



GENDER PREDICTION USING NAÏVE BAYES

AGENDA

Objective

Methodology

Tools Used

Prediction & Accuracy



OBJECTIVE

We are given names of people and have to predict their gender solely based on their names only using Excel and the Naïve Bayes Theorem. Now, How shall we go about it?



METHODOLOGY

The way to go about this problem is:

Extract the **last letter** of each name from the training data

Check how frequently a given letter occurs, corresponding to the gender

For e.g:

The letters '**a**' and '**i**' most frequently occur when the person is a **Female**

Use this information to predict the **Gender** based on the Names given in the test data

6	Kranti	Female	i				
7	Tulika	Female	a				
8	Aarushi	Female	i				
9	Abhicandr	Male	a				
10	Pratigya	Female	a				
11	Devak	Male	k				
12	Kashiprasa	Male	d				
13	Madhavi	Female	i				
14	Charusila	Female	a				
15	Chithayu	Male	u				
16	Manmayi	Female	i				
17	Mahajabe	Female	n				
18	Krishnakur	Male	r				
19	Kailas	Male	s				
20	Nidhyathi	Female	i				
21	Nainika	Female	a				

TOOLS USED

RIGHT- Returns the specified number of characters at the end of a string, in this case, it returns the last character.

PIVOT TABLE- To easily arrange and summarize data and to generate insights.

VLOOKUP- To look for the gender value based on the last letter from the pivot table(created on test data) and use that data to predict the Gender based on the new names in the test data

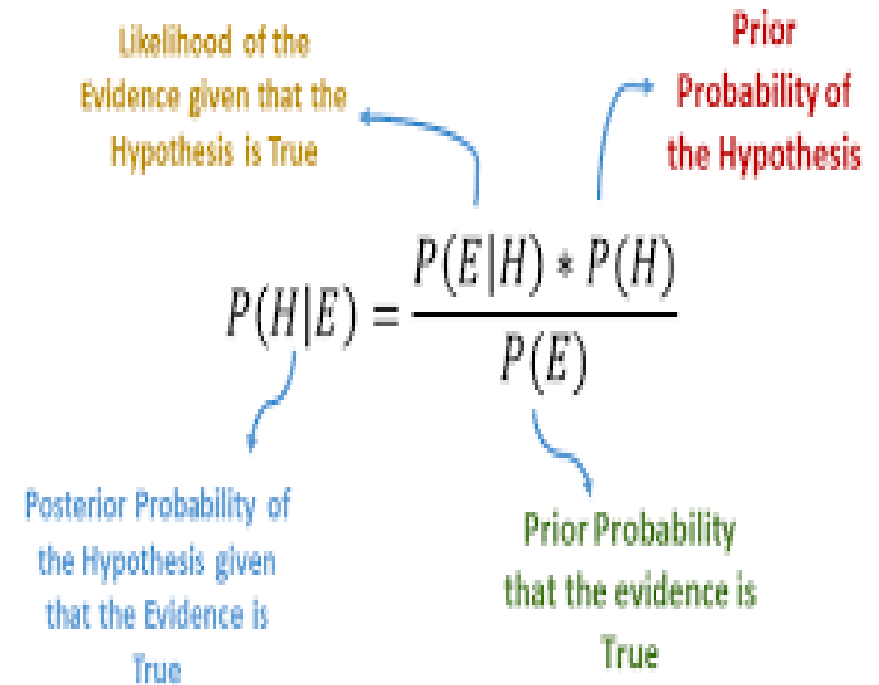
IF- To compare the probabilities of the person being male or female and generate a result based on the higher probability

COUNTIF- Counts the values only if they meet a certain condition



NAÏVE BAYES THEOREM

- Here, we need to obtain the probability that a person is a Male or Female based on the last letter($P(\text{gender}/\text{letter})$)
- So we use the Naïve Bayes formula and find the relative probability
- $P(\text{gender}/\text{letter}) = P(\text{letter}/\text{gender}) * P(\text{gender})$
- We can obtain this by creating a pivot table and setting rows as the actual gender, columns as the predicted gender, and values as the count of gender. The values are set as percentage of column total



The diagram illustrates the Naïve Bayes formula with labels for each component:

$$P(H|E) = \frac{P(E|H) * P(H)}{P(E)}$$

Labels and arrows:

- Likelihood of the Evidence given that the Hypothesis is True** (orange text) points to $P(E|H)$.
- Prior Probability of the Hypothesis** (red text) points to $P(H)$.
- Posterior Probability of the Hypothesis given that the Evidence is True** (blue text) points to $P(H|E)$.
- Prior Probability that the evidence is True** (green text) points to $P(E)$.

