El futuro de España: ¡Predigamos la tasa de crecimiento de la población!

Lucas Manuel Herencia Solís
dpto. Ciencias de la Computación e Inteligencia Artificial
Universidad de Sevilla
Sevilla, España
luchersol@alum.us.es , lmherencia2003@gmail.com

Daniel Galván Cancio
dpto. Ciencias de la Computación e Inteligencia Artificial
Universidad de Sevilla
Sevilla, España
dangalcan@alum.us.es , megamagolas@gmail.com

***Resumen*— A lo largo de este documento expondremos nuestro trabajo relacionado con las redes bayesianas realizado para la asignatura de Inteligencia Artificial. Nos hemos propuesto la siguiente pregunta como punto de partida: "¿Cómo influyen los indicadores económicos y demográficos en la tasa de crecimiento de la población en España?", y para dar respuesta nos haremos uso de las redes bayesianas.**

Tras el estudio hemos concluido que para este caso es mejor emplear la inferencia exacta porque obtiene resultados precisos y acorde con los datos reales. Además, podemos concluir que los datos de entrenamiento son buenos y lo suficientemente extensos como para obtener predicciones precisas, y que dichas predicciones se adaptan bien a la realidad.

Por último hemos obtenido como observación que una buena situación económica con buen gasto en salud y moderado en educación, poseyendo además una buena tasa de migración neta, dará lugar a una tasa de crecimiento de la población alta.

Palabras Clave—Inteligencia Artificial, redes bayesianas, inmigración, PIB, tasa de mortalidad, paro, ingreso nacional bruto, tasa de crecimiento, salud, fertilidad, etc

I. Introducción

Como sabemos, las redes Bayesianas son modelos de Inteligencia artificial que se basan en probabilidad y nos permiten hacer predicciones en base a ellas. Se usan en gran variedad de ámbitos, es más, citamos textualmente: "Las redes bayesianas son útiles para realizar inferencias y aprender de datos, y se aplican en una variedad de campos como la meteorología, el diagnóstico médico, la visión artificial y el aprendizaje automático" [1]

En nuestro caso hemos escogido hacer el trabajo de la asignatura sobre redes bayesianas porque redes Bayesianas es un tema interesante y con muchas posibilidades. No obstante, también queremos aprender y dar respuesta a una pregunta interesante: "¿Cómo influyen los indicadores económicos y demográficos en la tasa de crecimiento de la población española?". Nuestro cometido surgió porque estuvimos hablando sobre la tasa de natalidad de los países del primer mundo y apareció el debate de si está ligado el ritmo de vida de lugares como Estados Unidos u Occidente con una tasa de natalidad baja. A raíz de eso salió la propuesta de realizar dicho estudio enfocándonos en España y tras la validación del profesor nos pusimos a trabajar en ello.

Nuestro objetivo es saber en qué medida influyen ciertos factores sociales y económicos sobre la tasa de crecimiento de la población del país. A priori parece una pregunta sencilla de responder pero a medida que ahondamos en ella nos damos cuenta de que no lo es tanto. Nuestra hipótesis inicial fue que a mayor calidad económica, menor tasa de crecimiento de la población, ya que podemos ver que en países en vías de desarrollo la natalidad está por las nubes en comparación con otros mucho más competitivos económicamente, como Estados Unidos, Alemania, y otros países europeos.

En el presente artículo desarrollaremos todo el trabajo realizado así como las conclusiones obtenidas.

II. Preliminares

En esta sección se hace una breve introducción de las técnicas empleadas y también trabajos relacionados.

A. Métodos empleados

Para poder abordar la situación que se nos presenta, emplearemos las redes bayesianas. Una red bayesiana, o modelo probabilístico gráfico, es una representación gráfica de un conjunto de variables y sus dependencias condicionales mediante un grafo acíclico dirigido. Se basan en la teoría de la probabilidad y modelan problemas complejos. Cada nodo en el grafo representa una variable, mientras que las aristas dirigidas indican dependencias probabilísticas entre las variables.

Según Pearl (1988), proporcionan una forma compacta y eficiente de representar y calcular distribuciones conjuntas de variables. Esta capacidad de descomposición y factorización permite realizar inferencias (de las que hablaremos más en detalle posteriormente) de manera eficiente en sistemas con múltiples variables interdependientes. [2] Russell y Norvig (2010) explican que éstas son especialmente útiles en inteligencia artificial y aprendizaje automático, ya que facilitan el razonamiento bajo incertidumbre y la toma de decisiones en situaciones complejas y cambiantes. Las redes bayesianas pueden entrenarse con datos, permitiendo así que los modelos se adapten y mejoren conforme se dispone de más información. [3]

Para la documentación hemos hecho el documento basándonos en la plantilla proporcionada en la web de la asignatura, es decir, siguiendo el convenio definido por IEEE [4]. En cuanto a la implementación, hemos utilizado la librería pgmpy y como entorno de desarrollo VSCode y Jupyter [5].

