UNIVERSIDAD
DE GRANADA

Aplicación de Redes Bayesianas a datos genéticos

Álvaro Beltrán Camacho

14 de julio de 2021

Universidad de Granada

Índice

1. Introducción

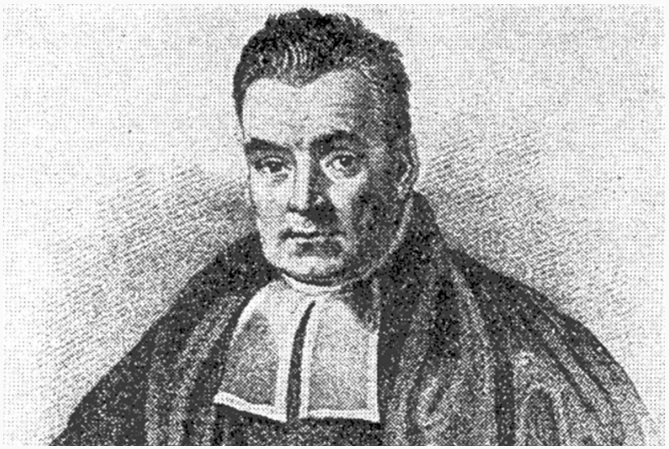
2. Teoría redes Bayesianas

3. Aplicación a datos Genéticos

Introducción

Introducción

Las **Redes Bayesianas** (RB) toman el nombre del teorema de Bayes, teorema propuesto en el siglo XVIII por el Reverendo Thomas Bayes. Pero no fue hasta finales de 1980s cuando Judea Pearl desarrolló la red Bayesiana para la creación de sistemas expertos a gran escala.



(a) Thomas Bayes

(b) Judea Pearl

Objetivos

Los objetivos fundamentales de este proyecto son:

- Estudio teórico de las RB.

Objetivos

Los objetivos fundamentales de este proyecto son:

- Estudio teórico de las RB.
  - Representación de la distribución.

Objetivos

Los objetivos fundamentales de este proyecto son:

- Estudio teórico de las RB.
  - Representación de la distribución.
  - Inferencia.

Objetivos

Los objetivos fundamentales de este proyecto son:

- Estudio teórico de las RB.
  - Representación de la distribución.
  - Inferencia.
  - Aprendizaje.

Objetivos

Los objetivos fundamentales de este proyecto son:

- Estudio teórico de las RB.
  - Representación de la distribución.
  - Inferencia.
  - Aprendizaje.
- Contrucción de forma automática una RB sobre la base de datos genética *Molecular Biology (Splice-junction Gene Sequences) Data Set*

Teoría redes Bayesianas

Estructura de red Bayesiana (RB)

Definición

Una **estructura de red Bayesiana** $\mathcal{G}$ es un grafo dirigido acíclico cuyos nodos representan variables aleatorias $X_1, ..., X_n$. Entonces $\mathcal{G}$ codifica el conjunto de relaciones de independencia condicionales, llamadas independencias locales, y denotadas como $\mathcal{I}_l(\mathcal{G})$:

$$\forall X_i : (X_i \perp NonDescendants_{X_i} | Pa_{X_i}^{\mathcal{G}}).$$

En otras palabras, cada nodo X_i es condicionalmente independiente de sus no descendientes dados sus padres.

Distribución de una RB

Definición

Sea $\mathcal{G}$ una estructura de RB sobre las variables $X_1, ..., X_n$. Se dice que una distribución P sobre el mismo espacio **factoriza según** $\mathcal{G}$ si puede ser expresado como el siguiente producto:

$$P(X_1, ..., X_n) = \prod_{i=1}^{n} P(X_i | Pa_{X_i}^{\mathcal{G}})$$

Distribución de una RB

Definición

Sea $\mathcal{G}$ una estructura de RB sobre las variables $X_1, ..., X_n$. Se dice que una distribución P sobre el mismo espacio **factoriza según** $\mathcal{G}$ si puede ser expresado como el siguiente producto:

$$P(X_1, ..., X_n) = \prod_{i=1}^{n} P(X_i | Pa_{X_i}^{\mathcal{G}})$$

Esta ecuación es llamada la regla de la cadena para RBs. El factor $P(X_i | Pa_{X_i}^{\mathcal{G}})$ es denominado distribución condicional de probabilidad (DPCs).

Definición RB

Definición

Se denomina **Red Bayesiana (RB)** al par $\mathcal{B} = (\mathcal{G}, P)$ donde P factoriza sobre $\mathcal{G}$ y donde P es especificada como un conjunto de DPCs asociadas con los nodos de $\mathcal{G}$.

Definición RB

Definición

Se denomina **Red Bayesiana (RB)** al par $\mathcal{B} = (\mathcal{G}, P)$ donde P factoriza sobre $\mathcal{G}$ y donde P es especificada como un conjunto de DPCs asociadas con los nodos de $\mathcal{G}$.

Propiedades:

- D-separación.

Definición RB

Definición

Se denomina **Red Bayesiana (RB)** al par $\mathcal{B} = (\mathcal{G}, P)$ donde P factoriza sobre $\mathcal{G}$ y donde P es especificada como un conjunto de DPCs asociadas con los nodos de $\mathcal{G}$.

Propiedades:

- D-separación.
- I-equivalencias.

Ejemplo RB

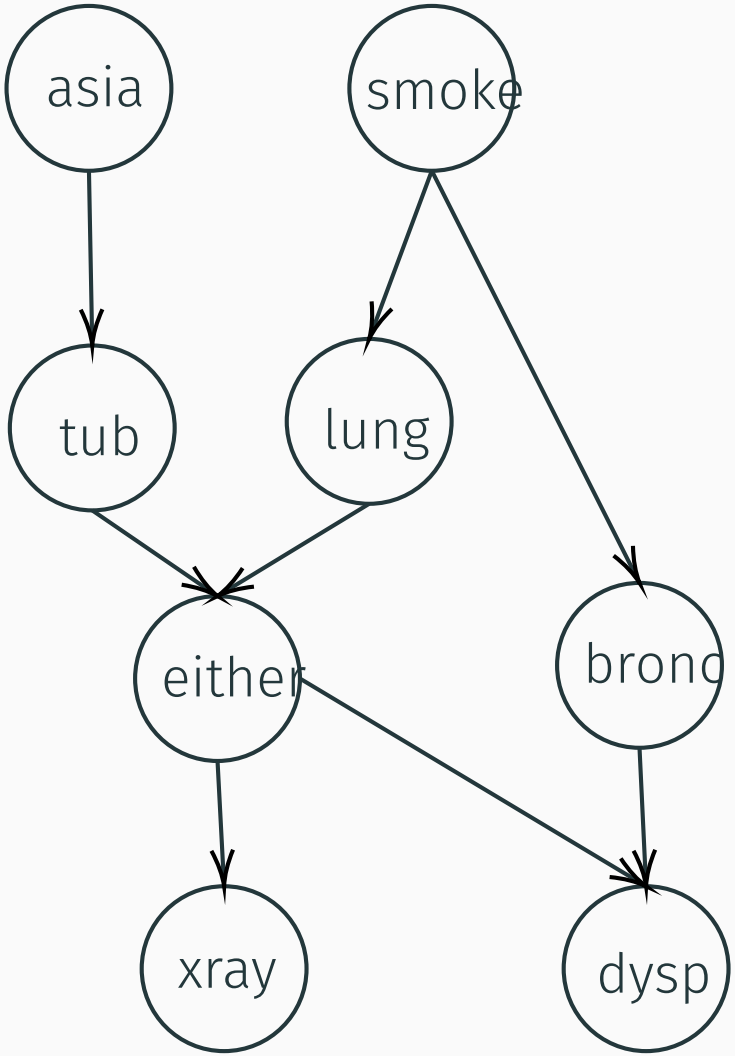


Figura 2: Red Asia

Representación de DPCs

Las formas de representar las DPCs son las siguientes:

