Interpretierbarkeit ausgelegt ist, liefert es viele Funktionen, mit welchen man sowohl das Dataset als auch die Ergebnisse des trainierten Modells betrachten kann. Auch beim Gaussian Naive Bayes kann man viele Funktionen anwenden, die die Explainability unterstützen, jedoch werden diese Funktionen von anderen Bibliotheken bereitgestellt und man muss sie oft für jede Variable einzeln ausführen.	

Explainability

Der Directed Acyclic Graph (DAG) zeigt die mittleren Werte der Merkmale in Bezug auf das Einkommen und bietet verständliche Einblicke in die Einflussfaktoren auf das Zielattribut. Diese Graphik unterstützt das erklärbare Maschinenlernen, indem sie wichtige Merkmale hervorhebt und deren Auswirkungen auf die Einkommensvorhersage veranschaulicht. Durch die Analyse von Daten und das Theorem ist bereits ein gewisser Grad an Explainability gegeben. Mithilfe von Bibliotheken können einige Outputs gemäß des EBM nachgebaut werden, was den Grad weiter erhöht.

Fazit
Der Naive Bayes Classifier ist für seine Einfachheit und leichte Implementierung bekannt, bietet jedoch nicht dasselbe Maß an Interpretierbarkeit wie einige andere Modelle, insbesondere solche, die ausdrücklich auf Interpretierbarkeit ausgelegt sind.
Die Nachstellbarkeit gewisser Funktionen führt zu einem hohen Grad an Explainability. Dennoch lassen sich ausgewählte Aspekte der Vorhersagen nicht vollständig erklären.