Además de esto, nuestro trabajo se puede encontrar en nuestro github. [6]

B. *Trabajo Relacionado*

Evidentemente, no somos los primeros en estudiar la evolución de la cantidad de población en relación con factores económicos y/o demográficos, es más, es un tema muy de moda en la actualidad. De hecho, podemos ver que hay noticias con respecto a estos temas al respecto, de donde vemos que en 2023, se han registrado 238.766 bebés, la cifra más baja desde 2019, un 2,87% menos que el año anterior. España tiene la segunda tasa de natalidad más baja de Europa con 7,6 nacimientos por cada 1000 habitantes, por debajo de la media europea (9,3). Madrid y Extremadura tuvieron el mayor número de nacimientos en septiembre, con 4.412 y 614 respectivamente. [7]

Con respecto al uso de modelos matemáticos, hemos encontrado un estudio que analiza la evolución de la natalidad en España titulado "Evolución de la natalidad en España. Análisis de la tendencia de los nacimientos entre 1941 y 2010" y fue publicado en la revista Anales de Pediatría. Analiza la evolución de la natalidad en España entre 1941 y 2010. Su objetivo fue examinar las tendencias de nacimientos en España y sus regiones a lo largo de 70 años empleando modelos de regresión de Joinpoint para calcular las tasas brutas de natalidad y detectar puntos de cambio y porcentajes anuales de cambio (PAC). Pudieron concluir que existen cambios en las tendencias de natalidad, cosa que puede permitir a las autoridades sanitarias planificar adecuadamente los recursos asistenciales pediátricos en el país. [8]

Debido a ello, para la realización de un estudio sobre ello deben ser estudiadas aquellas variables que puedan condicionar al ratio del crecimiento de la población. Por ello, el uso de una red bayesiana, que es mayormente utilizada para la predicción con conocimiento de probabilidades condicionadas, llega a ser la mejor opción para su desarrollo.

III. Metodología

Esta sección se dedica a la descripción del método implementado en el trabajo. Esta parte es la correspondiente a lo realmente desarrollado en el trabajo.

Como mencionamos anteriormente, hemos utilizado un entorno de desarrollo de python para trabajar con redes Bayesianas. La librería en cuestión se llama pgmpy. Antes de nada es necesario instalarla así como importar los módulos necesarios.

Para la creación de la red se debe usar el modelo de BayesianNetwork aportado por pgmpy. Cada nodo será representado como una cadena, mientras que los enlaces entre nodos serán una tupla (origen, destino). Dicho modelo permite añadir nodos a través de su método add_nodes_from (si introducimos los nodos como una lista de cadenas) o add_node (si deseamos añadirlos uno a uno), a su vez, los enlaces se pueden añadir a través de los métodos add_edges_from (si queremos añadir un conjunto de caminos a través de una lista de tuplas) o add_edge (si queremos introducirlos uno a uno, siendo el primer argumento el origen y el segundo argumento el destino). Se debe recordar que la red bayesiana debe ser tratada como un grafo acíclico dirigido, por lo que no debemos añadir caminos que produzcan algún ciclo y debemos tener en cuenta de la importancia de cuál es el origen y destino en los enlaces.

A continuación, describiremos cómo se calcula la inferencia y cómo entrenamos la red bayesiana. El cálculo de inferencias es la obtención de probabilidades de que suceda un fenómeno teniendo en cuenta las probabilidades de que sucedan otros fenómenos con los que esté relacionado. Por ejemplo, si "A" implica "B" y "C" implica "B", el cálculo de la inferencia de "B" tendrá en cuenta las probabilidades de que sucedan "A" y "C". Las inferencias son utilizadas para realizar predicciones de lo que pueda suceder.

El entrenamiento de una red bayesiana implica estimar las distribuciones de probabilidad condicional entre las variables, dadas las observaciones en los datos. Utilizando algoritmos como el de aprendizaje como el de BDeu. Este proceso permite que la red bayesiana capture las relaciones probabilísticas entre las variables y se ajuste a los datos disponibles, lo que la hace útil para inferir relaciones causales y hacer predicciones en entornos con incertidumbre, es decir, en los cuales que una cosa ocurra, no asegura que ocurra su consecuencia..

IV. Descripción del modelo

En esta sección describiremos el modelo que hemos planteado. Su descripción es en rasgos generales, ya que más adelante detallaremos las variables y sus valores, así como las relaciones entre ellas. El modelo escogido es una red bayesiana con un total de 11 nodos. Cada nodo representa una variable. En este caso los nodos son el porcentaje de personas mayores de 25 años con estudios iguales o superiores a la Educación Secundaria Obligatoria, porcentaje de población urbana, tasa de paro, ingreso nacional bruto, producto interior bruto, gasto educativo, gasto en salud, tasa de fertilidad, tasa de mortalidad, migración neta y tasa de crecimiento de la población. Es un grafo acíclico dirigido diseñado para poder estimar si la tasa de crecimiento de la población de España aumentará o disminuirá de acuerdo a ciertos parámetros. En este caso, el modelo se centra en factores que afectan el desarrollo socioeconómico y la salud de una población para estimar su tasa de crecimiento. Los valores de partida son el porcentaje de personas mayores de 25 años con al menos la ESO completa, el porcentaje de urbanización del país, la migración neta y el ingreso nacional bruto. Como salida obtenemos la tasa de crecimiento de la población del país. Nuestro objetivo es observar si puede predecirse el crecimiento de la población en España dependiendo de ciertos factores que aparentemente no están del todo relacionados. Los datos utilizados van desde 1975 hasta 2022, y nos han permitido entrenar la red. No obstante, nos habría gustado tener algunos más de ciertas variables ya que hay años de los que no tenemos valores para algunas de ellas. La mayoría de

datos han sido recabados del banco mundial de datos [9] . Con respecto a la implementación nos hemos basado sobre todo en el ejemplo de los tutoriales de la documentación de pgmpy titulado “Una red bayesiana para modelar la influencia del consumo de energía en los gases de efecto invernadero en Italia”. [10]

V. Descripción de la red

Nuestra red está diseñada de la siguiente manera. Véase la imagen a continuación.