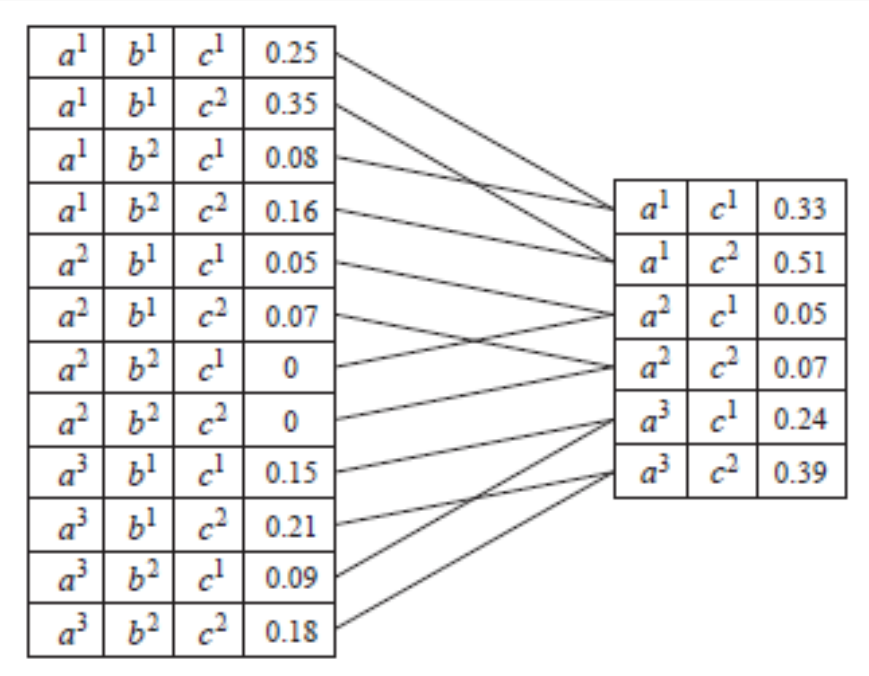
Representación de DPCs

Las formas de representar las DPCs son las siguientes:

- DPCs en formato tabla.
- DPCs deterministas.
- DPCs representadas mediante árboles.
- DPCs representadas mediante la multinomial.
- DPCs mediante gaussianas-lineales.
- DPCs para variables híbridas.

Inferencia

- Algoritmo de **eliminación de variables**.



a^1	b^1	c^1	0.25
a^1	b^1	c^2	0.35
a^1	b^2	c^1	0.08
a^1	b^2	c^2	0.16
a^2	b^1	c^1	0.05
a^2	b^1	c^2	0.07
a^2	b^2	c^1	0
a^2	b^2	c^2	0
a^3	b^1	c^1	0.15
a^3	b^1	c^2	0.21
a^3	b^2	c^1	0.09
a^3	b^2	c^2	0.18

a^1	c^1	0.33
a^1	c^2	0.51
a^2	c^1	0.05
a^2	c^2	0.07
a^3	c^1	0.24
a^3	c^2	0.39

Figura 3: ejemplo de marginalización de factores: sumar B

Aprendizaje

- **Función de pérdida.**
  - Estimación de la densidad.

$$\mathbb{E}_{\epsilon \sim P^*}[loss(\epsilon : \mathcal{M})]$$

  - Predicción
    - Error de clasificación.

$$\mathbb{E}_{(x,y)\sim\tilde{P}}[h_{\tilde{P}}(x) \neq y]$$

    - Criterio de probabilidad condicionada.

$$\mathbb{E}_{(x,y)\sim\tilde{P}}[log\tilde{P}(y|x)]$$

Aprendizaje

- **Función de pérdida.**
  - Estimación de la densidad.

$$\mathbb{E}_{\epsilon \sim P^*}[loss(\epsilon : \mathcal{M})]$$

  - Predicción
    - Error de clasificación.

$$\mathbb{E}_{(x,y)\sim\tilde{P}}[h_{\tilde{P}}(x) \neq y]$$

    - Criterio de probabilidad condicionada.

$$\mathbb{E}_{(x,y)\sim\tilde{P}}[log\tilde{P}(y|x)]$$

- Reescribir el aprendizaje como un problema de **optimización**

Estimación

Estimación

1. Aprendizaje estructural.

Estimación

1. **Aprendizaje estructural**.
    - Basado en restricciones.
        - *Grow-Shrink* (GS).
        - *Incremental Association Markov Blanket* (IAMB).

Estimación

1. **Aprendizaje estructural**.
    - Basado en restricciones.
        - *Grow-Shrink* (GS).
        - *Incremental Association Markov Blanket* (IAMB).
    - Basado en métricas.
        - *Hill-Climbing* (HC).
        - Métricas: AIC, BIC, *log-likelihood*, *K2*.

Estimación

1. **Aprendizaje estructural**.
    - Basado en restricciones.
        - *Grow-Shrink* (GS).
        - *Incremental Association Markov Blanket* (IAMB).
    - Basado en métricas.
        - *Hill-Climbing* (HC).
        - Métricas: AIC, BIC, *log-likelihood*, *K*2.
    - Algoritmos híbridos.
        - *Max-Min Hill-Climbing* (MMHC).

Estimación

1. **Aprendizaje estructural**.
    - Basado en restricciones.
        - *Grow-Shrink* (GS).
        - *Incremental Association Markov Blanket* (IAMB).
    - Basado en métricas.
        - *Hill-Climbing* (HC).
        - Métricas: AIC, BIC, *log-likelihood*, *K*2.
    - Algoritmos híbridos.
        - *Max-Min Hill-Climbing* (MMHC).
2. **Estimación paramétrica**.

Estimación

1. **Aprendizaje estructural**.
   - Basado en restricciones.
     - *Grow-Shrink* (GS).
     - *Incremental Association Markov Blanket* (IAMB).
   - Basado en métricas.
     - *Hill-Climbing* (HC).
     - Métricas: AIC, BIC, *log-likelihood*, *K2*.
   - Algoritmos híbridos.
     - *Max-Min Hill-Climbing* (MMHC).
2. **Estimación paramétrica**.
   - Estadístico máximo verosímil (EMV)

Aplicación a datos Genéticos

Aplicaciones de las Redes Bayesianas al análisis de datos genéticos

Para tener una visión más general a la hora de aprender una RB a partir de los datos genéticos que se van a estudiar, se han elegido los artículos:

Aplicaciones de las Redes Bayesianas al análisis de datos genéticos

Para tener una visión más general a la hora de aprender una RB a partir de los datos genéticos que se van a estudiar, se han elegido los artículos:

- *Using Bayesian Networks to Analyze Expression Data* [Friedman et al., 2000].

Aplicaciones de las Redes Bayesianas al análisis de datos genéticos

Para tener una visión más general a la hora de aprender una RB a partir de los datos genéticos que se van a estudiar, se han elegido los artículos:

- *Using Bayesian Networks to Analyze Expression Data* [Friedman et al., 2000].
- *Using Bayesian networks to discover relations between genes, environment, and disease* [Su et al., 2013].

Using Bayesian Networks to Analyze Expression Data

En este artículo los autores usan las redes Bayesianas para representar las dependencias estadísticas entre los genes. El estudio de las relaciones entre los genes surge del avance tecnológico que proporcionó el microarray de ADN.

Using Bayesian Networks to Analyze Expression Data

En este artículo los autores usan las redes Bayesianas para representar las dependencias estadísticas entre los genes. El estudio de las relaciones entre los genes surge del avance tecnológico que proporcionó el microarray de ADN.

En este proyecto se usan los datos sobre el ciclo celular de *S. cerevisiae* [Spellman et al., 1998]. Tratan todas las muestras como independientes y toman una variable aleatoria por gen, aparte de una para controlar el ciclo celular.

Using Bayesian Networks to Analyze Expression Data

En este artículo los autores usan las redes Bayesianas para representar las dependencias estadísticas entre los genes. El estudio de las relaciones entre los genes surge del avance tecnológico que proporcionó el microarray de ADN.

En este proyecto se usan los datos sobre el ciclo celular de *S. cerevisiae* [Spellman et al., 1998]. Tratan todas las muestras como independientes y toman una variable aleatoria por gen, aparte de una para controlar el ciclo celular.

Usaron dos modelos probabilísticos locales:

- **Modelo Multinomial**. Para el cual discretizaron los datos en : *underexpressed*, normal, y *over-expressed* .
- **Modelo Gaussiano-lineal**. Sin cambios para las variables continuas.

