Universidad de Talca
Facultad de Ingeniería
Departamento de Ciencias de la Computación

# Algoritmo Naive Bayes

Alumno: Miguel Arenas Santander
Profesor: Cesar Astudillo
Fecha: 29 de Septiembre de 2017.

# 1. Introducción

*En el presente trabajo de muestran los resultados de la implementación del algoritmo Naive Bayes. Este método es una técnica de clasificación y predicción que trabaja de forma supervisada la cual construye modelos que predicen la probabilidad de posibles resultados. Está basada en el teorema de Bayes que también es conocida como teorema de probabilidad condicionada, la cual constituye de una técnica supervisada porque necesita tener ejemplos clasificados para que funcione.*

# 2. Problema

*En la presente tarea el caso a analizar será el de predecir el inicio de la diabetes, para esto se utilizaron 768 observaciones medicas otorgadas por Vincent Sigillito de la National Institute of Diabetes and Digestive and Kidney Diseases.*
*Los registros contienen información de pacientes mujeres mayores de 21 años. Los atributos son de tipo numérico y sus unidades varían de atributo a atributo. Los nueve atributos son los siguientes:*

- Número de veces que ha estado embarazada.
- Concentración de glucosa a 2 horas en una prueba oral de tolerancia a la glucosa.
- Presión arterial diastólica.
- Espesor del pliegue cutáneo del tríceps.
- Insulina sérica de 2 horas.
- Índice de masa corporal.
- Función pedigree de la diabetes.
- Edad.
- Variable (0 ó 1).

*Cada registro tiene un valor que indica si el paciente tuvo una aparición de diabetes a 5 años de tomadas las mediciones (1) o no (0).*

# 3. Solución

*La tarea se ha desarrollado en lenguaje de programación Python 2.7 y fue trabajado en PyCharm 3.5.1. La solución propuesta consta de tres etapas, la primera es la carga y división de los*

*datos en una lista de entrenamiento con el 90 % de los datos y otra lista con el 10 % de datos restantes previamente desordenados de forma aleatoria para las pruebas.*

*La segunda etapa de la solución es la encargada del entrenamiento del algoritmo y a la vez de realizar la predicción utilizando las pruebas mediante el uso de probabilidades separando los tipos de datos, calculando el promedio y la desviación estándar, entregando el resultado en bruto.*

*La última etapa de la solución se enfoca al cálculo del porcentaje de predicción, pero también muestra los datos que fueron predichos correctamente y cuáles no. A la vez, se diseñó de tal manera que presentara la cantidad de positivos ciertos, positivos falsos, negativos ciertos y negativos falsos para así representar la matriz de confusión asociada al caso.*

## 4. Resultados

*Al realizar 100 pruebas, se logró constatar que el rango de predicción esta entre 69,7 % y 85,5 % mostrando así un nivel medio alto de predicción con un 90 % de los datos como entrenamiento. A continuación, se presentan los 10 porcentajes más comunes vistos con la respectiva matriz de confusión. Cabe destacar que los positivos y negativos ciertos como los negativos ciertos y falsos variaban generalmente, pero en la suma de cantidad predichas eran iguales que casos anteriores.*

*En cada una de las matrices de confusión se puede comprender de la siguiente forma. Considerando que positivos ciertos son 20, negativos falsos son 10, positivos falsos son 13 y negativos ciertos son 33, que existieron 20 casos en que se predijeron como positivos y estos fueron clasificados de manera correcta, mientras que, 10 casos se predijeron como negativos siendo que realmente eran positivos. Lo mismo se puede entender del caso de los casos negativos predichos de manera correcta, de los cuales 33 casos fueron acertados mientras que 13 de esto fueron erróneos. Este caso es el primero de los presentados a continuación, el cual obtuvo un 69,7 % de exactitud en la predicción.*

Porcentaje de exactitud: 73.6842105263%

| | | Clasificación | |
|---|---|---|---|
| | | Positivo | Negativo |
| Verdad | Positivo | 17 | 11 |
| | Negativo | 9 | 39 |

Porcentaje de exactitud: 84.2105263158%

| | | Clasificación | |
|---|---|---|---|
| | | Positivo | Negativo |
| Verdad | Positivo | 24 | 7 |
| | Negativo | 5 | 40 |

Porcentaje de exactitud: 75.0%

| | | Clasificación | |
|---|---|---|---|
| | | Positivo | Negativo |
| Verdad | Positivo | 16 | 10 |
| | Negativo | 9 | 41 |

Porcentaje de exactitud: 85.5263157895%

| | | Clasificación | |
|---|---|---|---|
| | | Positivo | Negativo |
| Verdad | Positivo | 21 | 7 |
| | Negativo | 4 | 44 |

Porcentaje de exactitud: 69.7368421053%

| | | Clasificación | |
|---|---|---|---|
| | | Positivo | Negativo |
| Verdad | Positivo | 20 | 10 |
| | Negativo | 13 | 33 |

Porcentaje de exactitud: 76.3157894737%

| | | Clasificación | |
|---|---|---|---|
| | | Positivo | Negativo |
| Verdad | Positivo | 21 | 11 |
| | Negativo | 7 | 37 |

Porcentaje de exactitud: 71.0526315789%

| | | Clasificación | |
|---|---|---|---|
| | | Positivo | Negativo |
| Verdad | Positivo | 14 | 9 |
| | Negativo | 13 | 40 |

Porcentaje de exactitud: 77.6315789474%

| | | Clasificación | |
|---|---|---|---|
| | | Positivo | Negativo |
| Verdad | Positivo | 21 | 10 |
| | Negativo | 7 | 38 |

Porcentaje de exactitud: 72.3684210526%

| | | Clasificación | |
|---|---|---|---|
| | | Positivo | Negativo |
| Verdad | Positivo | 19 | 11 |
| | Negativo | 10 | 36 |

Porcentaje de exactitud: 78.9473684211%

| | | Clasificación | |
|---|---|---|---|
| | | Positivo | Negativo |
| Verdad | Positivo | 18 | 8 |
| | Negativo | 8 | 42 |

## 5. Conclusión

*Naive Bayes es una potente herramienta en el ámbito de Machine Learning, el cual tiene una alta probabilidad de predicción en casos de tener una serie de datos supervisados. En tener conocimientos de la implementación de este algoritmo es de gran utilidad, para así tener una buena herramienta de fácil manejo y ejecución para casos supervisados.*

## 6. Referencias

*http://scikit-learn.org/stable/modules/naive$_b$ayes.html*
*https : //archive.ics.uci.edu/ml/datasets/pima + indians + diabetes*
*https : //archive.ics.uci.edu/ml/machine−learning−databases/pima−indians−diabetes/pima−indians − diabetes.data*