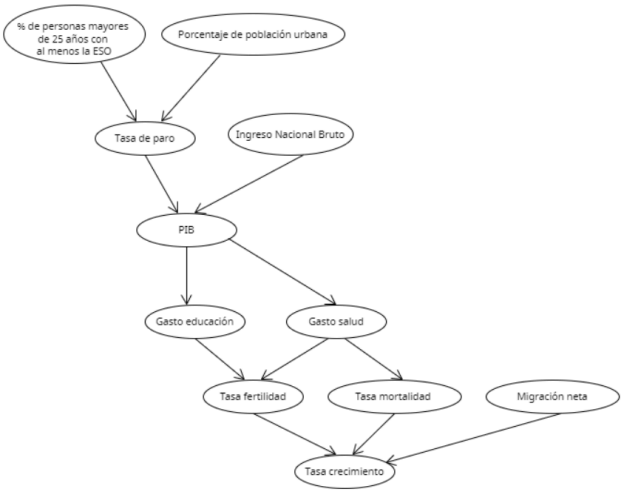


Fig. 1. Imagen de nuestra red bayesiana representada con UMLet [11]

Nuestra red consta de 11 nodos, los cuales describiremos en la sección siguiente. No obstante, aquí hablaremos de manera general la lógica seguida así como las entradas y la salida. Cabe destacar que nos hemos asegurado de que no posea ciclos a la hora de construirla. Esta a su vez ha sido diseñada y validada por el profesor antes de implementarla.

Como entrada la red posee 4 nodos, el porcentaje de urbanización del país, el porcentaje de personas mayores de 25 años con al menos la ESO, el ingreso nacional bruto y la migración neta. Los dos primeros influyen en la probabilidad de la tasa de paro. Una vez obtenida, y teniendo en cuenta el INB (Ingreso Nacional Bruto), observamos el PIB, el cuál influirá de manera directa sobre el gasto en educación y sanidad. El gasto en salud por un lado influye en la tasa de mortalidad y de fertilidad, mientras que el de educación solo afecta a la tasa de fertilidad. Por último, y teniendo en cuenta además la migración neta, obtenemos la salida de la red, la cuál es la tasa de crecimiento, que se ve influida por la tasa de fertilidad, la tasa de mortalidad y la migración neta.

VI. Nodos

En esta sección procederemos a explicar en detalle los nodos, es decir, las variables tenidas en cuenta para construir la red bayesiana. Antes de entrar en materia deberemos aclarar previamente que todas las variables están centradas en España, por lo que los datos de las mismas son anuales. Además de esto, es necesario clarificar también que todas ellas las trabajaremos con sus valores anuales. A continuación se procederá a realizar una descripción de cada uno de los nodos existentes de la red:

Porcentaje de personas mayores de 25 años con estudios iguales o superiores a la Educación Secundaria Obligatoria (RTE): Su nombre es bastante descriptivo. Esta variable representa el porcentaje de personas que tienen más de 25 años y que poseen un nivel de educación igual o superior a la ESO, es decir, al menos terminaron el instituto.

Porcentaje de población urbana (RTU): Esta variable lo que representa es la cantidad de población urbana que hay en España, es decir, el grado/nivel de urbanización del país.

Tasa de paro (RTP): Es el porcentaje de la población que carece de empleo.

Ingreso nacional bruto (INB): Es una variable obtenida del banco mundial de datos. También conocido como el INB, o el INB per cápita, refleja el promedio de ingresos de los ciudadanos de un país, y se calcula dividiendo el valor en dólares de los ingresos totales en un año de los residentes por el número de habitantes a mitad de año. El INB es una medida de la capacidad de un país para brindar bienestar a su población. [12]

Producto interior bruto (PIB): Es la cantidad total de dinero producida por el país en un año. Es decir, y citando la definición de PIB del Banco Santander: “El Producto Interior Bruto (PIB) mide el valor de todos los bienes y servicios producidos en un período -normalmente un año- en una economía. El PIB es un indicador que se utiliza para conocer la riqueza que genera un país.” [13]

Gasto educativo (GSE): Es el dinero total invertido por parte del Estado en educación durante el año.

Gastos en salud (GSS): Es el dinero total invertido por parte del Estado en salud durante el año.

Tasa de fertilidad (RTF): Se define como la cantidad de nacimientos por cada mujer en un año.

Tasa de mortalidad (RTM): Se define como la cantidad total de fallecidos registrados en un año. Esta es la única variable que hemos obtenido de una fuente distinta al banco mundial de datos, en este caso, del Instituto Nacional de Estadística (INE).

Migración neta (MGN): Es el resultado de calcular la diferencia entre la cantidad de personas que inmigran y la cantidad de personas que emigran del país.

Tasa de crecimiento de la población (RTC): Es una variable que se utiliza para medir la tasa de crecimiento de la población de un país, en este caso España. [14]

Hemos seleccionado estas variables y no otras porque son las más relevantes a la hora de calcular la tasa de crecimiento del país y también de las que más se habla.

VII. Enlaces

En la siguiente sección enumeraremos y describiremos la relaciones entre los nodos (enlaces), no sin antes dar un significado de lo que representan. Un enlace en la red bayesiana significa que un nodo cualquiera tiene una relación con otro. De este modo, los enlaces se representan como aristas en el grafo y relacionan nodos dos a dos. Si el enlace va dirigido desde el nodo A al nodo B, eso significa que el nodo A influye sobre el B. Esto no involucra que la relación se de en viceversa al tratarse de grafos dirigidos. Ahora pasemos a describir las relaciones:

Relación porcentaje de personas mayores de 25 años con estudios iguales o superiores a la Educación Secundaria Obligatoria con tasa de paso: Esta relación lo que nos viene a decir es que la cantidad de personas que tienen al menos la ESO cursada influye en la tasa de paro, ya que muchas veces para trabajar es necesario tener algún grado universitario o un grado superior.

Relación porcentaje de población urbana con tasa de paro: Esta relación nos representa que según el porcentaje de urbanización del país podemos tener una tasa de paro u otra. Realmente aquí hay muchos más factores a tener en cuenta pero entre ellos se encuentran los siguientes: En zonas rurales de la España vaciada es difícil encontrar trabajo, en grandes ciudades muchas personas intentan buscar oportunidades laborales, en zonas costeras y/o turísticas se reduce mucho el desempleo en temporadas de venida de turistas. Además, por norma general las empresas suelen establecer sus oficinas y sedes en ciudades, y construir fábricas en zonas más alejadas de la ciudad.