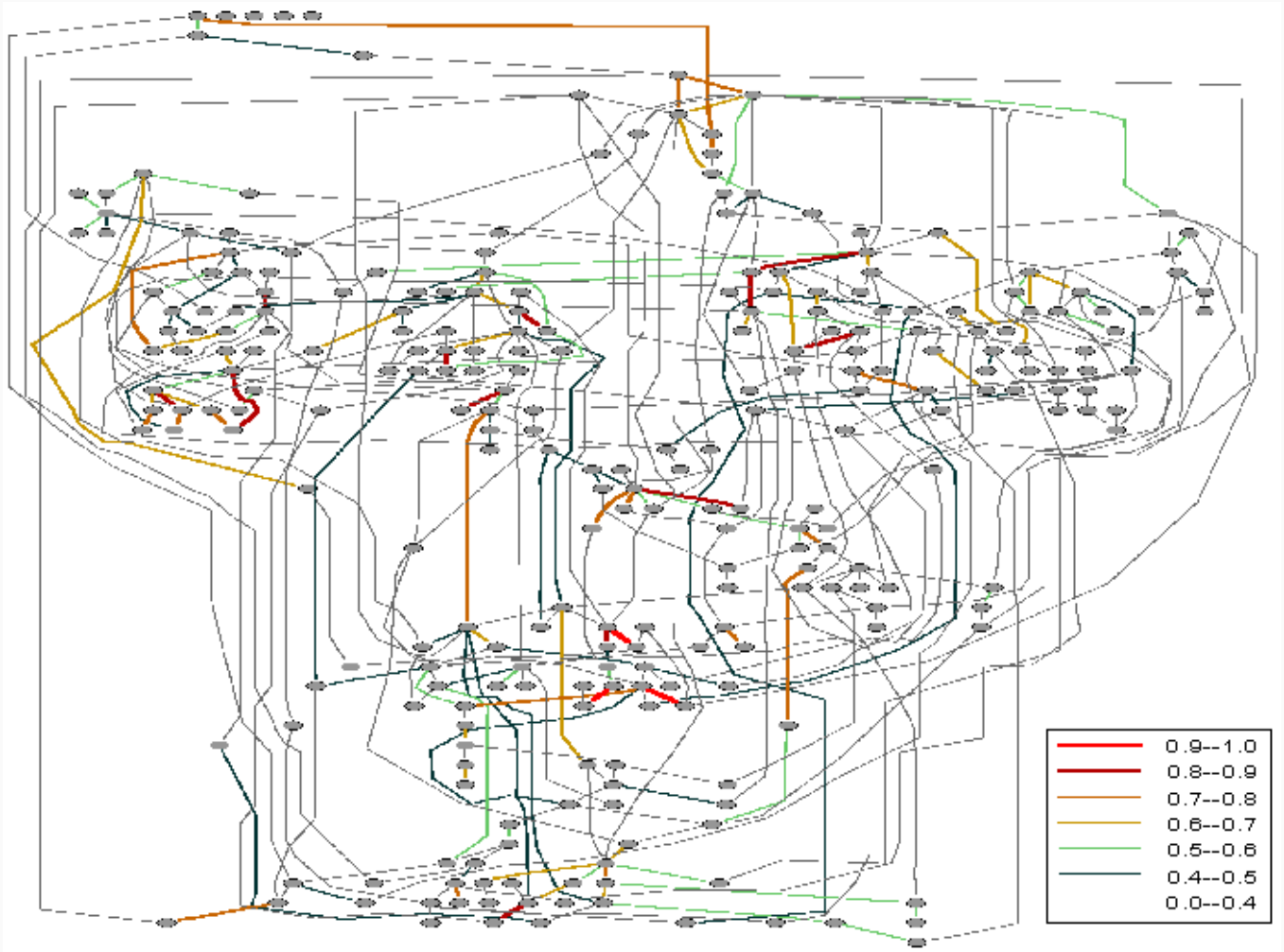


Figura 4: Red Bayesiana

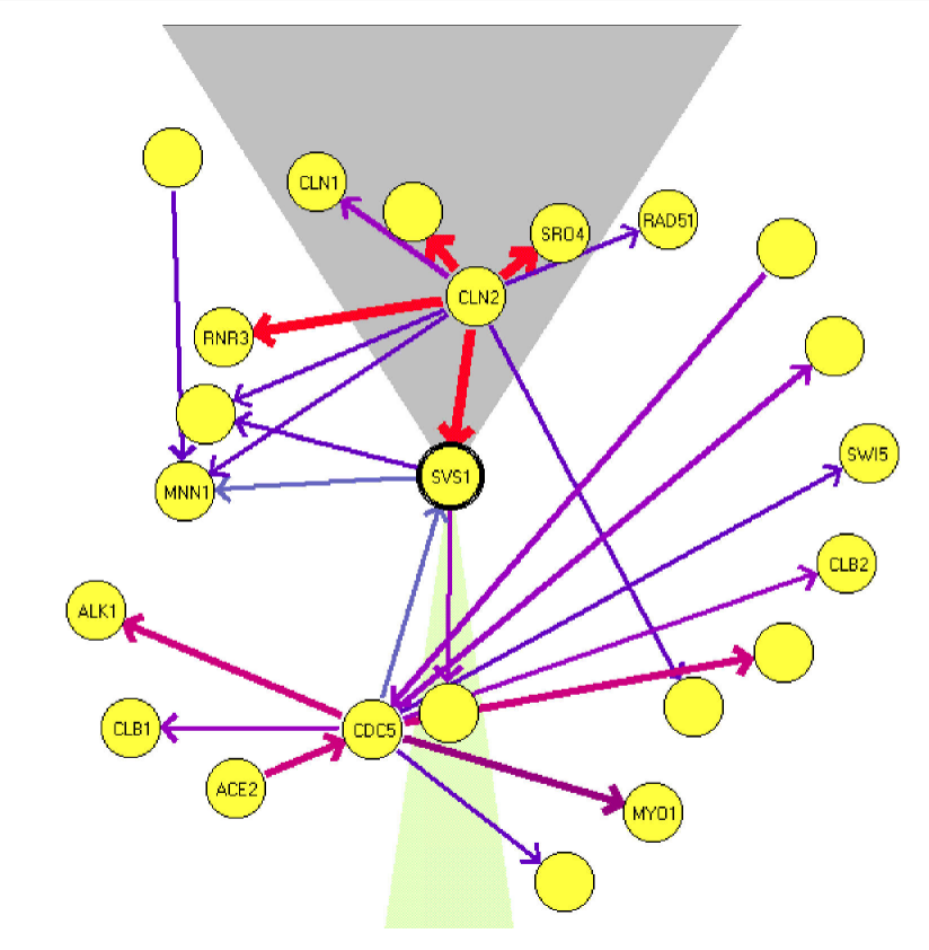


Figura 5: Red Bayesiana sobre el gen SVS1

Using Bayesian networks to discover relations between genes, environment, and disease

En este artículo se estudia las relaciones entre el cáncer de vejiga, la genética y el medio ambiente. El estudio de los genes se va a hacer mediante los polimorfismos de nucleótido único (SNPs).

Using Bayesian networks to discover relations between genes, environment, and disease

En este artículo se estudia las relaciones entre el cáncer de vejiga, la genética y el medio ambiente. El estudio de los genes se va a hacer mediante los polimorfismos de nucleótido único (SNPs).

Variables que se estudian:

- Fumador.
- Sexo.
- Cáncer de vejiga.
- Exposición al arsénico.
- Edad.
- Gen XRCC3 en las posiciones 03, 04 y 241.
- Gen ERCC2/XPD en las posiciones 03, 09 y 312.

Using Bayesian networks to discover relations between genes, environment, and disease

Los autores han decidido estudiar los algoritmos

- *Grow-Shrink* (GS).
- *Incremental Association Markov Blanket* (IAMB).
- *Hill-Climbing* (HC).
- *Max-Min Hill-Climbing* (MMHC).

Using Bayesian networks to discover relations between genes, environment, and disease

Los autores han decidido estudiar los algoritmos

- *Grow-Shrink* (GS).
- *Incremental Association Markov Blanket* (IAMB).
- *Hill-Climbing* (HC).
- *Max-Min Hill-Climbing* (MMHC).

Para los algoritmos basados en métricas usan la métrica $log(k2)$.

Using Bayesian networks to discover relations between genes, environment, and disease

Los autores han decidido estudiar los algoritmos

- *Grow-Shrink* (GS).
- *Incremental Association Markov Blanket* (IAMB).
- *Hill-Climbing* (HC).
- *Max-Min Hill-Climbing* (MMHC).

