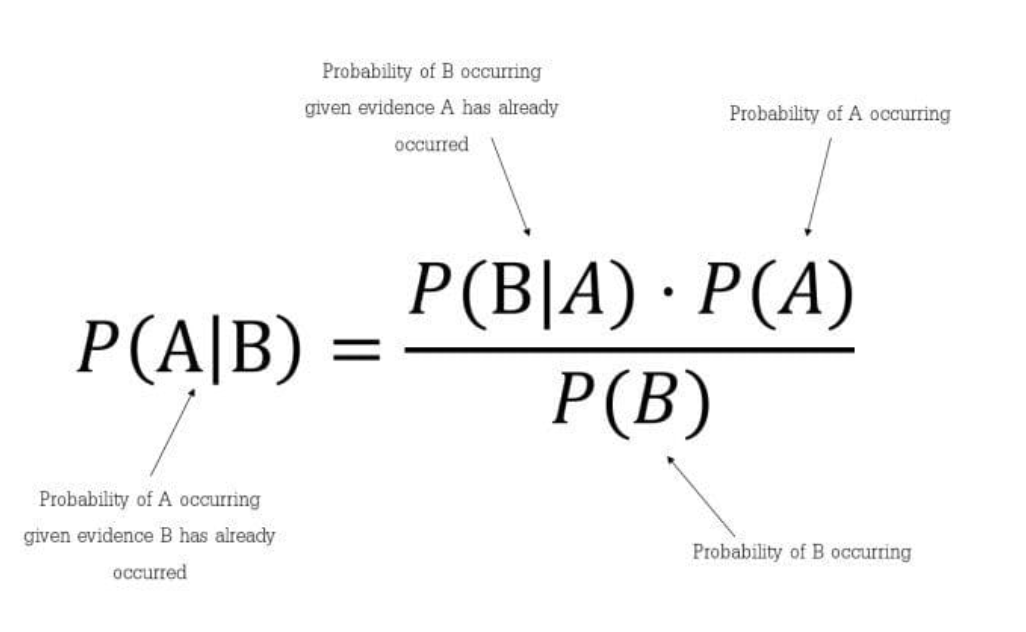
# **NAIVE BAYES**

**What is the Naive Bayes Algorithm?**

The Naive Bayes algorithm is a classification/categorization algorithm named after the mathematician Thomas Bayes. The foundation of the Naive Bayes classification method is based on Bayes' Theorem. For this reason, the algorithm is a probability-based classification method. The algorithm calculates all the probabilities for a given data and classifies other data according to the result with the highest probability. Naive Bayes is a type of lazy learning.

metin, elektronik yön işaretleri, yön işaretleri, çizgi içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.



(Nagesh Singh Chauhan, Ece Akdağlı)

P ( A | B ) = The probability of event A occurring given that event B has occurred  
P ( A ) = The probability of event A occurring  
P ( B | A ) = The probability of event B occurring given that event A has occurred  
P ( B ) = The probability of event B occurring

In Naive Bayes classification, a certain amount of labeled data is presented to the system beforehand, and the model uses this data to perform probability calculations. The data used in the training process must belong to specific categories. Based on the probabilities it learns, the model predicts which class new incoming data belongs to. As the amount of data used for training increases, the accuracy of the model also improves. In this method, what matters is not the content of the data, but how it is categorized based on certain features. Regardless of whether the data is numerical or textual, the critical point for the model is to correctly analyze the proportional relationships between the data.

(KodEdu)  
Email filtering is a good example of this. A model that detects spam analyzes emails previously marked as spam and calculates the probability of certain words being spam-related. For example, if words like "reward" or "bonus" appear more frequently in spam emails, and these words appear in a new email, the model predicts that the email might be spam.

**Quantitative Set**  
The features that distinguish spam emails from normal emails create a quantitative set. For example, since words like "bonus" and "reward" appear more frequently in spam emails, these words are considered parameters. The model assigns a higher spam probability to emails containing these types of words and classifies them accordingly. Mathematically:



 ∈ RL forms a quantity vector of dimension L.