Relación tasa de paro con PIB: Esta relación es intuitiva, y es que a mayor PIB, menor tasa de paro. No obstante, y sorprendentemente, esto está anunciado como una ley en la economía denominada "Ley de Okun". La Ley de Okun, formulada por el economista Arthur Okun, sugiere que hay una correlación negativa entre la tasa de crecimiento del PIB y la tasa de desempleo. Es decir, cuando el PIB crece a un ritmo más rápido de lo normal (por encima de su potencial), el desempleo tiende a disminuir. [15]. En resumen, a mayor paro, menor PIB.

Relación INB con PIB: En este caso la relación es más que directa, El INB incluye el PIB pero lo ajusta por los ingresos netos recibidos del exterior. Es decir, INB es la suma del PIB y los Ingresos Netos del Exterior. Es decir, a mayor INB, mayor PIB.

Relación PIB con gasto en educación: Normalmente la manera de representar los gastos en un país suele ser con porcentajes del PIB del mismo. En este caso, a mayor PIB, mayor cantidad de dinero disponible para invertir en educación.

Relación PIB con gasto en salud: La relación es la misma que con el gasto en educación pero aplicada a la salud. A mayor PIB payor cantidad de dinero disponible para gastar en salud.

Relación gasto en educación con tasa de fertilidad: Realmente esta relación puede ser un poco menos clara de ver y a la vez un poco polémica. El gasto en educación de algún modo repercute en nuestra vida sexual. Por un lado, un mayor gasto en educación hace que la población esté concienciada sobre el uso de métodos anticonceptivos, lo que previene muchos embarazos y reduce la tasa de natalidad [16], pero también da oportunidad a 2 cosas adicionales. El hecho de que las mujeres estudien y tengan acceso a universidades hace que todo ese tiempo que dedican a su carrera profesional no la estén dedicando a ser madres, lo cuál incrementa la edad a la que las mujeres tienen hijos. La Encuesta de Fecundidad 2018 del Instituto Nacional de Estadística de España encontró que conforme aumenta el nivel educativo se retrasa la edad a la maternidad de las mujeres con alto nivel de estudios. [17]

Relación gasto en salud con tasa de fertilidad: El gasto en salud que se hace en el país es importante a la hora de la fertilidad, ya que cuanto mejores sean las condiciones de parto a la hora de tener hijos, mejor y más eficientemente se asegurará la supervivencia de los neonatos. Además, hemos encontrado un estudio que demuestra que invertir dinero en técnicas de reproducción asistida favorece el crecimiento de la tasa de fertilidad. [18]

Relación gasto en salud con tasa de mortalidad: En este caso no hemos visto necesario mostrar estudios para esta relación ya que consideramos que es muy obvia. A mayor atención sanitaria (que se traduce en dinero que se invierte en salud) mejores tratamientos a enfermedades y mayor calidad de vida de la población, por tanto, menor tasa de mortalidad (mueren menos personas).

Relación tasa de fertilidad con tasa de crecimiento: Aunque puedan parecer lo mismo no son lo mismo. La tasa de fertilidad mide la cantidad de hijos promedios de una mujer, y la de crecimiento mide el porcentaje de la población que aumenta o disminuye durante un periodo de tiempo. Por tanto, la tasa de crecimiento tiene que tener en cuenta a la fertilidad para ser calculada, o lo que es lo mismo, la tasa de fertilidad influye en la tasa de crecimiento.

Relación tasa de mortalidad con tasa de crecimiento: Con esta relación ocurre algo parecido a lo que ocurre con la tasa de fertilidad. Como la tasa de mortalidad se tiene en cuenta al calcular la tasa de crecimiento, la relación es directa.

Relación migración neta con tasa de crecimiento: Finalmente, y como algo parecido a lo visto en las dos últimas relaciones, como la migración neta (diferencia entre inmigrantes y

emigrantes) influye en el cálculo de la tasa de crecimiento, la relación es directa.

VIII. Probabilidades tomadas

En esta ocasión, no ha sido necesario tomar ni decidir probabilidades, ya que la propia red es la encargada de calcular dichos valores según las relaciones entre los distintos atributos y los datos reales proporcionados.

No obstante, sí que fue necesario discretizar los datos relacionados a los nodos, ya que los datos que tomamos son obtenidos de un conjunto de datos numéricos continuos, mientras que para que la red bayesiana pueda funcionar de manera correcta todos los atributos deben estar discretizados.

IX. Modelo creado

El proceso de creación del modelo hemos seguido los siguientes pasos:

1. Importación de las librerías:

   Antes de nada, es necesario importar todas las librerías que vamos a utilizar.

2. Creación de la red, nodos y enlaces:

   Para la creación de la red se crea una instancia de la clase BayesianNetwork, a la misma se le van añadiendo sus nodos y enlaces correspondientes,

3. Lectura del csv:

   El csv lo leemos haciendo uso de la librería pandas, aplicando la función read_csv sobre la ruta relativa al conjunto de datos (“./csv/data.csv”).

```
There are 11 indicators in the dataframe.
```

	Year	1975	1976	1977	1978	1979	1980	1981	1982
0	Population growth (annual %)	1.081294	1.056852	1.029132	0.962877	0.881585	0.802964	0.710878	0.600390
1	GDP per capita growth (annual %)	-0.539097	2.217773	1.785657	0.490728	-0.836528	1.391315	-0.839887	0.640409
2	Unemployment, total (% of total labor force) (...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	Educational attainment, at least completed upp...	NaN	NaN	NaN	NaN	NaN	NaN	12.833750	NaN
4	Current health expenditure per capita (current...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Fig. 2. Trozo de la tabla obtenida al leer el csv

4. Limpieza y discretización de los datos:

   Como los datos aportados son datos continuos deberemos discretizarlos para poder trabajar con ellos correctamente. Cada variable continua ha sido discretizada de modo que pueda tomar solo 3 posibles valores (el equivalente a nivel conceptual de “alto, medio, bajo”), además de que los años donde no se han recopilado datos han sido obviados.