Para los algoritmos basados en métricas usan la métrica $log(k2)$.

Escogen la red resultante del **algoritmo HC** ya que consigue el mejor resultado.

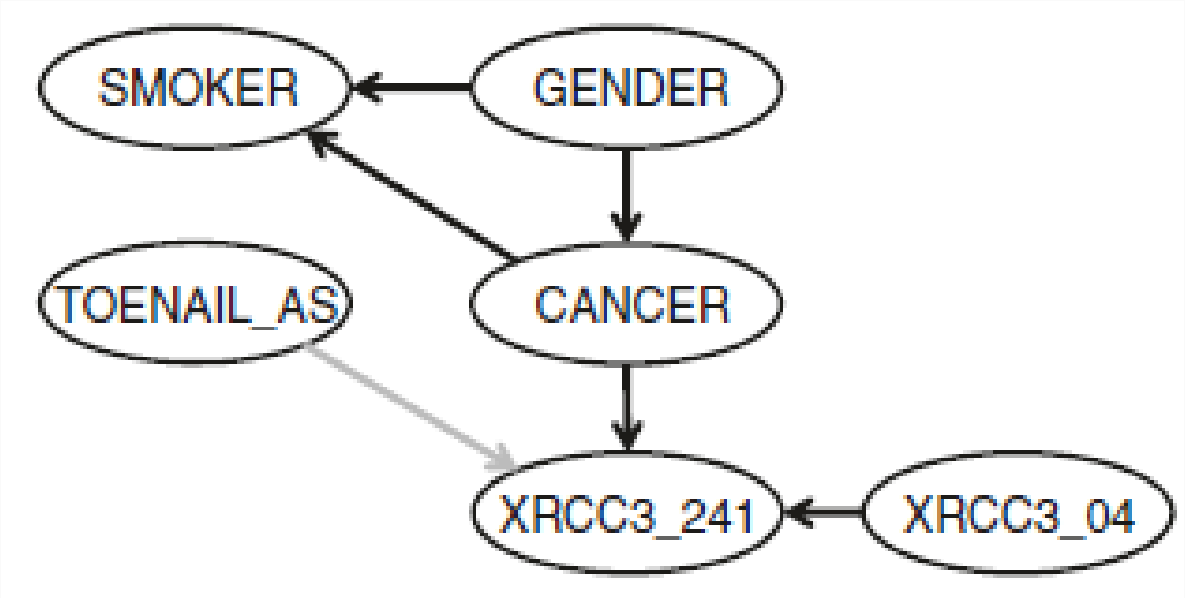


Figura 6: Red Bayesiana

Base de datos genética

La base de datos que se va a usar es ***Molecular Biology (Splice-junction Gene Sequences) Data Set*** [G Towell and Shavlik, 1991].

Empalme alternativo

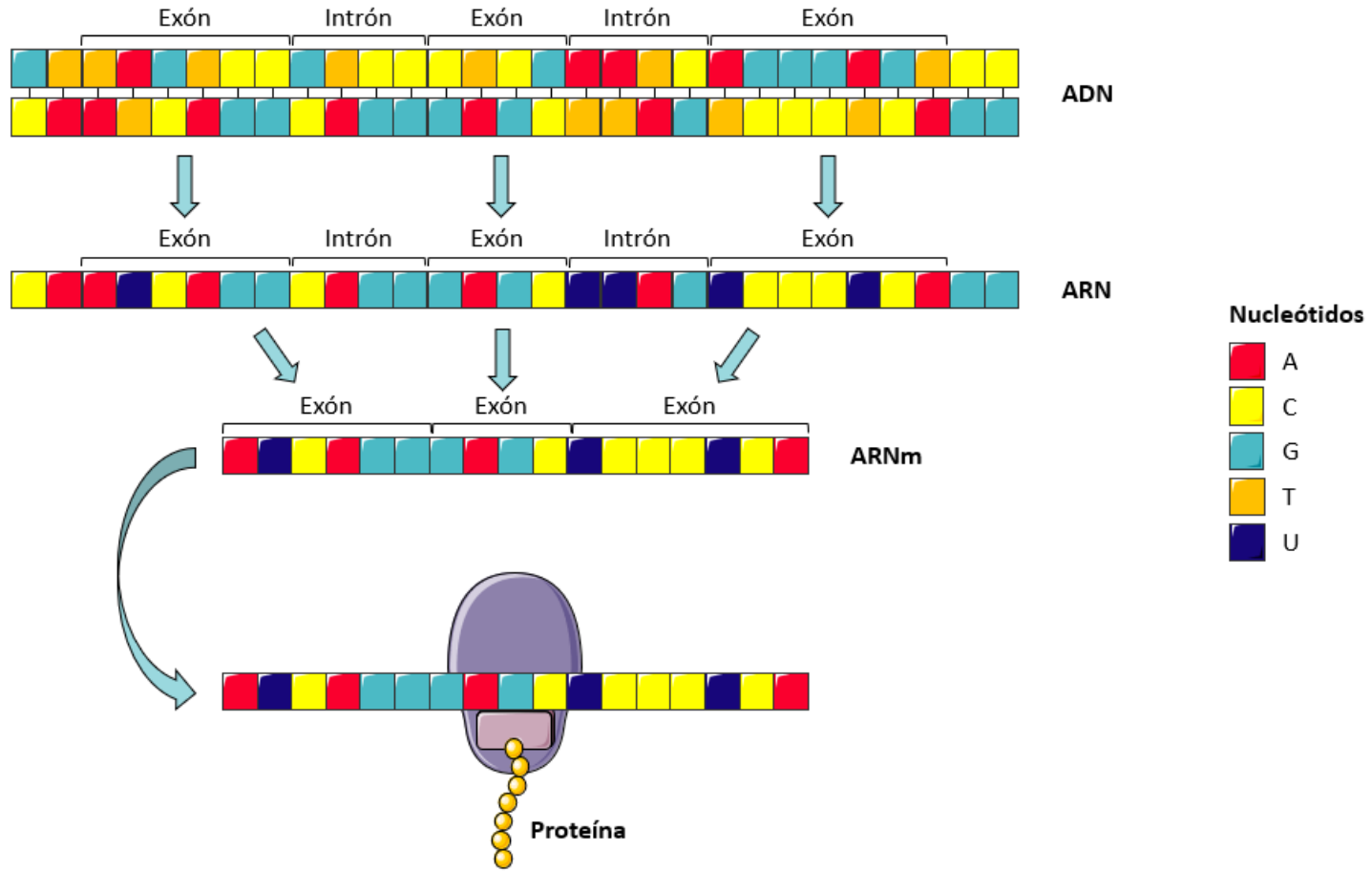


Figura 7: Empalme Alternativo

Explicación de la base de datos y análisis preliminar

La base de datos consta de 3190 instancias compuestas por secuencias de 60 posiciones de ADN. Cada una de las instancias etiquetadas mediante un tipo de región de empalme.

Explicación de la base de datos y análisis preliminar

La base de datos consta de 3190 instancias compuestas por secuencias de 60 posiciones de ADN. Cada una de las instancias etiquetadas mediante un tipo de región de empalme.

- **EI**. Se ha producido un cambio de exón a intrón.
- **IE**. Se ha producido un cambio de intrón a exón
- **N**. En el gen no hay región de empalme.

Explicación de la base de datos y análisis preliminar

Cada una de las posiciones de la secuencia del gen toma ocho posibles valores.