DATASET

The dataset consists of 2100 rows of data:

- A pivot table was created using the dataset and using VLOOKUP on that table a Gender Prediction was made
- The accuracy of these predictions was found using the Confusion Matrix

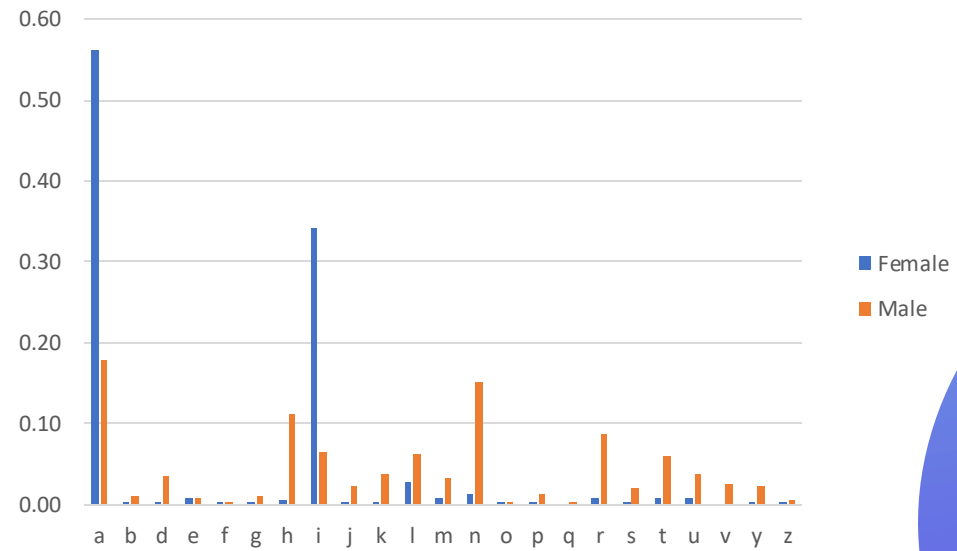


PIVOT TABLE

Row Labels	Female	Male	Grand Total
a	0.56	0.18	0.36
b	0.00	0.01	0.01
d	0.00	0.03	0.02
e	0.01	0.01	0.01
f	0.00	0.00	0.00
g	0.00	0.01	0.01
h	0.01	0.11	0.06
i	0.34	0.06	0.19
j	0.00	0.02	0.01
k	0.00	0.04	0.02
l	0.03	0.06	0.05
m	0.01	0.03	0.02
n	0.01	0.15	0.09
o	0.00	0.00	0.00
p	0.00	0.01	0.01
q	0.00	0.00	0.00
r	0.01	0.09	0.05
s	0.00	0.02	0.01
t	0.01	0.06	0.03
u	0.01	0.04	0.02
v	0.00	0.02	0.01
y	0.00	0.02	0.01
z	0.00	0.01	0.00
Grand Total	1.00	1.00	1.00



VISUALIZATION



PREDICTIONS

Using the VLOOKUP function and using the pivot table as a reference the gender was predicted and using the IF function the probabilities were compared.

Gender	LastLetter	LL/MALE	LL/FEMALE	P(MALE)	P(FEMALE)	Prediction
Male	h	0.1126322	0.00502152	0.02012	0.002817	MALE
Female	a	0.1785937	0.56097561	0.0319	0.3146936	FEMALE
Male	b	0.0099564	0.00215208	0.00178	0.0012073	MALE
Female	a	0.1785937	0.56097561	0.0319	0.3146936	FEMALE
Female	i	0.0647169	0.34074605	0.01156	0.1911502	FEMALE
Female	a	0.1785937	0.56097561	0.0319	0.3146936	FEMALE
Female	i	0.0647169	0.34074605	0.01156	0.1911502	FEMALE
Male	a	0.1785937	0.56097561	0.0319	0.3146936	FEMALE
Female	a	0.1785937	0.56097561	0.0319	0.3146936	FEMALE
Male	k	0.0373367	0.00286944	0.00667	0.0016097	MALE
Male	d	0.0348475	0.00215208	0.00622	0.0012073	MALE
Female	i	0.0647169	0.34074605	0.01156	0.1911502	FEMALE
Female	a	0.1785937	0.56097561	0.0319	0.3146936	FEMALE
Male	u	0.0373367	0.00789096	0.00667	0.0044266	MALE
Female	i	0.0647169	0.34074605	0.01156	0.1911502	FEMALE
Female	n	0.1512134	0.01219512	0.02701	0.0068412	MALE
Male	r	0.0877411	0.00645624	0.01567	0.0036218	MALE
Male	s	0.0205352	0.00286944	0.00367	0.0016097	MALE
Female	i	0.0647169	0.34074605	0.01156	0.1911502	FEMALE
Female	a	0.1785937	0.56097561	0.0319	0.3146936	FEMALE
Female	i	0.0647169	0.34074605	0.01156	0.1911502	FEMALE
Male	l	0.0616055	0.02797704	0.011	0.0156944	FEMALE
Male	a	0.1785937	0.56097561	0.0319	0.3146936	FEMALE
Female	a	0.1785937	0.56097561	0.0319	0.3146936	FEMALE
Female	l	0.0616055	0.02797704	0.011	0.0156944	FEMALE
Female	a	0.1785937	0.56097561	0.0319	0.3146936	FEMALE
Female	i	0.0647169	0.34074605	0.01156	0.1911502	FEMALE
Male	m	0.0329807	0.00789096	0.00589	0.0044266	MALE