Given x ∈ RL and S as the set of classes to be separated, according to Bayes' theorem, the following expression is written:



And

siyah, karanlık içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

* P(Si); The prior probability of Si, where i = 1, 2, ..., L
* P(Si|x); The posterior probability of Si given x
* p(x); The probability density function (PDF) of x
* p(x|Si); The conditional PDF of x given Si, where i = 1, 2, ..., L

**Bayes Decision Theorem**



It is a quantity vector of dimension L for a pattern whose class is unknown. Taking the example of spam emails, a new email whose spam status is unknown is a pattern with an unknown class.  
Again, the class Si to which x will be assigned is:  
According to Bayes' decision theory, x belongs to class Si





**Naive Bayes Classification**

In order to decide which class a given x belongs to, the Bayes decision theorem formulated above is used. If the assumption of statistical independence is utilized in this classification, it is called Naive Bayes classification. Mathematically, it can be expressed as:



siyah, karanlık içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

Thus, the Bayes decision theorem takes the following form. According to Bayes' decision theory, x belongs to class Si:

siyah, karanlık içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

P(Si) and P(Sj) are the prior probabilities of the i and j classes. These values can be easily calculated from the available dataset.

Although the application area of the Naive Bayes classifier seems limited, successful results can be obtained by relaxing the condition of statistical independence of the components of x (the quantitative set) in high-dimensional space and with a sufficient amount of data.

(Wikipedia)

The mathematical operations provided for the Naive Bayes algorithm are quite comprehensive. Beyond the mathematical foundations, there are also application and evaluation stages.

**Feature Selection and Preprocessing:**

* **Data Cleaning:** The Naive Bayes algorithm may perform poorly, especially when dealing with missing data or highly noisy data. Therefore, a data preprocessing process should be applied.
* **Feature Selection:** The selection of features and dimensionality reduction can affect the performance of Naive Bayes. Especially when there are many features, dimensionality reduction (e.g., PCA – Principal Component Analysis) can be useful.

(Kotsiantis, S. B. , Iglewicz, B., & Hoaglin, D. C.)

**Model Selection:**

There are different types of Naive Bayes, and depending on the dataset, different types may be more effective:

* **Gaussian Naive Bayes:** It is used especially when working with numerical data and assumes that the features follow a normal distribution.
* **Multinomial Naive Bayes:** It is used in cases where numerical data is related to counts and frequencies, such as in text classification.
* **Bernoulli Naive Bayes:** It is used particularly with binary data. It works on the presence or absence of features (e.g., whether certain words appear in text).

If the data is numerical and follows a normal distribution with continuous values, Gaussian Naive Bayes should be used.  
If the data involves counts, such as text data, Multinomial Naive Bayes is suitable. It is more commonly used in applications like text classification and sentiment analysis.  
For data with binary (0/1) features (such as yes/no, sick/not sick, etc.), Bernoulli Naive Bayes should be used.

(McCallum, A., & Nigam, K., Rish, I.)

**Application Areas:**

* Text classification
* Sentiment analysis
* Medical diagnosis
* Marketing
* Bioinformatics

**Performance Calculation**

The model assigns a category to the observations in the test dataset by comparing them with the observations in the training dataset. Since we know the class of the observations in the test dataset, the model's performance can be evaluated.  
The most commonly used method for performance evaluation is average accuracy.

metin, yazı tipi, beyaz, çizgi içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

* **TP (True Positive):** Correctly classified as positive.
* **TN (True Negative):** Correctly classified as negative.
* **FP (False Positive):** Incorrectly predicted as positive.
* **FN (False Negative):** Incorrectly predicted as negative.

**Advantages of the Naive Bayes Algorithm:**

* Simple, fast, and provides high performance with low data requirements.
* Successful in many application areas, such as text classification and spam filtering.
* Resistant to dataset imbalance and noisy data.

**Disadvantages of the Naive Bayes Algorithm:**

* Provides less flexibility due to having fewer hyperparameters.
* The assumption of independence between features is not always valid in the real world.
* Often fails to capture very complex structures, making it unsuitable for problems that require high precision, leading to lower performance.

(BTK Academy)

In the example study we conducted, we generated a random dataset with 1000 samples, consisting of 2 independent and 1 dependent variable, using the make\_classification function from the sklearn library in Python. We then performed classification using the Gaussian Naive Bayes algorithm (since we were working with numerical data, we used Gaussian Naive Bayes). The data was split into 80% training and 20% testing, and the algorithm learned the classification logic from the training data using its mathematical theory and tested it on the test data. From this example, we can clearly see how the Naive Bayes algorithm works.

ekran görüntüsü, renklilik içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

In this example, there are two groups, blue and red, with variables that range between -2 and 4. In the graph, correctly predicted groups are shown as circles with their respective colors, while incorrect predictions are shown as crosses.  
In the other visual, we can see the accuracy of our predictions using a confusion matrix.  
The accuracy of this example has been tested using classification model evaluation metrics, achieving a success rate of %91.