Carácter	Significado
A	Adenina
G	Guanina
T	Timina
C	Citosina

Carácter	Significado
D	A o G o T
N	A o G o C o T
S	C o G
R	A o G

Cuadro 1: Posibles valores de las posiciones de la secuencia genética

Estructura base de datos

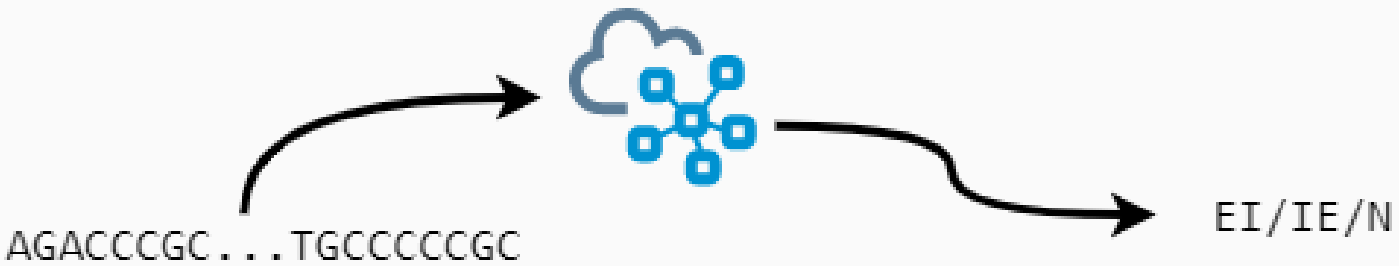
Nombre	Secuencia	Etiqueta
ATRINS-DONOR-521	CCAGCTGC...GCCAGTCTG	EI
ATRINS-DONOR-905	AGACCCGC...TGCCCCCGC	EI
BABAPOE-DONOR-30	GAGGTGAA...ACGGGGATG	EI
BABAPOE-DONOR-867	GGGCTGCG...GTTTTCCCC	EI
BABAPOE-DONOR-2817	GCTCAGCC...CTTGACCCT	EI
...	...	...

Cuadro 2: Ejemplo instancias base de datos

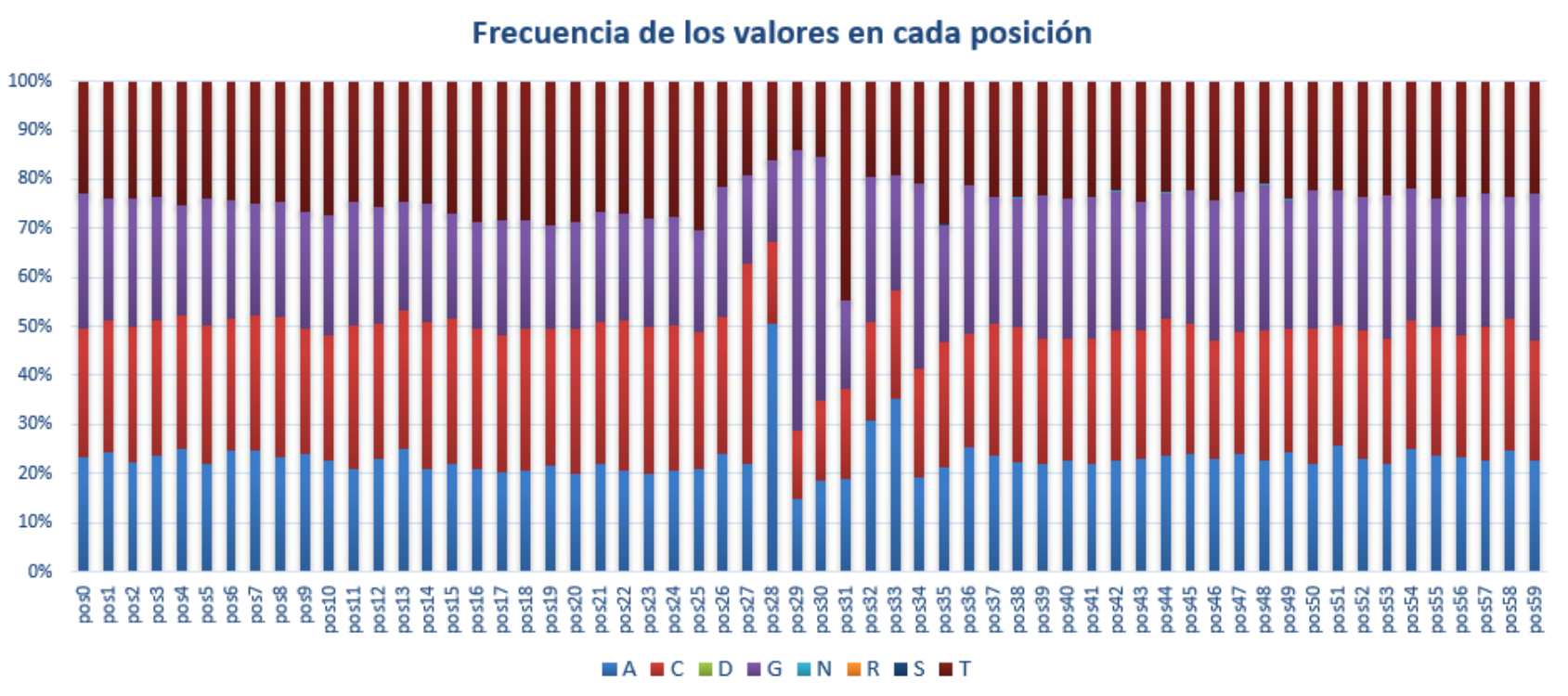
Definición del Problema

Definition

El problema que se va a tratar va a ser la **clasificación** de una secuencia en un tipo de región de empalme. Por tanto las posiciones de la secuencia serán las variables predictoras y la variable etiqueta o clase será la variable a predecir.



Análisis de la base de datos



Análisis de la base de datos

	Nº Genes
EI	767
IE	768
N	1655

Cuadro 3: Nº de genes por etiqueta

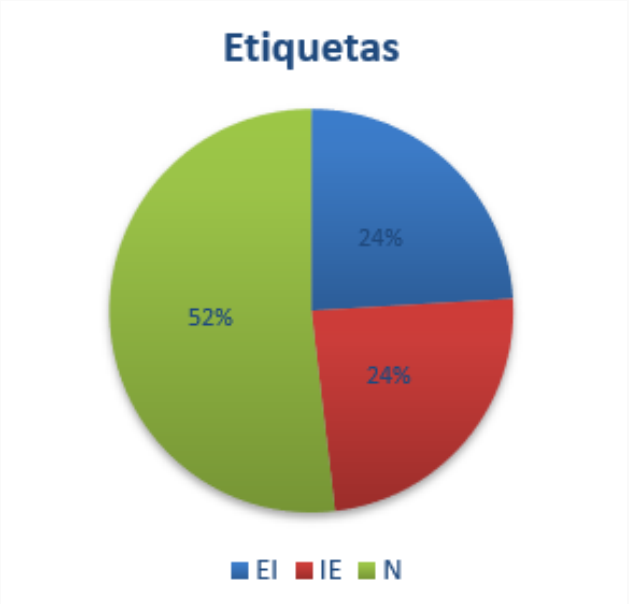


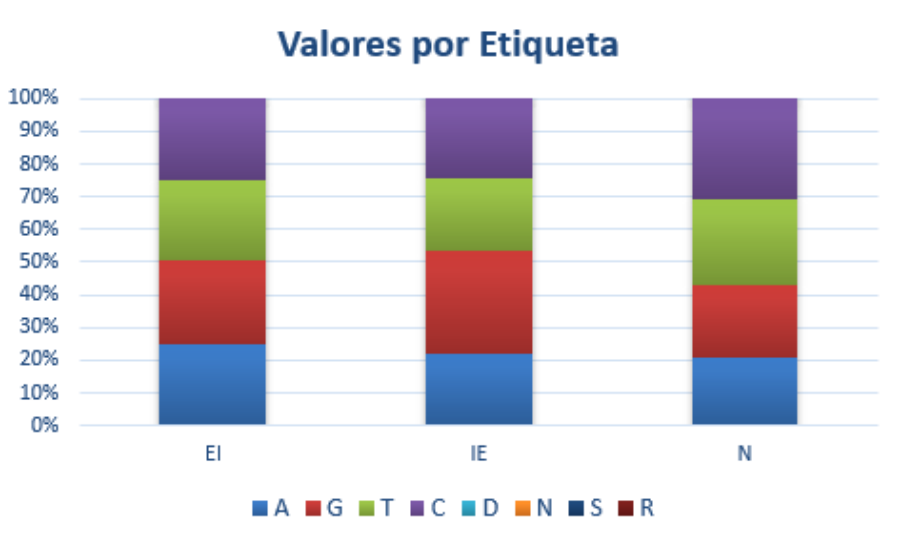
Figura 8: Frecuencias etiquetas

Análisis de la base de datos

	EI (%)	IE (%)	N (%)
A	24, 984	22, 1534	20, 577
G	25, 653	31, 415	22, 383
T	24, 273	21, 771	26, 445
C	25, 077	24, 561	30, 588