	RTC	PIB	RTP	GSE	GSS	INB	RTU	RTF	RTM	MGN	RTE
1975	1_ALTO	3_BAJO	NaN	NaN	NaN	3_BAJO	3_BAJO	1_ALTO	3_BAJO	2_MEDIO	NaN
1976	1_ALTO	2_MEDIO	NaN	NaN	NaN	3_BAJO	3_BAJO	1_ALTO	3_BAJO	2_MEDIO	3_BAJO
1977	1_ALTO	2_MEDIO	NaN	NaN	NaN	3_BAJO	3_BAJO	1_ALTO	3_BAJO	2_MEDIO	3_BAJO
1978	1_ALTO	3_BAJO	NaN	NaN	NaN	3_BAJO	3_BAJO	1_ALTO	3_BAJO	3_BAJO	3_BAJO
1979	1_ALTO	3_BAJO	NaN	NaN	NaN	3_BAJO	3_BAJO	1_ALTO	3_BAJO	3_BAJO	3_BAJO
1980	1_ALTO	2_MEDIO	NaN	NaN	NaN	3_BAJO	3_BAJO	1_ALTO	3_BAJO	2_MEDIO	NaN
1981	2_MEDIO	3_BAJO	NaN	3_BAJO	NaN	3_BAJO	3_BAJO	1_ALTO	3_BAJO	3_BAJO	NaN
1982	2_MEDIO	3_BAJO	NaN	NaN	NaN	3_BAJO	3_BAJO	1_ALTO	3_BAJO	3_BAJO	NaN
1983	2_MEDIO	2_MEDIO	NaN	NaN	NaN	3_BAJO	3_BAJO	1_ALTO	3_BAJO	3_BAJO	NaN
1984	2_MEDIO	2_MEDIO	NaN	NaN	NaN	3_BAJO	3_BAJO	1_ALTO	3_BAJO	3_BAJO	NaN

Fig. 3. Discretización de los datos

5. Entrenamiento:

   Entrenar la red ha sido relativamente simple. Bastaba con aplicar el método fit al modelo y proporcionarle los datos. Es bastante cómodo porque la propia red calcula los cpds, que son las probabilidades que hay sobre un nodo dependiendo de sus nodos padres o evidencias. Explicaremos todo esto más en detalle en una sección aparte.

6. Inferencia:

   Llegados a este punto calculamos las inferencias, para ambos casos hemos decidido tomar como variable a la que inferir nuestra variable objetivo, la tasa de crecimiento, mientras que para las evidencias tomadas hemos decidido tomar una de las posibles combinaciones existentes, eligiendo el caso para el que la migración neta y el ratio de mortalidad son bajas mientras que la tasa de fertilidad es alta.

   Por una parte, las inferencias exactas hemos decidido elegir la inferencia por Eliminación de Variable (Variable Elimination) y la inferencia por Propagación de Creencias (Belief Propagation). Al ser inferencias exactas, los resultados en ambos casos van a coincidir con los obtenidos en el modelo, por lo que no es posible realizar un estudio de cuál de ellas sería mejor en cuanto a precisión.

   Por otra parte, las inferencias aproximadas que hemos seleccionado nosotros han sido el Muestreo de Modelos Bayesianos (Bayesian Model Sampling) y la Inferencia Aproximada Utilizando Muestreo (Approximate Inference Using Sampling). Para estos casos hemos seleccionado que se haga un cálculo sobre 1000 ejemplos, resultando en que usar Bayesian Model Sampling daba un error de 0.0932, mientras que el error obtenido al usar Approximate Inference Using Sampling es del 0.0532. Esto provoca que entre las inferencias de aproximación, la más óptima sea Approximate Inference Using Sampling.

Tasa de crecimiento de la población	phi(Tasa de crecimiento de la población)
Tasa de crecimiento de la población(1_ALTO)	0.1868
Tasa de crecimiento de la población(2_MEDIO)	0.3626
Tasa de crecimiento de la población(3_BAJO)	0.4506