| **Class** | **Precision** | **Recall** | **F1-Score** | **Support** |
| --- | --- | --- | --- | --- |
| **Class 0** | 0.88 | 0.94 | 0.91 | 104 |
| **Class 1** | 0.93 | 0.86 | 0.90 | 96 |
| **Accuracy** |  |  | **0.91** | **200** |
| **Macro Avg** | 0.91 | 0.90 | 0.90 | 200 |
| **Weighted Avg** | 0.91 | 0.91 | 0.90 | 200 |

**Accuracy Score: 0.91**

# **References**

* Nagesh Singh Chauhan, KDnuggets on April 8, 2022 in Machine Learning
* <https://ece-akdagli.medium.com/makine-öğrenmesinde-naive-bayes-yöntemi-98d28cfe46de>
* <https://tr.wikipedia.org/wiki/Naive_Bayes_sınıflandırıcısı>
* <https://kodedu.com/2014/05/naive-bayes-siniflandirma-algoritmasi/>
* Iglewicz, B., & Hoaglin, D. C. (1993). *How to detect and handle outliers.* Sage Publications.
* Kotsiantis, S. B. (2007). "Supervised machine learning: A review of classification techniques," *In Proceedings of the 5th international conference on information & communication technologies: from theory to applications (ICTTA)*, 1-6.
* McCallum, A., & Nigam, K. (1998). "A comparison of event models for Naive Bayes text classification," *In Proceedings of the 15th international conference on machine learning (ICML-98)*, 41-48.
* Rish, I. (2001). "An empirical study of the Naive Bayes classifier," *In IJCAI-01 workshop on empirical methods in artificial intelligence*, 41-46.
* BTK Academy, Machine Learning Courses.

# NAİVE BAYES ALGORİTMASI

## **Naive bayes Algoritması Nedir?**

Adını Matematikçi Thomas Bayes'den alan bir sınıflandırma/ kategorilendirme algoritmasıdır. Naive Bayes sınıflandırma yönteminin temeli Naive Bayes teoremine dayanmaktadır. Bu sebeple, algoritma ,olasılığa dayalı bir sınıflandırma yöntemidir. Algoritma, bir veri için tüm olasılıkları hesaplar ve olasılık değeri en yüksek çıkan sonuca göre diğer verilerin sınıflandırmasını yapar. Naive Bayes algoritması tembel bir öğrenme türüdür.

metin, elektronik yön işaretleri, yön işaretleri, çizgi içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

metin, yazı tipi, diyagram, ekran görüntüsü içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

(Nagesh Singh Chauhan, Ece Akdağlı)

*P ( A | B )* = B olayı gerçekleştiğinde A olayının gerçekleşme olasılığı  
*P ( A )*= A olayının gerçekleşme olasılığı  
*P ( B | A )*= A olayı gerçekleştiğinde B olayının gerçekleşme olasılığı  
*P ( B )* = B olayının gerçekleşme olasılığı

Naïve Bayes sınıflandırmasında, sisteme önceden belirli miktarda etiketlenmiş veri sunulur ve model bu verileri kullanarak olasılık hesaplamaları yapar. Eğitim sürecinde kullanılan veriler belirli kategorilere ait olmalıdır. Model, öğrendiği bu olasılıklara dayanarak yeni gelen verilerin hangi sınıfa ait olduğunu tahmin eder. Eğitim için kullanılan veri sayısı arttıkça, modelin doğruluğu da artar. Bu yöntemde önemli olan, verinin içeriğinden çok, hangi özelliklere göre ayrıştırıldığıdır. Sayısal veya metinsel fark etmeksizin, model için kritik nokta, veriler arasındaki oransal ilişkileri doğru analiz etmektir.

(KodEdu)

E-posta filtreleme buna iyi bir örnektir. Spam tespiti yapan bir model, geçmişte spam olarak işaretlenmiş e-postaları analiz ederek, belirli kelimelerin spam olma olasılığını hesaplar. Örneğin, "ödül", "ikramiye" gibi kelimelerin spam e-postalarda daha sık geçtiği belirlenirse, yeni bir e-postada bu kelimeler bulunuyorsa, model onun spam olabileceğini tahmin eder.

**Nicelik Kümesi**

Spam e-postalar ile normal e-postalar arasındaki farkı belirleyen özellikler, bir nicelik kümesi oluşturur. Örneğin, "ikramiye", "ödül" gibi kelimeler spam e-postalarda daha sık yer aldığı için bu kelimeler birer parametre olarak kabul edilir. Model, bu tür kelimelerin geçtiği e-postalara daha yüksek spam olasılığı atayarak sınıflandırma yapar. Matematiksel olarak;



 ∈ RL L-boyutlu nicelik vektörünü oluşturur.  
x ∈ RL verildiğine göre ve *S* ayrıştırılacak sınıflar kümesiyse, Bayes teoremine göre aşağıdaki ifade yazılır.