	EI (%)	IE (%)	N (%)
D	0, 001	0	0, 002
N	0, 01	0, 01	0
S	0	0	0, 002
R	0	0	0, 002

Cuadro 4: Frecuencia valores posiciones por etiqueta



Estudio experimental

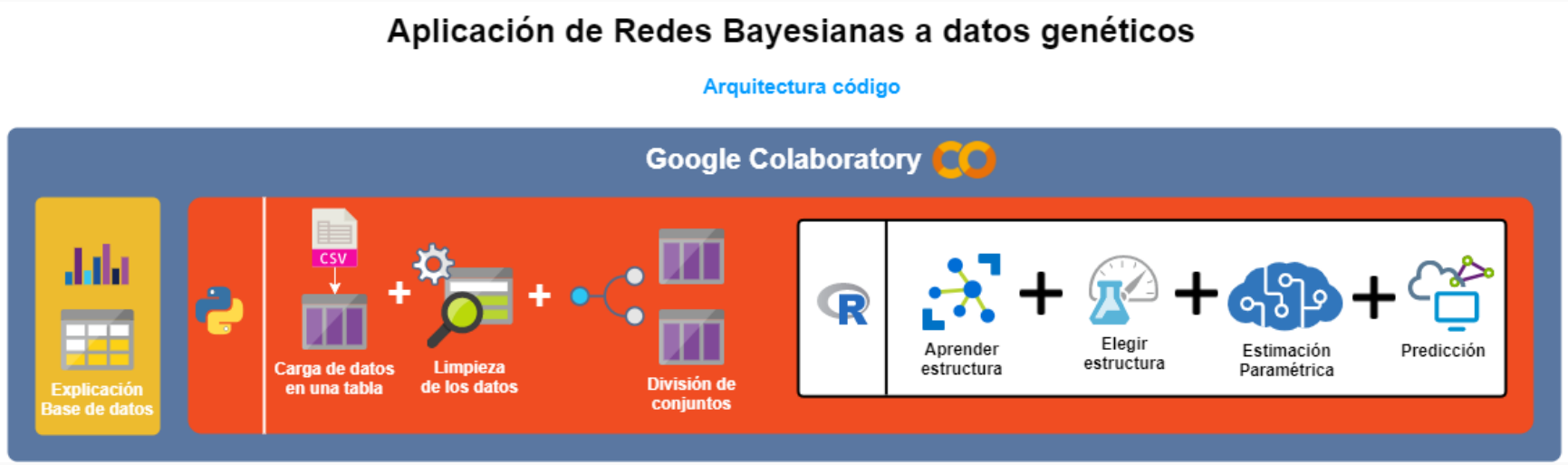
Este estudio va a consistir en un preprocesamiento de los datos aplicando algunas técnicas estadísticas; para luego estimar la estructura y los parámetros de la distribución asociada a los datos. Y finalmente, valorar la bondad del modelo frente al problema de clasificación comentado.

Estudio experimental

Este estudio va a consistir en un preprocesamiento de los datos aplicando algunas técnicas estadísticas; para luego estimar la estructura y los parámetros de la distribución asociada a los datos. Y finalmente, valorar la bondad del modelo frente al problema de clasificación comentado.

Para la implementación de la experimentación se ha desarrollado un cuaderno *Python*

Arquitectura del cuaderno



Carga de datos

Para la carga de datos se usa la librería ***Pandas*** de *Python*.

Carga de datos

Para la carga de datos se usa la librería ***Pandas*** de *Python*.

- Formato de entrada.

```
class,Name,Sequence
EI,ATRINS-DONOR-521, CCAGCTGCATC...CGAGCCAGTC
EI,ATRINS-DONOR-905, AGACCCGCCGG...CCGTGCCCCC
EI,BABAPOE-DONOR-30, GAGGTGAAGGA...GGCACGGGGA
EI,BABAPOE-DONOR-867, GGGCTGCGTTG...TCGGTTTTCC
EI,BABAPOE-DONOR-2817, GCTCAGCCCCC...GCCCTTGACC
```

Carga de datos

- Formato de almacenamiento.

class	Name	Sequence
EI	ATRINS-DONOR-521	CCAGCTGCATCACAGGAG...CTTCGAGCCAGTC
EI	ATRINS-DONOR-905	AGACCCGCCGGGAGGCGG...CCTCCGTGCCCCC
EI	BABAPOE-DONOR-30	GAGGTGAAGGACGTCCTT...GGGGGCACGGGGA
EI	BABAPOE-DONOR-867	GGGCTGCGTTGCTGGTCA...TGCTCGGTTTTCC
EI	BABAPOE-DONOR-2817	GCTCAGCCCCCAGGTCAC...CCGGCCCTTGACC
...	...	...

Cuadro 5: Formato datos cargados

Limpieza de los datos

- Cada posición de la secuencia es una variable.
- La variable *name* no aporta información.
- Es necesario fijar el tipo de los datos a *category*, ya que son datos categóricos.

Limpieza de los datos

- Cada posición de la secuencia es una variable.
- La variable *name* no aporta información.
- Es necesario fijar el tipo de los datos a *category*, ya que son datos categóricos.

class	*pos*0	*pos*1	*pos*2	*pos*3	*pos*4	*pos*5	...	*pos*53	*pos*54	*pos*55	*pos*56	*pos*57	*pos*58	*pos*59
EI	C	C	A	G	C	T	...	C	A	G	T	C	T	G
EI	A	G	A	C	C	C	...	C	C	C	C	C	G	C
EI	G	A	G	G	T	G	...	G	G	G	G	A	T	G
EI	G	G	G	C	T	G	...	T	T	T	C	C	C	C
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Cuadro 6: Formato tabla de variables

Limpieza de los datos

- Cada posición de la secuencia es una variable.
- La variable *name* no aporta información.
- Es necesario fijar el tipo de los datos a *category*, ya que son datos categóricos.

class	*pos*0	*pos*1	*pos*2	*pos*3	*pos*4	*pos*5	...	*pos*53	*pos*54	*pos*55	*pos*56	*pos*57	*pos*58	*pos*59
EI	C	C	A	G	C	T	...	C	A	G	T	C	T	G
EI	A	G	A	C	C	C	...	C	C	C	C	C	G	C
EI	G	A	G	G	T	G	...	G	G	G	G	A	T	G
EI	G	G	G	C	T	G	...	T	T	T	C	C	C	C
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Cuadro 6: Formato tabla de variables

- Eliminar instancias irrelevantes. La tabla pasa de tener 3190 instancias a 3175.

Limpieza de los datos

- Cada posición de la secuencia es una variable.
- La variable *name* no aporta información.
- Es necesario fijar el tipo de los datos a *category*, ya que son datos categóricos.

class	*pos*0	*pos*1	*pos*2	*pos*3	*pos*4	*pos*5	...	*pos*53	*pos*54	*pos*55	*pos*56	*pos*57	*pos*58	*pos*59
EI	C	C	A	G	C	T	...	C	A	G	T	C	T	G
EI	A	G	A	C	C	C	...	C	C	C	C	C	G	C
EI	G	A	G	G	T	G	...	G	G	G	G	A	T	G
EI	G	G	G	C	T	G	...	T	T	T	C	C	C	C
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Cuadro 6: Formato tabla de variables

- Eliminar instancias irrelevantes. La tabla pasa de tener 3190 instancias a 3175.
- Selección de variables.

Selección de variables

Para selección de variables se hace lo siguiente:

1. Codificar los valores de las variables predictoras.

valor	A	C	G	T
codificación	0	1	2	3

Selección de variables

Para selección de variables se hace lo siguiente:

1. Codificar los valores de las variables predictoras.

valor	A	C	G	T
codificación	0	1	2	3

2. Realizar el **test de independencia** χ^2.

Selección de variables

Para selección de variables se hace lo siguiente:

1. Codificar los valores de las variables predictoras.

valor	A	C	G	T
codificación	0	1	2	3

2. Realizar el **test de independencia** χ^2.
3. Realizar el **test de varianzas**.

División de conjuntos

Una vez realizada la división, el conjunto de entrenamiento queda con 1905 instancias, el de validación con 635 y el de comprobación con 635.

Aprendizaje de la estructura

- *Grow-Shrink* (GS).
- *Incremental Association Markov Blanket* (IAMB).
- *Hill-Climbing* (HC).
- *Max-Min Hill-Climbing* (MMHC).