CHECKING PREDICTION ACCURACY

USING BASIC PROBABILITY

The gender predictions and the actual gender are compared and counted using the COUNTIF function and the total count of values is found using the COUNT function. The accuracy is predicted using simple division.

Number of Accurate Predictions	1693
Total Count	2100
Accuracy	80.62%

USING F1 SCORE

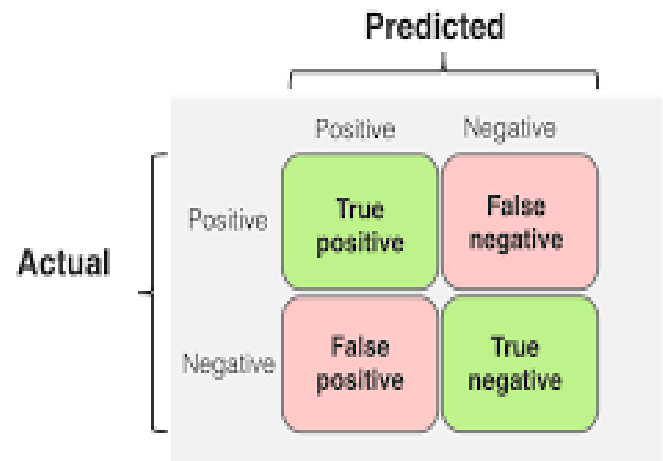
- The previous method is not ideal for predicting the model's accuracy as it is susceptible to the amount of data used and does not balance the accuracy based on false predictions.
- We need to establish a better method that balances the accuracy based on the amount of data used and the true and false predictions. That's where the **F1 Score** comes in.
- **TP**- True Positive, when its value is positive and predicted as positive
- **FP**- False Positive, when its value is positive and predicted as negative
- **TN**- True Negative, when its value is negative and predicted as negative
- **FN**- False Negative, when its value is negative and predicted as positive

USING F1 SCORE CONTINUED....

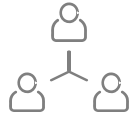
- **Precision(Positive class)**- The ratio of true positive values to our predicted true positive values($TP/(TP+FP)$)
- **Recall(Positive class)**- The ratio of true positive values to the actual number of true positive values($TP/(TP+FN)$)
- **Precision(Negative class)**- The ratio of true negative values to our predicted true negative values($TN/(TN+FP)$)
- **Recall(Negative class)**- The ratio of true negative values to the actual number of true negative values($TN/(TN+FN)$)
- **F1 Score**- The harmonic mean between precision and recall which establishes the accuracy of the model in a much better manner($2*Precision*Recall/(Precision+Recall)$)
- **Macro-averaged F1 Score**: Unweighted mean of F1 Scores of the positive and negative class
- **Weighted F1 Score**: The average F1 Score based on the weight of each category of data($F1 \text{ Score Female} * \text{count of Female} + F1 \text{ Score Male} * \text{count of Male}$)
- To depict the values we create a table called a **confusion matrix**

CONFUSION MATRIX AND F1 SCORE

To create the confusion matrix we use a pivot table on the validation set and using the values in the table we predict the precision, recall, and F1 Score respectively.



Count of Gender		Column Labels	
Row Labels		FEMALE	MALE
Female		923	64
Male		343	770
Grand Total		1266	834
Precision(female)		72.91%	
Recall(female)		93.52%	
F1 Score(female)		81.94%	
Precision(Male)		92.33%	
Recall(Male)		69.18%	
F1 Score(Male)		79.10%	
Macro-Averaged F1 Score		80.52%	
Support(Female)		0.470	
Support(Male)		0.530	
Weighted F1 Score		80.43%	



Thank You