Ve

siyah, karanlık içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

* P(Si); Si'nin öncel olasılığı i = 1, 2, . . ., L,
* P(Si|x); Si'nin ardıl olasılığı
* p(x); x in Olasılık yoğunluk fonksiyonu (oyf)
* p(x|Si); i =1 = 2, . . ., L, x'in koşullu oyf'si

## **Bayes karar teoremi**



sınıfı belli olmayan örüntünün L-boyutlu nicelik vektörüdür. Spam e-posta örneğinden gidecek olursak spam olup olmadığını bilmediğimiz yeni bir e-posta sınıfı belli olmayan örüntüdür.  
Yine Si x'in atanacağı sınıf ise;  
Bayes karar teorisine göre x sınıf Si'ya aittir.





**Naive Bayes sınıflandırması**

Verilen bir x'in hangi sınıfa ait olup olmadığına karar vermek için kullanılan yukarıda formüle edilen Bayes karar teoreminde istatistik olarak bağımsızlık önermesinden yararlanılırsa bu tip sınıflandırmaya Naive bayes sınıflandırılması denir. Matematiksel bir ifadeyle



siyah, karanlık içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

böylece Bayes karar teoremi aşagıdaki şekli alır. Bayes karar teorisine göre x sınıf Si'ya aittir.

siyah, karanlık içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

İ ve j sınıflarının öncel olasılıklarıdır. Elde olan veri kümesinden değerleri kolayca hesaplanabilir.

Naive bayes sınıflandırıcının kullanım alanı her ne kadar kısıtlı gözükse de yüksek boyutlu uzayda ve yeterli sayıda veriyle x'in (nicelik kümesi) bileşenlerinin istatistik olarak bağımsız olması koşulu esnetilerek başarılı sonuçlar elde edilebilinir.

Naive Bayes algoritması hakkında verdiğiniz matematiksel işlemler oldukça kapsamlı. Matematiksel temeller dışında, uygulama ve değerlendirme aşamaları da vardır.

**Özellik Seçimi ve Ön İşleme:**

* **Veri Temizleme:** Naive Bayes algoritması, özellikle eksik verilerle ya da yüksek oranda gürültülü verilerle çalışırken zayıf performans gösterebilir. Bu nedenle, veri ön işleme süreci uygulanmalıdır.
* **Özellik Seçimi:** Özelliklerin seçimi ve boyut indirgeme, Naive Bayes’in performansını etkileyebilir. Özellikle çok sayıda özellik olduğunda, boyut indirgeme (örneğin, PCA – Principal Component Analysis) kullanmak faydalı olabilir.

(Kotsiantis, S. B. , Iglewicz, B., & Hoaglin, D. C.)

**Model Seçimi:**

Naive Bayes'in farklı türleri vardır ve veri setine göre farklı türler daha etkili olabilir:

* **Gaussian Naive Bayes:** Özellikle sayısal verilerle çalışırken kullanılır ve özelliklerin normal dağılıma yakın olduğunu varsayar.
* **Multinomial Naive Bayes:** Metin sınıflandırma gibi, sayısal verilerin sayım ve frekansla ilgili olduğu durumlarda kullanılır.
* **Bernoulli Naive Bayes:** Özellikle ikili (binary) verilerle çalışırken kullanılır. Özelliklerin varlığı ya da yokluğu üzerinde çalışır (örneğin, metinlerde kelimelerin olup olmaması gibi).

Sayısal veriler varsa ve normal dağılıma giderken veriler sürekli ise gaussian naive bayes kullanılmalıdır.

Sayım verileri metin verisi gibi verilerde ise multinominal naive bayes uygundur. Daha çok metin sınıflandırma duygu analizi gibi uygulamalarda kullanılır.

İkili (binary) özelliklere sahip verilerde (0/1 gibi var veya yok, hasta/ hasta değil vb.) bernoulli naive bayes kullanılmalıdır.

(McCallum, A., & Nigam, K., Rish, I.)

## **Uygulama Alanları**

* Metin sınıflandırma
* Duygu analizi
* Tıbbi teşhis
* Pazarlama
* Biyoinformatik

## **Performans hesaplama**

Test veri kümesindeki gözlemleri eğitim veri kümesindeki gözlemlerle karşılaştırarak bunlara bir kategori atar. Test veri kümesindeki gözlemlerin sınıfını bildiğimiz için modelin performansı değerlendirilebilir.