Algoritmo HC

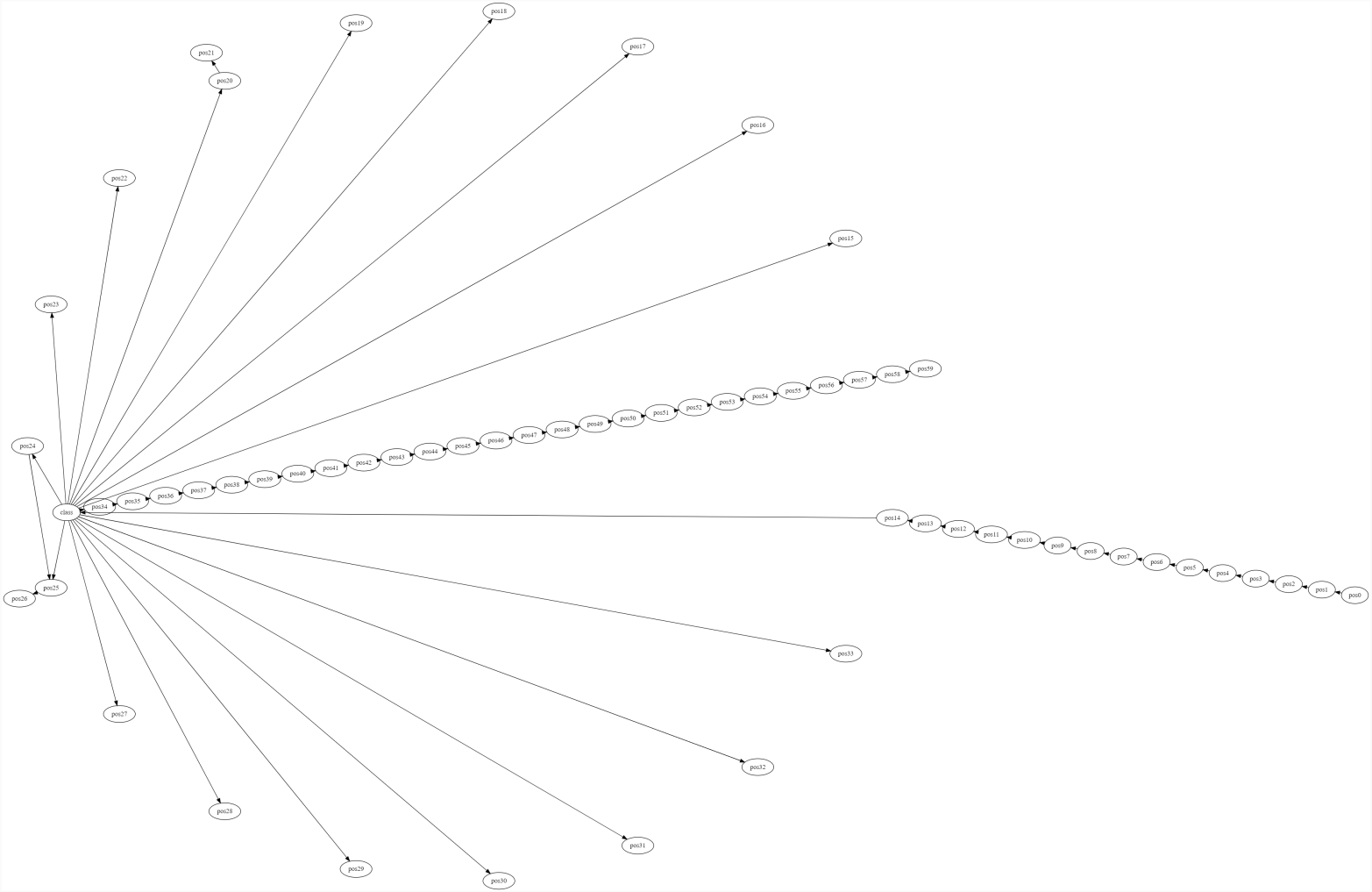


Figura 9: Estructura algoritmo HC

Algoritmo MMHC

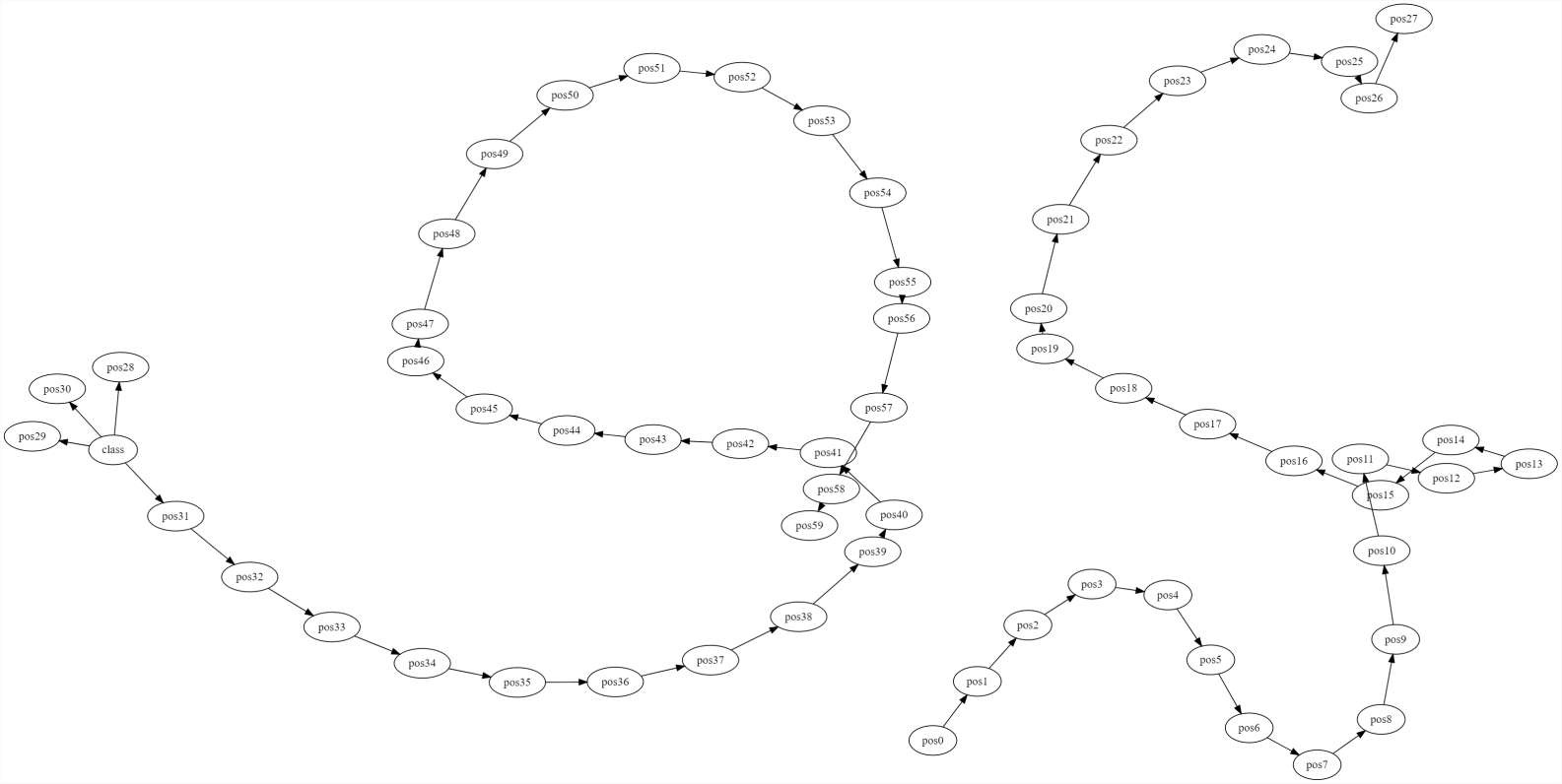


Figura 10: Estructura algoritmo MMHC

Algoritmo IAMB

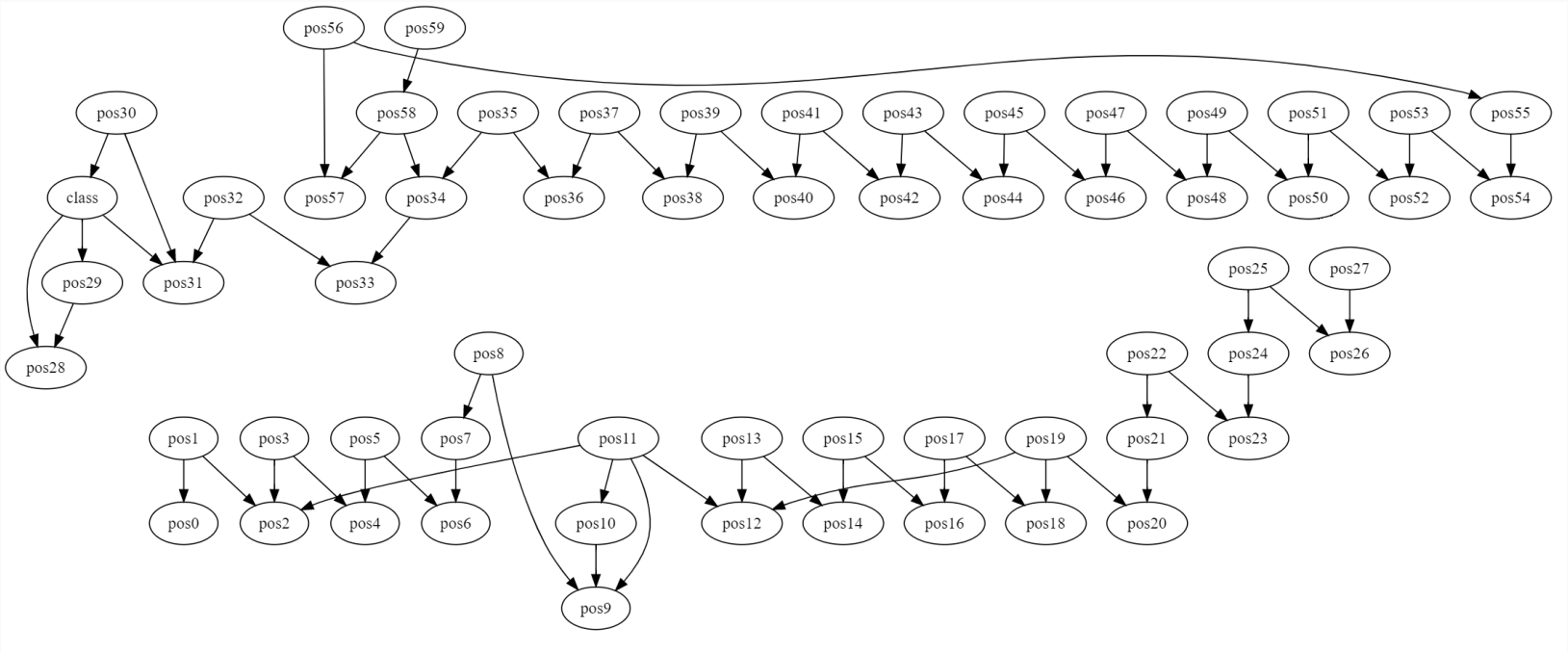


Figura 11: Estructura algoritmo IAMB

Algoritmo GS

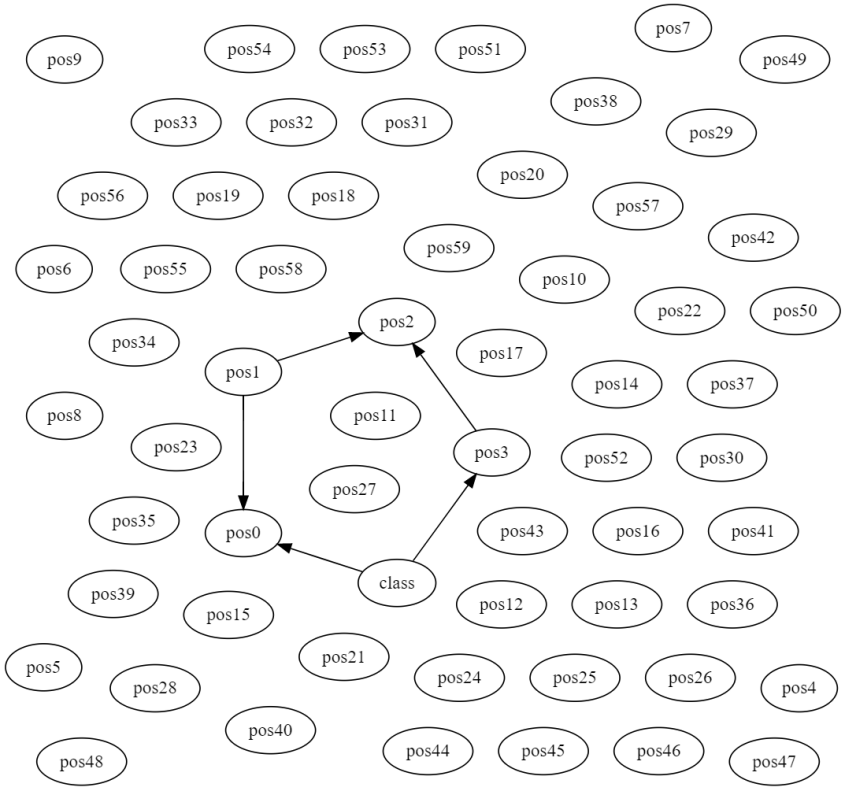


Figura 12: Estructura algoritmo GS

Elegir estructura

Algoritmos	*loglik*	*AIC*	k2	BIC
HC	-50237.382946	-50929.382946	-51655.753168	-52850.457047
MMHC	-50527.272868	-51228.272868	-51956.548474	-53174.332038
IAMB	-49739.979075	-51682.979075	-52244.343777	-56009.675761
GS	-52638.233160	-52904.233160	-53233.494716	-53496.565285

Cuadro 7: Métricas estructuras algoritmos