Fig. 4. Resultado de la inferencia exacta en los dos casos

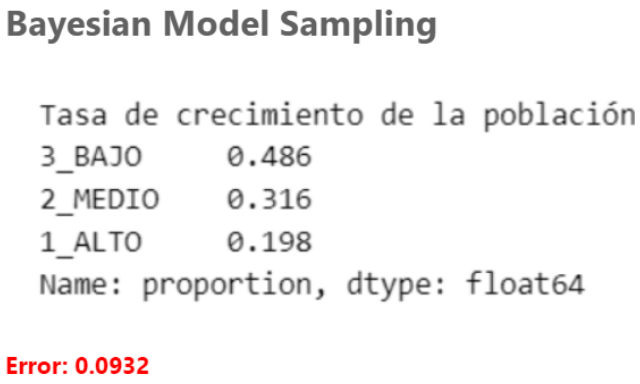


Fig. 5. Resultado de la inferencia aproximada con muestreo de modelo Bayesiano

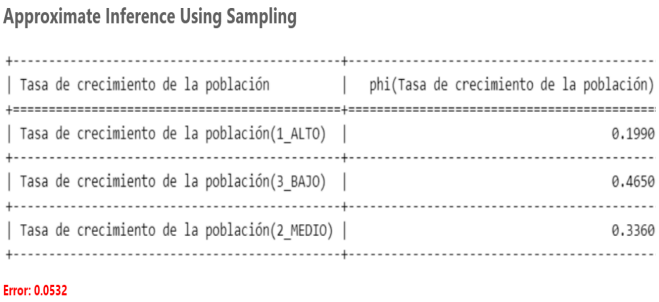


Fig. 6. Resultado de la inferencia aproximada con muestreo

X. Inferencia elegida

Durante el estudio de las inferencias hemos podido observar que entre las inferencias aproximadas parece ser que la Approximate Inference Using Sampling parece ser la más óptima, siendo que su error absoluto tomado es casi la mitad del que se obtiene aplicando la inferencia de Bayesian Model Sampling. No obstante, debido a que nuestra red no lleva una gran complejidad llega a ser más conveniente el uso de alguna de las inferencias exactas. Siendo en este caso la que elegiremos la Variable Elimination.

XI. Aprendizaje

Para el entrenamiento del modelo se utilizan todos los datos obtenidos previamente del csv de datos. A su vez, en dicho entrenamiento, aunque se puede utilizar el estimador por defecto MaximumLikelihoodEstimator, se ha decidido el uso como estimador el BayesianEstimator. Dicho estimador permite aprovechar una distribución previa conocida de datos. En el proceso de aprendizaje, el uso de BDeu permite que se generen N muestras uniformes para cada variable para calcular los pseudoconteos (en nuestro caso N=10), por lo tanto, las probabilidades estimadas en CPT son más conservadoras que las obtenidas a través de MLE (es decir, probabilidades cercanas a 1 o 0 se suavizan).

XII. Dificultades y decisiones de diseño

A lo largo de la realización del trabajo hemos presentado algunos problemas y hemos tomado decisiones de diseño. En esta sección los comentaremos.

Realmente no ha habido grandes problemas salvo el hecho de aprender a utilizar la página del banco mundial de datos y un problema que nos ha surgido en el entrenamiento debido a los cambios respecto las nuevas versiones de la librería al quedarse desfasado la clase que servía como modelo, BayesianModel. Esto se debe a que al hacer el modelo, el proceso que tomamos de ejemplo usaba BayesianModel como base, mientras que nosotros debíamos usar el modelo BayesianNetwork. En una primera instancia parecería que no debería dar problemas, pero una de las variables de la función fit tuvo modificaciones en la librería y dicho parámetro fue eliminado en BayesianNetwork. Debido a ello nos aparecieron problemas en nuestro código que pudimos solucionar con la eliminación de dicho parámetro. También tuvimos problemas en familiarizarnos con la inferencia, ya que era un concepto totalmente nuevo para nosotros y que no dominábamos bien. No obstante, al final pudimos calcular las inferencias seleccionadas de manera exitosa.

Con respecto a las decisiones de diseño, en un inicio no incluimos el tema de la inmigración y emigración, ya que no estábamos seguros de si sería necesario, y porque era añadir 2 nodos más y no sabíamos si podíamos tener más de 10 nodos. Sin embargo, tras validarlo con el profesor, añadimos la migración neta, que incluye dichos datos relacionados a la inmigración y emigración, aportando así ese factor adicional a la hora de tener en cuenta la tasa de crecimiento de una población, ya que en una sociedad también tienen hijos los extranjeros que viven en ella. Además de ello, también se decidieron añadir otras relaciones más entre los nodos que inicialmente no habíamos contemplado. Relacionado a lo anteriormente nombrado, como consejo del profesor, se ha decidido usar como modelo para nuestro código el ejemplo en los tutoriales de la página con el título “Una red bayesiana para modelar la influencia del consumo de energía en los gases de efecto invernadero en Italia”. Debido a ello, el entrenamiento de la red ha sido aplicado de manera similar en nuestra red.

XIII. Resultados

En esta sección se detallará tanto los experimentos realizados como los resultados conseguidos:

Tras haber expuesto todo lo comentado con anterioridad, podemos decir que hemos conseguido varias cosas. Por un lado hemos recopilado datos suficientes como para que nuestro modelo pueda calcular cpds sensatos y fiables. Dichos datos abarcan desde 1975 a 2022, sin embargo, al escoger un

menor conjunto para los datos de prueba hemos tenido que coger un conjunto de datos que fuese válido, en nuestro caso, de 1980 a 2005. Esto debido a que el banco de base de datos no proporciona suficientes datos para años más antiguos y produce el salto de una excepción al no proporcionar suficientes de ellos para alguna de las variables.

Para la obtención de resultados hemos realizado pruebas con el uso de dos inferencias exactas y otras dos inferencias aproximadas, tal y como se ha llegado a nombrar con anterioridad. A su vez hemos probado el modelo entrenado con una inferencia exacta, siendo la elegida la de por eliminación de variables, y la mejor inferencia aproximada, por inferencia aproximada sobre el muestreo, sobre un conjunto menor de datos de prueba. Esto nos ha dado como resultado datos bastantes similares a lo esperado, existiendo una mayor aproximación por parte de la inferencia exacta.

XIV. Conclusiones

Finalmente, se dedica la última sección para indicar las conclusiones obtenidas del trabajo. A lo largo de nuestro trabajo hemos podido obtener 4 conclusiones, las cuales enumeraremos y explicaremos a continuación.

1. Para este caso lo mejor es usar la inferencia exacta: Es cierto que es más costosa de calcular, pero no sólo los dos métodos para calcular la inferencia exacta obtienen el mismo resultado, sino que además ambas se corresponden con la realidad. Además, tenemos bastantes datos y aún así no tarda demasiado, por lo que la inferencia exacta es sin dudas la que mejor se adapta a nuestro modelo.