Performans değerlendirmede kullanılan en yaygın yöntem ortalama doğruluk olarak kullanılabilir.

metin, yazı tipi, beyaz, çizgi içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

* **TP (True Positive)**: Doğru şekilde pozitif olarak sınıflandırılanlar.
* **TN (True Negative)**: Doğru şekilde negatif olarak sınıflandırılanlar.
* **FP (False Positive)**: Yanlış pozitif tahmin edilenler.
* **FN (False Negative)**: Yanlış negatif tahmin edilenler.

## **Naive Bayes Algoritmasının Avantajları**

* Basit, hızlı ve düşük veri gereksinimiyle yüksek performans verir.
* Metin sınıflandırma ve spam filtreleme gibi birçok uygulama alanlarında başarılı.
* Veri seti dengesizliğine ve gürültülü veriye karşı dayanıklı.

## **Naive Bayes Algoritmasının Dezavantajları**

* Daha az hiperparametreye sahip olması sebebiyle daha az esneklik sağlar.
* Öznitelikler arası bağımsızlık varsayımı gerçek dünyada her zaman geçerli değildir.
* Genellikle çok karmaşık yapıları yakalayamaz, yüksek hassasiyet gerektiren problemlerde tercih edilemez, düşük performans gösterir

(BTK Akademi)

Yaptığımız örnek çalışmada Python üzerinden sklearn kütüphanesini kullanarak make\_classification fonksiyonu ile 2 bağımsız 1 bağımlı değişkenli 1000 örneklemli rastgele bir veri oluşturup bu veri üzerinden Gaussian Naive Bayes algoritması kullanarak (sayısal verilerle işlem yaptığımız için Gaussian Naive Bayes Kullandık) sınıflandırma işlemi yapılmıştır. Verinin yüzde 80 train yüzde 20 test verisi olmak üzere ayırıp train verisinden algoritmanın matematiksel teoremini kullanarak sınıflandırma mantığını öğrenip test verisinde bunları denemiştir. Bu örneğe bakıldığında Naive Bayes algoritmasının çalışma mantığını daha açık şekilde görmekteyiz.

ekran görüntüsü, renklilik içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

Burada mavi ve kırmızı iki grup ve -2 ile 4 arasında değerler alan değişkenlerimiz bulunmakta grafikte doğru tahmin yapılan gruplar renkleriyle yuvarlak ile gösterilirken yanlışlar çarpı ile gösterilmektedir.

Diğer görselde ise karmaşıklık matrisiyle tahminlerimizin doğruluğunu görmekteyiz.

Bu örneğin doğruluğu sınıflandırma modellerinin değerlendirme metrikleri yöntemleriyle test edilmiştir. Ve %91’ik bir başarı sağlanmıştır.

| **Sınıf** | **Precision** | **Recall** | **F1-Score** | **Support** |
| --- | --- | --- | --- | --- |
| **Sınıf 0** | 0.88 | 0.94 | 0.91 | 104 |
| **Sınıf 1** | 0.93 | 0.86 | 0.90 | 96 |
| **Accuracy** |  |  | **0.91** | **200** |
| **Macro Avg** | 0.91 | 0.90 | 0.90 | 200 |
| **Weighted Avg** | 0.91 | 0.91 | 0.90 | 200 |

**Doğruluk Oranı: 0.91**

Kaynakça

* Nagesh Singh Chauhan, KDnuggets on April 8, 2022 in Machine Learning
* <https://ece-akdagli.medium.com/makine-öğrenmesinde-naive-bayes-yöntemi-98d28cfe46de>
* <https://tr.wikipedia.org/wiki/Naive_Bayes_sınıflandırıcısı>
* <https://kodedu.com/2014/05/naive-bayes-siniflandirma-algoritmasi/>
* Iglewicz, B., & Hoaglin, D. C. (1993). *How to detect and handle outliers.* Sage Publications.
* Kotsiantis, S. B. (2007). "Supervised machine learning: A review of classification techniques," *In Proceedings of the 5th international conference on information & communication technologies: from theory to applications (ICTTA)*, 1-6.
* McCallum, A., & Nigam, K. (1998). "A comparison of event models for Naive Bayes text classification," *In Proceedings of the 15th international conference on machine learning (ICML-98)*, 41-48.
* Rish, I. (2001). "An empirical study of the Naive Bayes classifier," *In IJCAI-01 workshop on empirical methods in artificial intelligence*, 41-46.
* BTK Akademi, Makine Öğrenmesi Kursu