Estimar la distribución

Listing 1: DPCs de la distribución en formato tabla

```
Bayesian network parameters

  Parameters of node class (multinomial distribution)

Conditional probability table:

     pos14
class          A         C         G         T
   EI  0.2549020 0.2539130 0.2729258 0.1788793
   IE  0.1519608 0.3286957 0.1419214 0.3081897
   N   0.5931373 0.4173913 0.5851528 0.5129310
...
```

Bondad del modelo

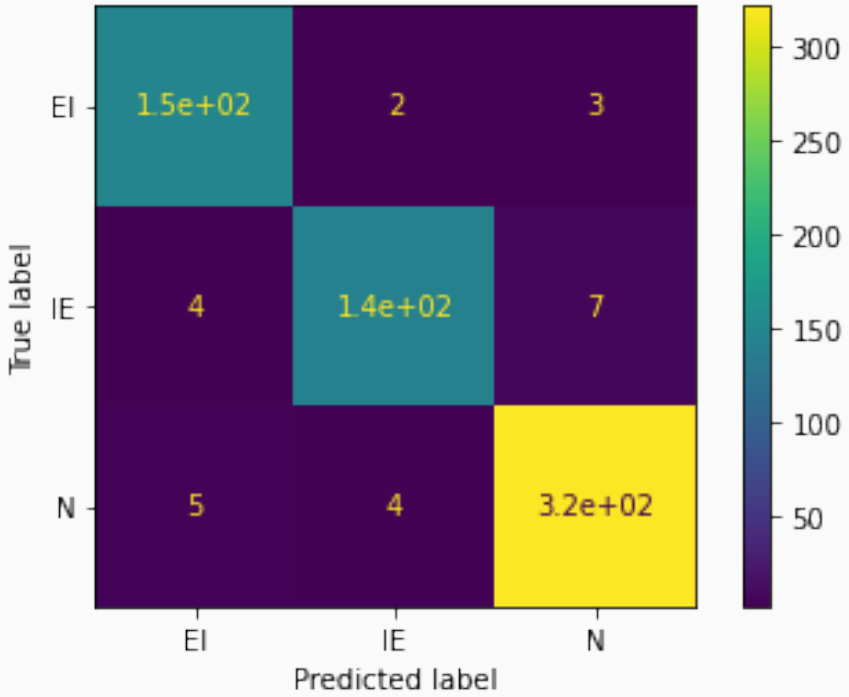


Figura 13: Matriz de confusión sobre el conjunto de comprobación

Bondad del modelo

Métrica	*F1-score*	*Recall score*
Conjunto validación (val)	0,96057336	0,96062992

Cuadro 8: Métricas sobre el problema de clasificación.

Conclusión

Fin