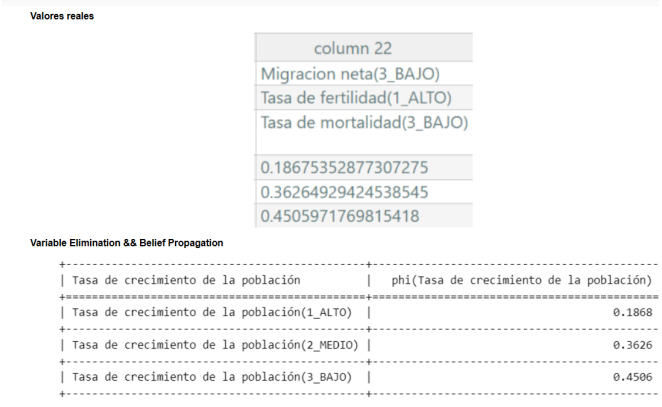
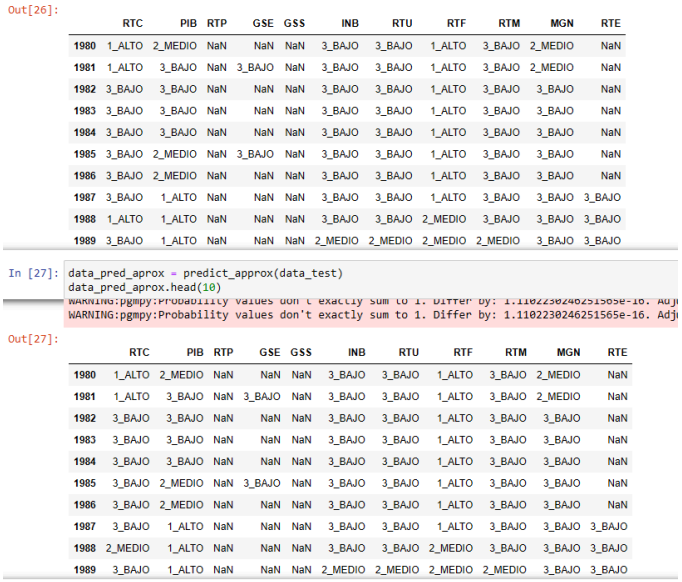


Fig. 7. Comparación valores reales con los predichos con inferencia exacta.

2. Por otra parte, podemos concluir que tenemos datos suficientes, porque hemos conseguido predecir de manera bastante acertada los valores de la variable objetivo (tasa de crecimiento del país). Esto lo hemos comprobado haciendo una prueba con un conjunto de datos menor (desde 1980 a 2005) y se puede observar que obtenemos los resultados similares que los que hay en los datos de prueba, por lo que la red se encuentra bien entrenada y predice bien lo que se pide.



	RTC	PIB	RTP	GSE	GSS	INB	RTU	RTF	RTM	MGN	RTE
1980	1_ALTO	2_MEDIO	NaN	NaN	NaN	3_BAJO	3_BAJO	1_ALTO	3_BAJO	2_MEDIO	NaN
1981	1_ALTO	3_BAJO	NaN	3_BAJO	NaN	3_BAJO	3_BAJO	1_ALTO	3_BAJO	2_MEDIO	NaN
1982	3_BAJO	3_BAJO	NaN	NaN	NaN	3_BAJO	3_BAJO	1_ALTO	3_BAJO	3_BAJO	NaN
1983	3_BAJO	3_BAJO	NaN	NaN	NaN	3_BAJO	3_BAJO	1_ALTO	3_BAJO	3_BAJO	NaN
1984	3_BAJO	3_BAJO	NaN	NaN	NaN	3_BAJO	3_BAJO	1_ALTO	3_BAJO	3_BAJO	NaN
1985	3_BAJO	2_MEDIO	NaN	3_BAJO	NaN	3_BAJO	3_BAJO	1_ALTO	3_BAJO	3_BAJO	NaN
1986	3_BAJO	2_MEDIO	NaN	NaN	NaN	3_BAJO	3_BAJO	1_ALTO	3_BAJO	3_BAJO	NaN
1987	3_BAJO	1_ALTO	NaN	NaN	NaN	3_BAJO	3_BAJO	1_ALTO	3_BAJO	3_BAJO	3_BAJO
1988	1_ALTO	1_ALTO	NaN	NaN	NaN	3_BAJO	3_BAJO	2_MEDIO	3_BAJO	3_BAJO	3_BAJO
1989	3_BAJO	1_ALTO	NaN	NaN	NaN	2_MEDIO	2_MEDIO	2_MEDIO	2_MEDIO	3_BAJO	3_BAJO

	RTC	PIB	RTP	GSE	GSS	INB	RTU	RTF	RTM	MGN	RTE
1980	1_ALTO	2_MEDIO	NaN	NaN	NaN	3_BAJO	3_BAJO	1_ALTO	3_BAJO	2_MEDIO	NaN
1981	1_ALTO	3_BAJO	NaN	3_BAJO	NaN	3_BAJO	3_BAJO	1_ALTO	3_BAJO	2_MEDIO	NaN
1982	3_BAJO	3_BAJO	NaN	NaN	NaN	3_BAJO	3_BAJO	1_ALTO	3_BAJO	3_BAJO	NaN
1983	3_BAJO	3_BAJO	NaN	NaN	NaN	3_BAJO	3_BAJO	1_ALTO	3_BAJO	3_BAJO	NaN
1984	3_BAJO	3_BAJO	NaN	NaN	NaN	3_BAJO	3_BAJO	1_ALTO	3_BAJO	3_BAJO	NaN
1985	3_BAJO	2_MEDIO	NaN	3_BAJO	NaN	3_BAJO	3_BAJO	1_ALTO	3_BAJO	3_BAJO	NaN
1986	3_BAJO	2_MEDIO	NaN	NaN	NaN	3_BAJO	3_BAJO	1_ALTO	3_BAJO	3_BAJO	NaN
1987	3_BAJO	1_ALTO	NaN	NaN	NaN	3_BAJO	3_BAJO	1_ALTO	3_BAJO	3_BAJO	3_BAJO
1988	2_MEDIO	1_ALTO	NaN	NaN	NaN	3_BAJO	3_BAJO	2_MEDIO	3_BAJO	3_BAJO	3_BAJO
1989	3_BAJO	1_ALTO	NaN	NaN	NaN	2_MEDIO	2_MEDIO	2_MEDIO	2_MEDIO	3_BAJO	3_BAJO

Fig. 8. Resultado de predecir los resultados con un conjunto de prueba más pequeño.

3. Valores de los cpds: Una vez hemos entrenado a la red y comprobado que predice bien, nos hemos puesto a analizar los cpds (influencias que tienen los nodos padres sobre los hijos). Hemos podido observar algunas cosas muy curiosas, como que un gasto alto en educación propicia un alto PIB, Además, es curioso porque el Ingreso Nacional Bruto y la migración neta no son tan relevantes a la hora de calcular la tasa de crecimiento de la población, ya que las probabilidades se distribuyen de una forma equitativa (siendo 1⁄3 cada una) lo que hace que el valor de la variable no influya de forma significativa en los cálculos. Aún así hay que tenerlas en cuenta.

	column 1	column 2
1	Migracion neta(1_ALTO)	0.3333333333333333
2	Migracion neta(2_MEDIO)	0.3333333333333333
3	Migracion neta(3_BAJO)	0.3333333333333333

Fig. 9. Cpd de la migración neta

4. Coherencia con la realidad: Por último, podemos concluir que las predicciones tienen sentido con la realidad. Esto se puede afirmar porque no sólo los datos predichos tienen relación con los verídicos, sino también porque siguen la lógica planteada en la definición de las relaciones, es decir, la población crece cuando la situación económica no es demasiado alta, y a medida que aumenta el gasto en educación y el nivel de urbanización, la tasa de crecimiento de la población disminuye. A lo largo tanto de los datos predichos como los reales podemos observar que la

tasa de crecimiento del país empezó a aumentar y con el tiempo ha ido decrementando.

XV. Mejoras y Trabajo Futuro

Como último punto, vamos a describir una serie de propuestas de mejoras y puntos de extensión para futuros trabajos relacionados. Alguna propuesta de mejora sería contrastar los resultados obtenidos con los resultados al trabajar con datos de otros países, como Francia o Portugal. Además sería interesante incluir más nodos como la edad media de la población del país, de modo que se puedan tener más cosas en cuenta. Además, como posible trabajo futuro podríamos intentar predecir la tasa de crecimiento de la población en los próximos años con predicciones de los valores de los nodos entrada de la red.

Referencias y Bibliografía

[1] Página que usamos para informarnos sobre qué son las redes Bayesianas: https://datascience.eu/es/matematica-y-estadistica/redes-bayesianas/,

[2] Pearl, J. (1988). Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. San Mateo, CA: Morgan Kaufmann.

[3] Russell, S., & Norvig, P. (2010). Artificial Intelligence: A Modern Approach (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

[4] Este documento se ha confeccionado siguiendo el formato de conferencias del IEEE.

[5] El entorno tecnológico utilizado ha sido jupyter, (https://jupyter.org/) y nos hemos apoyado en la librería pgmpy (https://pgmpy.org/)

[6] Enlace a nuestro proyecto de github: https://github.com/luchersol/Redes-Bayesianas

[7] Artículo sobre la natalidad en España vozpopuli.com/espana/ine-natalidad-bebes-bajo-2023-anos.html (Vozpopuli)

[8] Artículo sobre la natalidad: https://www.analesdepediatria.org/es-evolucion-natalidad-espana-analisis-tendencia-articulo-S1695403314001660

[9] Fuente de donde obtuvimos nuestros datos para realizar el trabajo: https://databank.worldbank.org/reports.aspx?source=2&type=metadata&series=SP.POP.GROW#advancedDownloadOptions y , https://www.ine.es/jaxiT3/Tabla.htm?t=6546

[10] Ejemplo en el que nos hemos basado para realizar el trabajo: https://pgmpy.org/detailed_notebooks/11.%20A%20Bayesian%20Network%20fo%20model%20the%20influence%20of%20energy%20consumption%20on%20greenhouse%20gases%20in%20Italy.html

[11] Herramienta para representar la red bayesiana: https://www.umlet.com/

[12] Definición de Ingreso nacional bruto: https://www.bancomundial.org/es/news/press-release/2015/07/01/new-world-bank-update-shows-bangladesh-kenya-myanmar-and-tajikistan-as-middle-income-while-south-sudan-falls-back-to-low-income#:~:text=El%20INB%20per%20c%C3%A1pita%20refleja,brindar%20bienestar%20a%20su%20poblaci%C3%B3n.

[13] Definición de PIB: https://www.bancosantander.es/glosario/pib-producto-interior-bruto

[14] Definición de tasa de crecimiento: https://www.indexmundi.com/es/datos/indicadores/sp.pop.grow

[15] Ley de Okun: https://es.wikipedia.org/wiki/Ley_de_Okun

[16] Relación anticonceptivos y educación feminina con la tasa de natalidad: https://www.healthdata.org/sites/default/files/2024-03/TL%20Capstones%20global%20fertility_ES.pdf

[17] Relación nivel educativo y tasa de fertilidad de las mujeres: https://ine.es/prensa/ef_2018_d.pdf

[18] Relación reproducción asistitda con tasa de fertilidad: https://academica-e.unavarra.es/bitstream/handle/2454/39818/Aparicio%20Morcillo%2C%20Andrea_TFG.pdf